

## Example based machine translation – a review and commentary

JOHN HUTCHINS

89 Christchurch Road, Norwich NR2 3NG, England (E-mail: WJHutchins@compuserve.com)

*Recent advances in example-based machine translation.* Edited by Michael Carl and Andy Way. Dordrecht: Kluwer Academic Publishers, 2003. xxxi, 482pp. (Text, Speech and Language Technology, vol. 21) ISBN: 1-4020-1400-7 (hardback), 1-4020-1401-5 (paperback).

### 1. General introduction

In the last decade the dominant models of machine translation (MT) have been data-driven or corpus-based. This is in sharp contrast to the dominant framework of the 1980s and previous decades, which was 'rule-based' (RBMT). In general, a distinction is made between, on the one hand, statistical machine translation (SMT), based primarily on word frequency and word combinations, and on the other hand, example-based machine translation (EBMT), based on the extraction and combination of phrases (or other short segments of texts). In both cases the corpora comprise bilingual texts (originals and their translations).

The origin of EBMT can be dated precisely to a conference paper in 1981 by Makoto Nagao (1984). Research, however, did not begin until the late 1980s at the same time as the first appearance of the translation memory (TM) as a translator's tool and the first research on SMT. The latter in particular gave rise to much dispute in the early 1990s. EBMT was associated with SMT as both were seen as variants of corpus-based approaches to MT systems, and during the 1990s both became familiar at MT conferences. In recent years, SMT has become the dominant (almost 'mainstream') approach in MT (as witnessed by the proceedings of almost any conference in the field of computational linguistics), and EBMT systems are less evident than SMT (but now more prevalent than RBMT).

The overall conception of SMT is now familiar – in essence, virtually all described models derive from the design first formulated in 1988 by the IBM group (Brown et al. 1988). Sentences of the bilingual corpus are first aligned, then individual words or word sequences (called 'phrases' or 'clumps' in SMT literature) of source language (SL) and target language (TL) texts are aligned, i.e. brought into correspondence. On the basis of these alignments are derived a 'translation model' of SL-TL frequencies and a 'language model' of TL word sequences. Translation involves the selection of most probable TL output for each input word or phrase and the determination of the most probable sequence(s) of words in the TL.

By contrast, the EBMT model is less clearly defined than the SMT model. Basically (if somewhat superficially), an MT system is an EBMT system if it uses segments (word sequences (strings) and not individual words) of source language (SL) texts extracted from a text corpus (its example database) to build texts in a target language (TL) with the same meaning. The basic units for EBMT are sequences of words (phrases, or 'fragments'), and the basic techniques are the matching of input strings against SL strings in the database, the extraction of corresponding TL strings and the 'recombination' of the strings as acceptable TL sentences. However, there is a multiplicity of techniques, many derived from other approaches, including methods used in RBMT systems, methods found in SMT, techniques used in translation memories (TM) etc., and there seems to be no clear consensus on what the basic 'model' (or design framework) of EBMT is and what it is not.

### 2. Contents of the collection

The aim of this collection is not only to present good exemplars of EBMT practice and to demonstrate the latest developments within this MT 'paradigm' but also to introduce some order and cohesion now that "the dust has settled" since the early 1990s when advocates of opposing sides ('corpus-based' vs. 'rule-based'), notably at the TMI conference in 1992 (TMI 1992), argued inconclusively about the merits of their respective approaches. The editors (Michael Carl and Andy Way) admit that even now there is no clear definition of what EBMT is. However, they emphasize that complete clarity in defining a field is not essential for its progress.

The editors introduce readers to the great variety of EBMT approaches and methods and provide the justification for the arrangement of papers in this collection. They adopt the distinction described by Turcato and Popowich (see below) between systems which make direct use of bilingual data during the actual translation processes and systems which derive information from corpora before application during translation. The former are labelled "run-time approaches", the latter are presented as "compiled approaches" of two basic kinds, which are distinguished by the types of derived and pre-compiled data: "template-driven EBMT" and EBMT using "derivation trees".

The articles in the collection are divided into four parts: Foundations, Run-time systems, Template-driven systems, and Systems using tree derivations. This sequence reflects (as the editors note) “a progression from approaches which are least rule-based to those which are most rule-based” or, alternatively, from those with least similarity (in techniques and methods) to ‘traditional’ RBMT systems to those with most similarity to RBMT systems. Indeed, the “non-run-time” approaches represent as much continuations of older RBMT models as they represent the newer EBMT framework.

The first part (*Foundations of EBMT*) begins with two papers by Harold Somers and by Davide Turcato and Fred Popowich which, together with the editors’ introduction, discuss the nature of EBMT, its distinctive features and how it does or does not differ from RBMT and SMT.

Somers, in an expanded version of his article in this journal (Somers 1999), provides an excellent overview of the methods and problems of EBMT (chap.1), covering also TM and SMT and drawing parallels with RBMT. He describes the basic features of example databases, various approaches to the ‘matching’ processes, extraction of TL equivalent examples and their ‘re-combination’, including problems of adaptation and ‘boundary friction’ (i.e. where combining TL fragments results in morphological or syntactic mistakes). After considering how EBMT might be evaluated, he summarises the basic features of EBMT: use of real language data, data-driven rather than theory-driven (no complex grammars, problems of rule conflict minimized), overcoming constraints of structure preservation (i.e. the almost inevitable reflection of SL structures in TL output despite the known differences between SL and TL discourse), expandability of corpora and the possibility of rapid development.

