Towards a definition of example-based machine translation

John HUTCHINS
89 Christchurch Road
Norwich NR2 3NG, UK
WJHutchins@compuserve.com

Abstract

The example-based approach to MT is becoming increasingly popular. However, such is the variety of techniques and methods used that it is difficult to discern the overall conception of what example-based machine translation (EBMT) is and/or what its practitioners conceive it to be. Although definitions of MT systems are notoriously complex, an attempt is made to define EBMT in contrast to other MT architectures (RBMT and SMT).

1 Introduction: why a definition is needed

The dominant framework until the late 1980s was what is now known as 'rule-based' machine translation (RBMT). Since then, research has been dominated by corpus-based approaches, among which the primary 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 parts of texts).

The overall conception of SMT is now fairly familiar – in essence, all described models derive from the design first formulated in 1988 by the IBM group (Brown et al. 1988).[1] Sentences of the bilingual corpus are first aligned, and then individual words of SL and 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 words for each input word and the determination of the most probable sequence of those selected words in the TL. The basic units for SMT systems are words; but recently longer segments are being taken into account (see section 8 below.)

The EBMT model is less clearly defined than the SMT model. Basically (if somewhat superficially), a 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 thus sequences of words (phrases).

Within EBMT there is however a plethora of different methods, a multiplicity of techniques, many of which derive from other approaches: methods used in RBMT systems, methods found in SMT, some techniques used with translation memories (TM), etc. In particular, there seems to be no clear consensus on what EBMT is or what it is not. In the introduction to their collection of EBMT papers (Carl & Way 2003), the editors – probably wisely – refrain from attempting a definition, arguing that scientific fields can prosper without clear watertight frameworks, indeed may thrive precisely because they are not so defined.

2 Original conceptions of EBMT

As a preliminary definition, we may identify the basic processes of EBMT as: the alignment of texts, the matching of input sentences against phrases (examples) in the corpus, the selection and extraction of equivalent TL phrases, and the adaptation and combining of TL phrases as acceptable output sentences.

In its original conception (e.g. Nagao 1984), EBMT seems to have been regarded primarily as a means of overcoming the deficiencies of RBMT systems, namely their weaknesses when translating between languages of greatly differing structures, such as English and Japanese, and therefore in generating good quality output – particularly in the treatment of collocations, e.g. translations of *yabureru* in: The bag *was broken* and The president *was defeated* in the election, and the different translations of Japanese *kakeru* according to 'context': *hang* something on a tree, *put* something on/over one's shoulder, *cover* someone or something.[2] Examples were thus to be treated like other SL and TL data, i.e. as tree representations. Hence, input sentences were analysed as far as possible, and transfer using examples was initiated when rules and trees failed.

[1] This SMT 'model' is not the only possibility, but others have rarely, if ever, been investigated.

[2] The first examples come from Nagao 1984, the second ones from Sato and Nagao 1190.

Two tendencies emerged: some researchers (e.g. Sumita et al. 1990) used examples to supplement (improve) RBMT systems and were unsure whether EBMT could or should deal with the whole process of translation, while others (e.g. Sato and Nagao 1990) were encouraged to investigate 'pure' EBMT systems, where the basic process was founded on finding examples of TL sentences "analogous to" input SL sentences, and rules were applied only when examples could not be found in the database.

These two tendencies persist. On one hand, example-based methods are used in what are basically RBMT systems and are essentially seen as developments of the MT tradition, and on the other hand, there is the conviction that EBMT represents in itself a new 'paradigm' – much as SMT researchers argue that their MT architecture represents a new paradigm. (Personally, I am reluctant to use the term 'paradigm' since it suggests the near complete overturn and virtual rejection of all preceding research in the field – as Kuhn (1962) originally conceived the term in connection with theories in the pure sciences, specifically physics. While some SMT researchers may see their approach as completely new, others in recent years have begun to incorporate methods from older periods. In the case of EBMT, most researchers appear to see their efforts as continuations of traditional approaches and readily acknowledge their predecessors. For such reasons, I prefer to refer to new 'architectures' or 'frameworks'.[3])

One argument for exploring EBMT approaches is that since it is based on actual texts, output translations should be more readable and more sensitive to contexts than RBMT systems, i.e. of higher quality in appropriateness and idiomaticity. A second argument is that EBMT systems can be more easily improved, by the addition of more examples from bilingual corpora; whereas the improvement of RBMT systems involves the modification and addition of complex rules and lexical entries. A third is that EBMT does not involve the complexities of lexical and structural transfer found in (most) recent RBMT systems, i.e. that the basic architecture of EBMT is simpler and less prone to failure than RBMT. As a fourth point, it is argued that EBMT can deal with cases of translation involving complex structural differences and subtle lexical choices that RBMT often fails at. In general, the argument in favour of EBMT is its potential to improve the generation of TL sentences.