Davide Turcato and Fred Popowich set out (chap.2) to identify what makes a system example-based as opposed to rule-based. First they argue that use of a database of examples in a MT system is in itself no justification for labelling the system EBMT, since the ways in which system knowledge is acquired or expressed is irrelevant; what matters is how this knowledge is used in operation. On this basis, they compare ‘linguistically-principled’ EBMT systems and one type of transfer-based RBMT system (lexicalist ‘shake-and-bake’, e.g. Whitelock 1994). They conclude that only when EBMT has access to and makes use of the original full database of examples during the translation process itself that EBMT is clearly distinguished from RBMT systems. In other words, the original conception of ‘translation by analogy’ (as initially proposed by Nagao, 1984) represents “the most characteristic technique of EBMT” and the only true EBMT systems are those where the information is not pre-processed, where it is available intact and unanalysed throughout the matching and extraction processes.

Bróna Collins and Harold Somers (chap. 4) seek to establish a theoretical (principled) foundation for EBMT methodology on the technique of case-based reasoning (CBR). The parallels with EBMT, though obvious, have not been exploited and the paper seeks to redress this neglect by illustrating its application to processes of ‘adaptation’.

In the other article of this first section (chap. 3) Reinhard Schäler, Michael Carl and Andy Way describe how translation memories could be made more effective by searching below the sentence level, reducing “redundant, ambiguous or wrong” examples, improving searches for ‘fuzzy’ matches, and helping translators in the recombination of extracted TL segments. Such enhancements are claimed to be most effective and achievable in controlled environments, and they argue that EBMT is ideally suited for applications in restricted domains (sublanguages) and with controlled languages. The paper ends with a discussion of various ‘models’ for integrating different approaches (TM as well as RBMT and EBMT), including multi-engine systems and hybrid systems.

Part II (*Run-time approaches to EBMT*) begins with a paper by Emmanuel Planas and Osamu Furuse (chap. 5) which continues the TM theme, describing a method of ‘fuzzy matching’ (involving superficial lemmatization and shallow parsing) – with potential application in a “rudimentary” (word by word) EBMT system.

The paper by Eiichiro Sumita (chap.6) describes a full run-time EBMT system, using dynamic-programming matching, and thesauri for calculating semantic distances, and illustrated by Japanese to English speech translation (at ATR in Japan).

The article by Francis Bond and Satoshi Shirai (chap.7) introduces an EBMT method suitable for texts where parallelism of content is rare but where corpora contain many similar SL-TL examples at phrase levels. Alignment and template-building are both run-time processes and thus repeated for every new input sentence. As yet the system has no mechanisms for smoothing ‘boundary friction’ or for deleting extraneous output, and non-matching input words and sentences are translated by an RBMT system (ALT/JE), i.e. a hybrid approach to make best use of the strengths of RBMT and EBMT.

The last paper in this section by Tantely Andriamanankasina, Kenji Araki and Koji Tochinni (chap.8) describes an EBMT system intended for languages with unreliable or weak dictionaries and

with limited corpus resources. The only tool required is a POS tagger; it is not, therefore, a ‘pure’ run-time system, even though the pre-compiled data is much less than in other non-run-time EBMT systems. Output is checked and corrected manually, and the system then constructs links for the new SL and TL sentences, thus expanding the database with new example pairs. It is one of the few papers in this collection illustrating EBMT’s oft-claimed merit of expandability.

The last two articles in the previous section describe processes for the run-time construction of templates. In part III (*Template-driven EBMT*), papers describe methods of building templates from bilingual example corpora in advance of translation processes. Ilyas Cicekli and Altay Güvenir (chap.9) use templates in the form of words/lemmas with POS tags for a system with English as SL and Turkish as TL. The similarity metric for matching input against SL examples is claimed to be applicable to any pairs of languages and for translation in any direction. Human evaluators select the best results for retention (they refer to their system as “human-assisted EBMT”).

Ralf Brown describes (chap.10) the induction of transfer rules in the form of templates of word strings (of any length), which are then either interpreted as rules of a transfer grammar or added as new examples to the original corpus, and four variant algorithms are evaluated. The transfer rule induction used by Brown is somewhat simpler than that used by Cicekli and Güvenir in the previous paper, and Brown believes that adding a small amount of ‘seed knowledge’ to the grammar induction process could probably improve performance.

Kevin McTait (chap.11) also describes the derivation of “translation patterns” (templates) which resemble transfer rules in RBMT systems. Various linguistic resources are used (morphological analysis, POS tagging) to improve accuracy and recall – although they add to computational complexity. The algorithm seeks to overcome a weakness of dynamic-programming in respect of non-adjacent long-distance dependencies. Although the full system is not described in detail, it is sufficient to illustrate the computational complexity of recombination in EBMT systems. The author reviews the merits of three variants, and compares results (French-English) with the *Babelfish* (Systran) system.

The final essay in this part (chap.12) is by Michael Carl. He proposes an algorithm to induce a “translation grammar” from bracketed alignments of the bilingual (German-English) corpus texts, aided by a ‘seed’ dictionary with morphological information. The derived patterns are annotated with morphosyntactic information, and wrong alignments and bracketing are filtered out. The result is an EBMT transfer component which is strikingly similar to transfer grammars in RBMT systems.

The last four papers of the collection (in Part IV *EBMT and derivation trees*) are devoted to the pre-compiled preparation of templates with more structure than in the previous section. Kaoru Yamamoto and Yuji Matsumoto describe two studies extracting knowledge from an English-Japanese parallel corpus of business texts (chap.13). In the first study, word and phrase correspondences are derived using a statistical dependency parser, and three variants are evaluated. The second study compares the statistical dependency model with methods using word segmentation (plain n-gram) and ‘chunk’ boundaries; it is concluded that this method is most useful for preparing bilingual dictionaries in new domains (particularly for identifying compound nouns) while statistical dependency is most useful for disambiguation.

The paper by Hideo Watanabe, Sadao Kurohashi and Eiji Aramaki (chap.14) is devoted also to the extraction of translation patterns (templates) from paired dependency structures. Parsing errors are corrected by human post-editing. The aim is to construct the SL translation pattern (dependency structure) which would have produced the correct TL patterns. The results are added to the corpus, with potential improvements of overall performance.

Arul Menezes and Stephen Richardson discuss the extraction of structured transfer rules from bilingual corpora (chap.15). The corpus is analysed by a rule-based logical-form parser (designed to be language-independent), the representations for SL and TL strings in the corpus are then aligned by lexical correspondences and by structural information (based on POS tags), and frequency counts are computed for all SL-TL mappings. TL sentences are produced by a rule-based generation component from the output logical forms. A Spanish-English version of the system showed better quality output than *Babelfish* – although the authors rightly point out that it was not (and is not) possible to customize *Babelfish* to specific domains. In contrast to many other papers in the collection, Menezes and Richardson convey a picture of the whole Microsoft MT system – a hybrid approach combining rule-based, example-based and statistical techniques, similar in conception to the one presented by Bond and Shirai (chap.7).

The concluding paper is by Andy Way (chap.16), who proposes that example sentences of a corpus should be analysed into structural representations; in his case, using a variant of Lexical Functional Grammar based on data-oriented parsing. The main objective is to overcome problems when retrieving TL examples that may be ill-formed or incomplete and/or present difficulties of boundary friction.

Superficially, descriptions in this section seem to be in the old RBMT tradition, but the fundamental difference is that structures and rules are derived from actual examples in bilingual databases and not from formalisms and theories by system developers.

### 3. General comments and observations

At the end of this book does the reader have any clearer idea of what constitutes an EBMT system? Possibly not. The variety of methods and techniques, of the ways in which they interact, are all indicators of a thriving and productive research framework, but they do not make its definition any easier. The editors rightly point out that there is no single technique which can be labelled uniquely example-based, and there is no NLP method which can be excluded *a priori* from application in EBMT. The borrowings from RBMT and SMT are manifold. Indeed, the editors state (p. xxi) that “EBMT subsumes those approaches and systems which people decide to call ‘example-based’” The editors may – with some justification – say that definition is unnecessary in practice since the lack of clear definitions does not hamper research in many areas of the computer sciences. Indeed, a definition may be premature. On the other hand, it can be argued that a clearly identifiable framework or model helps not only outside observers to understand the aims of EBMT but that it also provides researchers a firm foundation for assessing progress and identifying areas in need of further investigation.

The first point to be made (with Somers, chap.1) is that example-based methods can be implemented in systems which are not themselves EBMT systems. One of the initial impetuses for research in example-based approaches was the recognition that rule-based methods have major problems with non-compositionality and collocational expressions. Researchers at ATR (Sumita et al. 1990) on a speech translation system found that the definition of rules for translating Japanese *no* into English were both highly complex and often inadequate. More satisfactory was the collection of examples of actual usage (phrases containing the word in context) and their English equivalents. The approach was seen as supplementary to and an expansion of the rule-based system. Later, however, the approach was adopted as the basis for a new type of MT system – example-based MT (EBMT) – and from the late 1980s there have essentially been two strands of EBMT research: the creation of example-based methods applicable in any MT system, and the establishment of EBMT as an MT model in its own right. The question is then how EBMT systems can be defined (see also Hutchins 2005), and what components can be categorized as being example-based.

#### 3.1. Defining EBMT

It is not sufficient to say that EBMT is ‘data-driven’ in contrast to ‘theory-driven’ RBMT and that EBMT is ‘symbolic’ in contrast to ‘non-symbolic’ SMT. The first distinction implies that EBMT has no theoretical stance regarding the use of data and that RBMT ignores data which may conflict with theory; the second ignores the fact that many EBMT systems include statistical (non-symbolic) methods (alignment, fuzzy matching) as well as symbolic methods, and that some SMT systems include symbolic treatments (lemmatization, parsing) as well as statistical ones. It is also not enough to define EBMT as simply MT which makes use of databases of translation examples; as Turcato and Popowich stress (chap.2): what matters is how the data is used in translation operations.