3 Definitions of EBMT by Somers, and by Turcato and Popowich

As a starting point for approaching a definition of EBMT, we shall consider the article by Harold Somers (1999), reprinted in revised form in the Carl-Way collection. In this excellent overview of EBMT, he provides outline characterisations of the chief processes and methods encountered in EBMT research. These include the content, size and organisation of databases of parallel bilingual text corpora – how they are selected (e.g. for a domain, as controlled texts) and edited (e.g. to reduce redundancy and potentially disruptive 'unusual' examples), how they are aligned, whether texts are tagged, analysed as tree representations, etc. Likewise, there are options in the processes of matching (character based, word based, structure based), measures of similarity (e.g. statistical and/or by reference to thesauri), the adaptation of extracted examples and their 'recombination' to produce TL sentences. He points out that 'recombination', despite its crucial role for EBMT (whose major objective is to generate better quality output than RBMT), is the most neglected area of EBMT research – and the Carl-Way collection (2003) confirms this relative neglect. Finally, Somers outlines the actual and potential applications of EBMT (or EBMT-like) techniques and approaches in other MT architectures, specifically the derivation of dictionaries and grammar rules for RBMT systems, and the role of EBMT in multi-engine and 'hybrid' systems.

Somers rightly points out that the use of what are claimed to be 'EBMT methods' does not mean that systems are EBMT systems. 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. What does Somers see as the essence? Firstly, "the use of a bilingual corpus is part of the definition, but this is not sufficient", since almost all current MT research (including RBMT systems) make use of text corpora to define and limit or constrain the range of data they are aiming to cover – at least in the initial stages of development. As a closer definition, Somers offers: "EBMT means that the

[3] It could be argued that corpus-based approaches as a whole represent a new departure in contrast to the preceding rule-based approaches. In so far as previous work is reconceptualised and reformulated in new frameworks the 'sudden' introduction of corpus-based MT in the late 1980s could be termed a 'paradigm shift' in the Kuhnian sense. This could be true, even though corpus-based approaches were quite common in the earliest days of MT research (e.g. the Rand project), before the rise of grammatical formalisms (Bar-Hillel, Harris, Chomsky, etc.) led to the domination of rule-based architectures in MT research.

main knowledge base stems from examples". But, example sentences can be used in RBMT systems as source data from which generalized rules and patterns can be derived,[4] and the databases of SMT systems are also derived from corpora of 'example' texts. A more restrictive and defining characteristic for EBMT is that "the examples are used at run-time". As Somers comments, this definition excludes SMT from the EBMT framework, since the data used in SMT is derived in advance of the translation process. In addition, the 'run-time' condition appears to exclude many of the EBMT systems described in the Carl-Way collection.