A definition can be based on considering how EBMT differs from SMT and RBMT in both the core bilingual operations (the conversion of SL strings into TL strings) and in the ancillary monolingual operations. In the case of RBMT, systems are commonly distinguished by their core bilingual operations: whether SL-TL conversion operates via an intermediary language-neutral representation (interlingua-based MT), via structure transduction from SL representation to TL representation (transfer-based MT), or via piece by piece conversion of SL fragments into TL fragments using dictionaries and rules (‘direct translation’ or transformer-based MT). The ancillary operations in RBMT include the preparatory “analysis” stage and the succeeding “synthesis” (or generation) stage. In the case of SMT the core bilingual process is the ‘translation model’, based on statistics derived from bilingual corpora, which substitutes TL words or phrases for SL words or phrases. It is preceded by the preparation of the database (alignment of bilingual corpora, producing correlations of SL strings and TL strings) and it is succeeded by sequencing of TL fragments using a monolingual ‘language model’ (also prepared in advance). In TM, the core ‘translation’ operation is performed by human translators who select equivalent TL phrases from possibilities presented to them in a database (the translation memory). In the case of EBMT the core bilingual process is the matching of SL fragments (from an input text) against SL fragments, and the retrieval of equivalent TL fragments (as potential partial translations). Other processes are ancillary. The preceding analysis stage in EBMT may be trivial (as it in SMT, i.e. simple decomposition into words) but, most often, it is more complex: identifying strings as comparable to aligned elements in the database, creating potential templates by generalising some strings as variables, or parsing input strings as structured

representations. The succeeding process of recombination is also not part of the core EBMT process since it is a monolingual process, and its nature is determined by the form in which TL fragments are extracted. Likewise, the alignment of bilingual corpora is a secondary process since it is a consequence of the requirement that the matching process has available a database of equivalent SL-TL fragments.

We may go further. Differences in the kinds of methods used in the core processes itself are secondary aspects (as far as the definition of EBMT as such is concerned). In other words, whether the matching and retrieval (transfer) involves pre-compiled fragments (papers in parts III and IV), whether the fragments are derived at 'run-time' (papers in part II), and whether the fragments are templates containing variables (as in part III) or structured (part IV), are all secondary factors. It follows also that the use of variables, of 'fuzzy matching', of templates and patterns, of statistical methods, of RBMT-like linguistic methods, etc., are all ancillary techniques subordinated to the core EBMT process. What distinguishes EBMT from RBMT and SMT is the nature of the core bilingual processes (matching, retrieval, adaptation).

EBMT systems have much in common with both RBMT and SMT. The closest parallels between EBMT and RBMT are found when systems use structural transformations (in analysis, matching, extraction, recombination/synthesis), but they are present also whenever individual SL lexical items are substituted by individual TL lexical items (e.g. in templates). There is also the parallel (mentioned by Somers in ch.1) between the 'traditional' three-stage transfer-based RBMT model (analysis, transfer, generation) and the three-stage EBMT model (analysis/matching, extraction/retrieval, recombination). The parallels with SMT are found when EBMT systems use statistical methods for matching and retrieval, and also if SMT systems use syntactic parsing (e.g. Charniak et al. 2003). Where SMT differs from both is the use of a two-stage model with similarities to the earliest RBMT 'direct translation' systems: SL to TL lexical substitution ('translation model') and TL word sequence rearrangement ('language model') – the difference being, of course, that the old RBMT model was rule-based (SL-TL dictionaries), while SMT is statistics-based.

What distinguishes EBMT from both RBMT and SMT is that the basic processes of EBMT are analogy-based, i.e. the search for phrases (fragments) in the database which are similar to input SL strings (isolated by segmentation) and their adaptation and recombination as TL phrases and sentences. (This was the conclusion of Turcato and Popowich (chap.2).) Neither RBMT nor SMT seek 'similar' strings, both search for exact matches of input words and strings and produce sequences of words and strings as output. The original conception of Nagao (1984) is therefore maintained in current EBMT research; it is seen most clearly in those systems based on matching and extracting templates – strings with variables – where both matching and extraction seeks similarities of one string and another. It is seen also in the use of thesauri to aid 'fuzzy' matching, and in the use of structured representations where extraction is based on partial matches of parsed strings.

Is EBMT a variant of SMT? Or is SMT a variant of EBMT? Both of the 'foundation' articles by Somers and by Turcato and Popowich regard SMT as 'non-symbolic' variants of EBMT. Irrespective of the fact that this would not be the view of SMT researchers (indeed, they might regard EBMT as a 'more symbolic' version of SMT), and that it ignores the use of 'non-symbolic' methods in EBMT and the use of 'symbolic' methods in some SMT systems, the discussion above indicates that the two can be clearly distinguished by their approaches to the basic processes of SL to TL conversion. In EBMT, it is seen as a search for analogous examples to guide production of TL forms equivalent to the SL input. In SMT it is seen as the maximization of the probable correlations between SL forms and TL forms. Put somewhat simplistically, EBMT is analogy-based and SMT is correlation-based. Neither is a variant of the other.

Is there a pure EBMT design? If there is one, it would probably be one which makes exclusive use of analogy-based techniques and makes no use of pre-compiled templates or structured databases. This is the essence of Turcato's and Popowich's argument in chap.2, namely that the only pure EBMT design is 'run-time EBMT', where operations are based on access to the full database of examples, i.e. not adjusted or selected or analysed in any form. Examples of the approach are Jones (1992), who argues for a connectionist model; and, more recently, Lepage and Denoual (2005), who demonstrate the feasibility (practically and not just theoretically) of basing a system purely on unanalysed examples, in a small experiment with short sentences from the tourism domain (English-Japanese). However, their suggested improvements involve the use of dictionaries and paraphrases, which would diminish the 'purity' of the approach – as, indeed, the pure SMT design of Brown et al. (1988), formalized in Brown et al. (1993), has been 'compromised' by the use of morphological and syntactic information. In fact, purity of design is difficult to maintain, and some form of hybridization seems to be inevitable.