In an article following Somers' overview, Davide Turcato and Fred Popowich take issue with Somers' definition. Their aim is to set out a framework for defining the core processes of EBMT, i.e. to identify or isolate what makes a system example-based as opposed to rule-based. First they agree that use of a database of examples in a MT system is in itself no justification for labelling the system EBMT, since (they argue) the ways in which system knowledge is acquired or expressed is irrelevant; what matters is how 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') – since this type (unlike other RBMT systems) also avoids structural transfer. The aim is to clarify the status of example databases. If EBMT can be shown to be equivalent in operation with a system (such as lexicalist RBMT) which makes no use of an example database, then either EBMT has to be defined in terms which make no reference to an example database or the characterization of EBMT rests upon knowledge acquisition rather than knowledge use – which, with Somers, they have already rejected as a valid defining characteristic. A crucial question is how sentences are decomposed during the EBMT matching process in comparison with decomposition (i.e. analysis) in lexicalist RBMT. Any MT system has to deal with constructions which cannot be translated compositionally; it needs to have access to a repository of 'non-monotonic contexts' (examples). In RBMT, the repository is extracted (created) from dictionary or text sources; in EBMT the repository used in operation *may* also be extracted from the resource (the example database) as 'explicit knowledge'. In this case, the "linguistic information used by EBMT is indistinguishable from the information used by lexicalist MT." However, in other EBMT architectures, there may be direct reference to the example database during the processing of sentences (i.e. during translation). In this case the repository is used as an 'implicit knowledge' database. Turcato and Popowich argue that it is only when EBMT has access to and makes use of the original full database of examples *during* the translation process that EBMT is clearly distinguished from RBMT systems. In other words, the original conception of 'translation by analogy' (as initially proposed by Nagao) represents "the most characteristic technique of EBMT" and it is "the one where the use of entire examples is most motivated." Such complete access can only be available if the EBMT system has *not* already processed examples (as 'explicit knowledge'). In other words, they suggest that the only true EBMT systems are those where the information is not pre-processed, is available intact and unanalysed throughout the matching and extraction processes, i.e. as the systems in the Carl-Way collection using example databases as 'implicit' knowledge during 'run time'. Even such use does not finally define EBMT since 'translation by analogy' could also "in principle… be an extension to a traditional transfer MT system, to solve cases of lexical ambiguity for which no direct evidence is found in a translation database".[5] In effect, Turcato and Popowich imply that a close definition of EBMT is unimportant; the main thing is to make good MT systems.

However, there are two major problems with such conclusions. Firstly, it does not help observers and indeed other MT researchers if it is said by EBMT practitioners themselves that there is no definition of EBMT; they need to know how EBMT differs from other MT architectures. Secondly, restriction of EBMT to the use of 'implicit knowledge' at run time only would seem to be too narrow, since it would exclude much of the research reported in the Carl-Way collection and at recent conferences. On the other hand, to say simply that, in effect, a system is an EBMT system if its authors say it is, is not the answer.

4 EBMT in the context of MT in general

The attempt here to define EBMT starts from a broader perspective, starting from identifying the core processes and components of *any* MT system and how these differ in RBMT, EBMT and SMT.

In any MT system the core must be the process by which elements (entities, structures, words, etc.)

[4] Carbonell et al. 2002 and Lavoie et al. 2001 describe current RBMT systems which induce rules from corpora.

[5] This was the motivation for the EBMT work of Sumita et al. 1990.

of the input (SL) text are converted into equivalent elements for the output (TL) text, where the output text means the same (or is functionally equivalent to) the input text.[6] In all cases there are processes of 'analysis' preceding this core conversion (or 'transfer') and processes of 'synthesis' (or 'generation') succeeding conversion.

1. In RBMT, the core process is mediated by bilingual dictionaries and rules for converting SL structures into TL structures, and/or by dictionaries and rules for deriving 'intermediary representations' from which output can be generated. The preceding stage of analysis interprets (surface) input SL strings into appropriate 'translation units' (e.g. canonical noun and verb forms) and relations (e.g. dependencies and syntactic units). The succeeding stage of synthesis (or generation) derives TL texts from the TL structures or representations produced by the core 'transfer' (or 'interlingual') process.

2. In SMT, the core process involves a 'translation model' which takes as input SL words or word sequences ('phrases') and produces as output TL words or word sequences. The following stage involves a 'language model' which synthesises the sets of TL words in 'meaningful' strings which are intended to be equivalent to the input sentences. In SMT the preceding 'analysis' stage is represented by the (trivial) process of matching individual words or word sequences of input SL text against entries in the translation model. More important is the essential preparatory stage of aligning SL and TL texts from a corpus and deriving the statistical frequency data for the 'translation model' (or adding statistical data from a corpus to a pre-existing 'translation model'.) The monolingual 'language model' may or may not be derived from the same corpus as the 'translation model'.

3. In EBMT, the core process is the selection and extraction of TL fragments corresponding to SL fragments. It is preceded by an 'analysis' stage for the decomposition of input sentences into appropriate fragments (or templates with variables) and their matching against SL fragments (in a database). Whether the 'matching' involves pre-compiled fragments (templates derived from the corpus), whether the fragments are derived at 'run-time', and whether the fragments (chunks) contain variables or not, are all secondary factors. The succeeding stage of synthesis (or 'recombination' as most EBMT authors refer to it) adapts the extracted TL fragments and combines them into TL (output) sentences. As in SMT, there are essential preparatory stages which align SL and TL sentences in the bilingual database and which derive any templates or patterns used in the processes of matching and extracting.

We may note that in practice clear distinctions between stages may not be present, or some stages may even appear to be absent. In many RBMT systems there is a conflation of transfer and generation; some indeed conflate analysis and generation in a single 'transfer' process (in the transformer or 'direct translation' model). In various EBMT systems (or proposals) we see a conflation of matching and extraction – indeed, it could be argued that 'matching' is not a part of 'analysis' since it does not involve decomposition (or rather it follows decomposition) but is an integral part of the core (conversion or 'transfer') stage. In many EBMT systems, analysis may be as trivial as in SMT, consisting simply of the dividing of sentences into phrases or word strings on the basis of 'markers' (e.g. prepositions, conjunctions, punctuation; see e.g. Gough and Way 2004). In most cases, however, parts of the derived segments are further converted into templates or tree structures (i.e. 'normalised') before the matching process.

5 The database

However, the definition is not yet complete. Essential for any translation – a consequence of the aim to maintain 'meaning equivalence' – is access to information about correspondences of vocabulary in the SL and the TL. The information contained in a database may be derived from a variety of resources (bilingual and monolingual texts, bilingual and monolingual dictionaries, grammars, thesauri, etc.)

Before the arrival of corpus-based approaches (SMT and EBMT) it would be assumed that an MT system has to have a bilingual dictionary of some kind and a set of rules to deal (at very least) with differences of word order between SL and TL. In SMT, the dictionary is *largely* replaced by a bilingual text corpus (aligned in order to correlate SL sentences and words and TL sentences and words) and the rules are replaced by information about frequencies of correlations between SL words and TL words ('translation model') and collocations of TL words in texts ('language model'). In EBMT the dictionary is *largely* replaced by an aligned bilingual text corpus (the set of 'examples') and the rules are replaced by examples of TL strings in the text corpus. In both SMT and EBMT there may also be supplementary

[6] Meaning equivalence is the aim, but in practice MT output can be useful when falling short of this ideal, e.g. in contexts where readers need only to understand and grasp the 'essence' of messages and/or where output can be edited (post-edited) to produce appropriate and acceptable texts.

use of traditional bilingual dictionaries, and perhaps also of monolingual thesauri. If it is acknowledged that dictionaries represent generalisations of analyses by linguists and language users, culled from previous readings of texts, then bilingual RBMT dictionaries are also derived from text corpora.[7] In this light, the distinctions between RBMT on the one hand and SMT and EBMT on the other regarding the use of dictionaries and bilingual corpora also become secondary.

Can we go further and argue that it is essential also to have access to information necessary for decomposing (analysing) and combining (generating) sentences? Before EBMT and SMT it was assumed that systems require knowledge about the morphology and syntax (and probably also semantics) of both SL and TL. The rules used in RBMT were derived (explicitly or implicitly and indirectly) from observations of pattern frequencies between and within languages. In EBMT and SMT, information about well-formedness of sentences and strings is implicitly incorporated in the bilingual databases. The information is implicitly 'extracted' for matching and conversion in so far as input strings have to conform to the practices of the SL, otherwise matches will not be found. Likewise information is implicitly utilised in the synthesis stages by reference to a monolingual 'language model' (in SMT) and by the extraction of well-formed TL fragments (in EBMT). In sum, knowledge about sentence formation, explicit in RBMT, is still present implicitly in EBMT and SMT.

6 The essence of EBMT: a definition

If it is agreed that the essence of any MT system is to be located in the method(s) used to convert a SL string into a TL string, then this would locate the defining essences of MT architectures where they are most distinctive. RBMT systems are commonly distinguished by whether SL-TL transformation 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). Likewise, the comparable operation in SMT is the 'translation model' based on statistics derived from bilingual corpora which substitutes TL words or phrases for SL words or phrases. In TM systems, the comparable operation is performed by human translators who select equivalent TL phrases from the possibilities presented to them in a database (the translation memory).