### **3.2. Methods and problems in EBMT**

What makes a procedure an example-based method? Is it sufficient for a method to be an EBMT method if it derives information for its procedures from a database of examples? Should it not also be a requirement that the procedures involve the manipulation (matching, extraction, adaptation) of whole segments of actual examples from the corpus? The alignment of texts in a bilingual database would not in itself be an example-based method, since the results could be used in RBMT, SMT or TM. The analysis of a database of examples for the derivation of rules would not in itself constitute an example-based method, since the results could be used in a conventional RBMT system.

The use and derivation of templates in EBMT systems is undoubtedly a truly example-based process, albeit one which may have parallels in earlier RBMT systems – the difference being that the EBMT templates derive from actual texts and not (as they are often in RBMT) from the intuitions of system developers. Templates can be quite complex constructions which would be notoriously problematic in RBMT, such as (1) (simplified from Carl, chap.12, p.358), where N might stand for any name (Peter) or reference to a human, and NP for a phrase template such as (2):

(1) *Von N wird erwartet, dass er mit NP vertraut ist* ↔ *N is supposed to know NP*

(2) *den Gesetzen seines eignen Landes* ↔ *the laws of his own country*

The matching of input sentences becomes quite complex, particularly if different tenses and moods are present (*wurde, war, würde, wäre; was, would be, might be, etc.*) In most cases, however, simpler templates are derived (e.g. Cicekli and Güvenir, chap. 9),

(3) *X bought Y for Z,*

where variables stand for words or word sequences, e.g. *X* as *I, you, the man, the old man, the stranger in the bar, etc.*; *Y* as *wine, beer, medicine*; *Z* as another name; etc.

Variables of templates may also be derived with type constraints, e.g. grammatical and semantic constraints on the contexts of verb and tense usage (Cicekli 2005). Similarly, variables might stand for case/role functions, for parts of speech or for semantic features (as found in many RBMT systems: transitive verb, instrument, location, animate, name, etc.) and relate indirectly to example strings in the database. This is illustrated by Saha and Bandyopadhyay (2005) who use semantic tags in templates for news headlines (e.g. time, location, animal, artefact, etc.) One further step would be the replacement of all elements (words) in templates by variables, which would render them virtually identical to strings of categories (grammatical or lexical) in RBMT.

There are few overt examples of semantic disambiguation methods in EBMT: Kaji et al. (1992) use semantic categories (sport, instrument) as variables in templates for the ambiguous verb *play*, Richardson et al. (1993) use the semantic contexts of examples for word sense disambiguation, and Sumita (chap.6; also Sumita et al. 1993) uses a thesaurus to measure semantic distances when matching SL strings to database examples. In most cases, semantic equivalences are implied from the correlations of aligned words, phrases, templates or structures in the database, and disambiguation is thus performed via the use of templates and statistical information. In general, disambiguation does not extend beyond clause and sentence bounds – an exception is Brown (2005) who applies statistical methods for disambiguation over a wider context.

The use of semantic features and categories for templates and the use of disambiguation techniques distinguishes EBMT from SMT, since the latter may be characterized (at least in its present state) as being ‘semantics-free’, i.e. with operations based exclusively on surface forms of words and strings in texts. As yet, EBMT systems have not incorporated the more complex ‘deeper’ semantic analysis and representation methods found in RBMT systems. There is, however, no reason why an EBMT system could not fully annotate bilingual databases with semantic features for use in matching and retrieval procedures. Such semantic annotations would support measures of similarity and semantic distance and augment the use of thesauri and wordnets – all in accord with the basic analogy principle of EBMT.

While alignment, matching and retrieval in EBMT are identical in principle to comparable methods in SMT, other methods are to be found also in RBMT. Lemmatization, morphological analysis and syntactic markers are applied in the segmentation and decomposition of SL input (e.g. Cicekli and Güvenir (chap.9), McTait (chap.11); Gough and Way 2004b). Syntactic analysis (parsing) is found in those EBMT systems where examples are treated as structured representations: the parsing is applied both to input sentences and to SL and TL examples in the database (e.g. Watanabe et al. (chap.14), Menezes and Richardson (chap.15), Way (chap.16); Watanabe 1993; Langlais et al. 2005). The grammars applied are those found in recent RBMT research (dependency grammar, lexical functional grammar, case grammar, etc.). Dependency grammar is preferred (as illustrated in part IV of this collection); but as yet there seems to be no discussion of which formalisms are most suitable for EBMT.

The principal argument for structuring representations is that the processes of adaptation can be improved and that, in particular, many problems of boundary friction in recombination can be

averted. These are the processes most intrinsic to EBMT – they are the basis of the analogy approach – and they are, as Somers (chap.1) writes, the most difficult and relatively least investigated aspects of EBMT. As a simple example of boundary friction in recombination (from Way, chap.16), the database may include the aligned English-German sentences in (4)-(5).

(4) *A small dog eats a lot of meat* ↔ *Ein kleiner Hund frisst viel Fleisch.*

(5) *I have two ears* ↔ *ich habe zwei Ohren.*

The input sentence (6) is decomposed as *I have*X and *a small dog*. Combining the fragments in the database would give (7), which is ungrammatical as the German noun phrase corresponding to ‘a small dog’ is in the nominative case instead of the accusative (*einen kleinen Hund*).

(6) I have a small dog.

(7) *Ich habe ein kleiner Hund.*

Clearly, the TL fragment should be adapted for combination in the new context. Way’s solution is the syntactic parsing of both database examples and input sentences. In so doing, however, the database would have to be expanded with more examples and there would be greater complexity in matching and retrieval processes – as well, of course, in the recombination process. An alternative could be the morphological tagging of nouns (noun phrases, and verbs) in the database and of input sentences (fragments), and the inclusion in the database of information about noun declensions and verb conjugations – but that would be an RBMT solution and no longer an EBMT one. The problem of boundary friction is not limited to EBMT; it is present also in SMT whenever TL fragments are joined as putative sentences. In SMT, the solution is of course reference to large monolingual corpora in a language model providing information about most probable word sequences. This option could also be used in EBMT, as Somers suggests, and as Groves and Way (2005) have demonstrated – the result is a ‘hybrid’ EBMT-SMT system.

However, boundary friction may well not be the only area for problems in recombination. It is not unlikely that EBMT (and SMT) will suffer from the loss of information about relationships between fragments extracted from input sentences. After segmentation, SL fragments are treated as separate entities; they are matched against fragments in the database and their corresponding TL fragments are retrieved. The recombination procedure then attempts to join TL fragments without information about relationships between their original SL fragments. A number of EBMT systems (e.g. McTait (chap.11), Way (chap.16); Watanabe 1993) propose full parsing of SL input and this procedure is certainly capable of reducing the problems of recombination – at the cost of greater computational complexity.

The lack of any way for dealing with sentence-level relationships between TL fragments has various consequences. The production of fluent TL output requires more than effective recombination of fragments. Fluency is also a function of smooth transitions from clause to clause and sentence to sentence. The lack of information about intra- and intersentential relations affects also the aim of EBMT to get away from the constraints of structure preservation, one of the main failures of RBMT systems. And it means that EBMT cannot deal with anaphora at the sentence level (i.e. intrasentential anaphora) and at the paragraph and discourse level (intersentential anaphora) – there can be a solution only with some kind of discourse model. The treatment of anaphora has been the subject of substantial work in RBMT research albeit with minimal practical success, but whatever its value, it is difficult to see how it could be incorporated within the current framework of EBMT.

### 3.3. Complexity and expandability

One of the strengths of EBMT systems is claimed to be their capacity to improve by adding more examples to their databases (Somers (chap.1), Andriamanankasina et al. (chap.8), Watanabe et al. (chap.14)). However, some of the EBMT systems described here require substantial analysis of data, e.g. introduction of variables, structure parsing, which make expansion less easy. In fact, as Somers emphasizes, expansion does not necessarily improve performance: more text may mean superfluous examples, unusual and infrequent examples, misleading examples, etc. – on the problem of corpus quality see Denoual (2005) – and it may also limit the potential reusability of databases. On the other hand, as in SMT, there is the converse problem of inadequate data for some languages and there have been efforts to create EBMT databases without parallel corpora (e.g. Vandeghinste et al. 2005).

Expandability may not however be an important issue for EBMT. If the main application for EBMT is seen as the development of special-purpose systems (and not general-purpose systems where EBMT would compete with established RBMT systems and with SMT systems), then databases will be compiled for linguistic information within specific domains. Papers in the collection illustrate a few of such applications: stock market reports (chap.7), health (chap.11), business (chap.13), software documentation (chap.15). Domain restriction means that databases will not be expanded to cover wider ranges of vocabulary; there will be no need to add further examples.

The complex interaction of syntactic rules and the resulting unpredictability of outcomes has been a weakness of RBMT systems. While this particular complexity has been avoided in EBMT, other processes are potentially equally complex: algorithms for matching templates, for extracting and selecting TL examples and for adaptation and recombination. Researchers on EBMT are well aware of the dangers – in this collection, particularly Collins and Somers (chap.4), Planas and Furuse (chap.5), Cicekli and Güvenir (chap.9), Brown (chap.10), and McTait (chap.11). As yet there is no evidence for how much complexity might be expected in large-scale EBMT systems – much of course depends on which methods are adopted in particular systems. As far as statistical techniques are concerned there is evidence from SMT that the algorithms for matching and extracting in the translation model and for reordering and synthesizing in the language model require substantial computational resources.

As in RBMT, complexity can be greatly reduced by the use of controlled language input. Schäler et al. (chap.3) suggest that controlled language applications are most appropriate for EBMT, since here some of the main strengths of EBMT (quality control of example translations, context-specific matching and extraction) can be exploited. The results of an experiment by Gough and Way (2004a) (cf. also Way and Gough 2005b) suggest that controlled language EBMT can indeed lead to improved translations. However, controlled language applications of EBMT would not be without problems and implications. Controlled language EBMT would require augmentation or modification of database examples in order to conform to the ‘rules’ of the controlled language. This may involve exclusion of ‘unauthorized’ text fragments and the insertion of ‘authorized’ examples. Databases could not be easily derived automatically from substantial text corpora, except those already controlled – and how many of these are available? Templates would also have to conform to the norms of the controlled language. In either case, the result is no longer EBMT open to the heterogeneity of actual language – one of its major strengths. In so far as the output is regulated it is determined by operations (‘rules’) which the developers or users define. Controlled EBMT is thus a step closer to RBMT or, at least, to an EBMT-RBMT hybrid.

### **3.4. Hospitality and hybridization**

The openness of EBMT to a wide range of different techniques and methods – amply demonstrated in this collection – would seem to make EBMT the epitome of hybrid MT, i.e. with the potentiality to adopt the best and most effective methods and to minimize the worst effects of others within a common framework. As indicated already, there may in fact be many different versions of EBMT hybrids. We could talk of EBMT-RBMT hybrids (e.g. core example-based transfer with RBMT syntactic analysis) and of EBMT-SMT hybrids (e.g. core example-based transfer with SMT-like language model for recombination processes). We might likewise talk of basically RBMT and SMT systems with EBMT components (e.g. templates for dealing with collocations and idiomatic constructions), i.e. RBMT-EBMT and SMT-EBMT hybrids. In this collection, Bond and Shirai (ch.7) and Menezes and Richardson (ch.15) refer to their systems as hybrids, closer to the basic EBMT model than to RBMT or SMT. A later paper from the Microsoft research group (Menezes and Quirk 2005) illustrates even greater hybridization by merging EBMT-type parsing with a phrase-based SMT model. Given the closeness of EBMT and SMT techniques (particularly the use of aligned corpora and the use of statistical methods for matching and retrieval) it is surely inevitable that there will be more examples of EBMT-SMT hybrids such as the use of SMT ‘language models’ in a basically EBMT system and the augmentation of SMT aligned corpora with EBMT-like phrases (Groves and Way 2005).

Rather than seeing these various types of MT systems as ‘hybrids’ it may be preferable to define different MT models according to their mix of statistical, example-based and rule-based methods, as Wu (2006) does. It would still be possible to categorize systems as basically EBMT or SMT or RBMT according to the nature of their core SL-TL conversion processes (as proposed here), but with admixtures of a wide variety of techniques. The adoption and integration of many different methods has in any case been the general characteristic of MT research since the 1970s and earlier – it has only been the previous dominance of the interlingua- and transfer-based RBMT models and the recent dominance of SMT models which has obscured this general hospitality. But in EBMT it is a fundamental feature of the model itself, hence the suggestion above that EBMT may be the epitome of MT hybridization.

### **3.5. Evaluation and commercialization**

The real test for any MT system (whether RBMT, EBMT or SMT) is how well it performs on previously unseen texts. As Somers’ discussion of evaluation (chap.1) indicates, most EBMT evaluations are focused on the effectiveness of particular procedures or methods used in one system, particularly matching and retrieval: (cf. Sumita (chap.6), Cicekli and Güvenir (chap.9), Yamamoto and Matsumoto (chap.13), Brown, 2005; Denoual, 2005; Doi et al., 2005; Langlais et al. 2005; Menezes

and Quirk, 2005; Vandeghinste et al. 2005). However, as yet there seem to be no measures for evaluating the optimal size and content of EBMT databases; there are few evaluations of full EBMT systems; and relatively few comparative evaluations. In this collection the papers by McTait (chap.11) and Menezes and Richardson (chap.15) compare their systems with the RBMT system *Babelfish* (Systran); Gough and Way (2004a, 2004b) compare results with the RBMT system *Logomedia*; Way and Gough (2005a) compare EBMT results with a word-based SMT system, and Groves and Way (2005) compare results with a phrase-based SMT system. By contrast, in SMT it has become the norm for developers to test the efficacy of any changes in design, and to include comparisons with other MT systems (mainly RBMT and not EBMT). Evaluation is fully integrated into SMT research, and it ought to be in EBMT research as well. Many more comparative evaluations of EBMT systems are needed (with SMT as well as RBMT) before the impartial observer will be able to judge the value of example-based methods – and for this purpose more sensitive evaluation metrics than BLEU, NIST, WER, SER, etc., may be needed. Until such comparisons have been made and accepted, EBMT will remain a sideline to what is increasingly seen as the main thrust of MT research, namely investigations of SMT systems.

The fact that there are as yet no commercial or working EBMT systems in immediate prospect suggests that no EBMT model is considered robust enough. This is not surprising. It is in fact only very recently that SMT has become commercial – with the launch of *LanguageWeaver*'s systems – some 15 years after the model was first proposed by Brown et al. (1988). There had been a similar timescale for RBMT, since the first reasonably adequate working systems did not appear until the mid 1960s, about 10 years after the first demonstration system. EBMT has been an active research interest as long as SMT, so perhaps full working EBMT systems will come within the next few years.

Whatever the commercial prospects, EBMT represents an important branch of MT research – even if often overshadowed by SMT activity – and this collection must be required reading for anyone interested in current methods and techniques of MT research, whatever their inclinations. It will surely remain a basic text and source for EBMT for many years to come.

## References

- Brown, P., J.Cocke, S.Della Pietra, V.Della Pietra, F.Jelinek, R.Mercer and P.Roossin: 1988, 'A statistical approach to French/English translation'. *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*. Pittsburgh: Carnegie-Mellon University. 16pp.
- Brown, P.F., S.A.Della Pietra, V.J.Della Pietra and R.L.Mercer: 1993, 'The mathematics of statistical machine translation: parameter estimation.' *Computational Linguistics* **19** (2), 263-311.
- Brown, R.: 2005, 'Context-sensitive retrieval for example-based translation' *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp. 9-15.
- Charniak, E., K.Knight and K.Yamada: 2003, 'Syntax-based language models for statistical machine translation', *MT Summit IX: proceedings of the Ninth Machine Translation Summit, New Orleans, USA, September 23-27, 2003*; pp.40-46.
- Cicekli, I.: 2005, 'Learning translation templates with type constraints'. *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp. 27-33. [Revised version titled 'Inducing translation templates with type constraints' in this issue of *Machine Translation*.]
- Denoual, E.: 2005, 'The influence of example-data homogeneity on EBMT quality'. *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp. 35-42.
- Doi, T., H. Yamamoto and E.Sumita: 2005, 'Graph-based retrieval for example-based machine translation using edit-distance'. *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp. 51-58.
- Gough, N. and A.Way: 2004a, 'Example-based controlled translation'. In *Proceedings of the 9th EAMT Workshop: "Broadening horizons of machine translation and its applications"*, 26-27 April 2004, Valetta, Malta; pp.73-81.
- Gough, N. and A.Way: 2004b, 'Robust large-scale EBMT with marker-based segmentation'. In *TMI-2004: proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation*, October 4-6, 2004, Baltimore, Maryland, USA; pp.95-104.
- Groves, D. and A.Way: 2005, 'Hybrid example-based SMT: the best of both worlds?' In *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, Ann Arbor, Michigan, USA; pp. 183-190. [Extended revised version 'Hybrid data-driven models of machine translation' in this issue of *Machine Translation*.]

- Hutchins, J.: 2005, 'Towards a definition of example-based machine translation' *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp.63-70.
- Jones, D.: 1992, 'Non-hybrid example-based machine translation architectures'. In *TMI (1992)*; pp. 163-171.
- Kaji, H., Y.Kida and Y.Morimoto: 1992, 'Learning translation templates from bilingual text'. In *Coling-92: proceedings of the Fifteenth [sic] International Conference on Computational Linguistics*, Nantes, France; pp. 672-678.
- Langlais, P., F.Gotti, D.Bourigault and C.Coulombe: 2005, 'EBMT by tree-phrasing: a pilot study.' *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp. 71-80. [Revised version by Langlais and Gotti in this issue of *Machine Translation*.]
- Lepage, Y. and E.Denoual: 2005, 'The 'purest' EBMT system ever built: no variables, no templates, no training, examples, just examples, only examples'. *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp. 81-90. [Revised version in this issue of *Machine Translation*]
- Menezes, A. and C.Quirk: 2005, 'Dependency treelet translation: the convergence of statistical and example-based machine-translation?' *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp.99-108. [Revised version by Quirk and Menezes in this issue of *Machine Translation*.]
- 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* (Amsterdam: North-Holland); pp.173-180.
- Richardson, S.D., L.Vanderwende and W.Dolan: 1993, 'Combining dictionary-based and example-based methods for natural language analysis'. In: *TMI-93: the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, Japan, July 14-16, 1993. Proceedings; pp.69-79.
- Saha, D. and S.Bandyopadhyay: 2005, 'A semantics-based English-Bengali EBMT system for translating news headlines'. *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp. 125-133.
- Somers, H.: 1999, 'Review article: example-based machine translation'. *Machine Translation* **14** (2), 113-157.
- Sumita, E., H.Iida and H.Kohyama: 1990, 'Translating with examples: a new approach to machine translation'. In *Third International Conference on Theoretical and Methodological Issues in Machine Translation*, 11-13 June, 1990, University of Texas; pp.203-212.
- Sumita, E., O.Furuse and H.Iida: 1993, 'An example-based disambiguation of prepositional phrase attachment'. In *TMI-93: the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, Japan, July 14-16, 1993. Proceedings; pp.80-91.
- TMI: 1992, *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, June 25-27, 1992, Montréal, Canada. Laval (Québec): CWARC.
- Vandeghinste, V., P.Dirix and I.Schuurman: 2005, 'Example-based translation without parallel corpora: first experiment on a prototype.' *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Second Workshop on Example-Based Machine Translation*; pp. 135-142.
- Watanabe, H.: 1993, 'A method of extracting translation patterns from translation examples.' In *TMI-93: the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, Japan, July 14-16, 1993. Proceedings; pp.292-301.
- Way, A. and N.Gough: 2005a, 'Comparing example-based and statistical machine translation', *Natural Language Engineering* **11** (3), 295-309.
- Way, A. and N.Gough: 2005b, 'Controlled translation in an example-based environment: what do automatic evaluation metrics tell us?', *Machine Translation* **19** (1), 1-36.
- Whitelock, P.: 1994, 'Shake-and-bake translation'. In C.J. Rupp, M.Rosner and R.L.Johnson (eds.) *Constraints, language and computation* (London: Academic Press), pp.339-359.
- Wu, D.: 2006, 'MT model space: statistical vs. compositional vs. example-based machine translation'. *Machine Translation* [this issue]