In EBMT, therefore, the essence is the matching of SL fragments (from an input text) against SL fragments (in a database) and the extraction of the equivalent TL fragments (as potential partial translations). In this light, whether the 'matching' involves pre-compiled fragments (templates derived from the corpus), whether the fragments are derived at 'run-time', and whether the fragments (chunks) contain variables or not, are all secondary factors – however useful in distinguishing EBMT subtypes (as Carl and Way (2003) in their collection). Input sentences may be treated as wholes, divided into fragments or even analysed as tree structures; what matters is that in transfer (matching/extraction) there is reference to the example database and not, as in RBMT, the application of rules and features for the transduction of SL structures into TL structures. Consequently, the 'analysis' of SL input is secondary, its form dependent on the way examples are treated in the core 'transfer' process (and therefore stored in the database). Likewise, it can be argued that the operations of synthesis ('recombination'), perhaps the most difficult and complex in EBMT systems, are a consequence of the nature of the output from the matching/extraction process, i.e. because the input has been decomposed, because what are extracted from the database are not full sentences. Likewise, the alignment of bilingual corpora is a secondary process since it is a consequence of the requirement that the matching process has available sets of corresponding SL-TL fragments. Finally, in this framework, the use of variables, the use of 'fuzzy matching', of templates and patterns, etc., are all ancillary techniques in relation to the core EBMT process.

In the light of the definition being put forward here, the distinctions made by Turcato and Popowich (2003) between 'run time' EBMT and other systems are also secondary. In 'run time' systems, the full database of examples is made accessible and subject to any manipulation as required during matching and extracting processes (e.g. Sumita 2003). Such use of the database is ancillary (however essential) to the basic operation of converting SL input into TL output. In other EBMT systems, the analysis of the database is made in preparatory operations, before actual SL texts (input sentences) are processed for translation – i.e. as explicitly ancillary operations (e.g. McTait

[7] It follows that, as Somers and Turcato-Popowich point out, RBMT systems could also use bilingual corpora instead of (manually or automatically derived) bilingual dictionaries.

2003). The argument that systems which do not access the whole corpus during translation are not 'true' EBMT systems is no longer valid. What matters is the way SL fragments are converted into TL fragments in the core (transfer) process. The 'knowledge base', how it is derived and how it is structured, is secondary, albeit crucially important. Therefore, EBMT knowledge used during the core process can be either fully prepared in advance as 'explicit knowledge' or it can be adapted (adjusted) to the specific input as 'implicit knowledge' during translation operations. This may have important consequences computationally and for recall and precision in the retrieval and selection of examples, but choice between 'explicit' and 'implicit' knowledge remains secondary (as far as a definition of EBMT is concerned). What we have, therefore, are two sub-types of EBMT, both subsumed in the general framework outlined above. Indeed, if we consider the types of systems described in the Carl-Way collection we have probably more than two sub-types since it seems that clear differences are discernible between systems which use templates or patterns and systems which use derived (tree) structures.

To summarise the definition: MT systems are EBMT systems if the core 'transfer' (or SL-TL conversion) process involves the matching of SL fragments (sentences, phrases, strings) from an input text, the matching of such fragments against a database of bilingual example texts (in the form of strings, templates, tree representations), and the extraction of equivalent TL fragments (as partial potential translations). The databases of EBMT systems are derived primarily from bilingual corpora of (mainly) human translations, and are pre-processed in forms appropriate for the matching and extraction processes performed during translation (i.e. 'run-time' processes). The processes of analysis (decomposition) and synthesis (recombination) are designed, respectively, to prepare input text for matching against the database and to produce text from database output.

7 EBMT and RBMT

The proposed definition does not specify the structure of the 'knowledge base' (the database of examples) or the kinds of representations involved in the core 'transfer' process. However, whatever form they do have – simple 'surface' strings, strings with variables, templates, or structured (tree) representations – the crucial point is that they are derived from actual examples of SL and TL sentences.

However, when these representations are in forms similar to (or even identical with) those found in RBMT systems, the question arises whether their inclusion is stretching the framework of EBMT too far. The more input is analysed and the more structured the examples in the database, the less EBMT appears to differ from traditional RBMT.

There are clearly gradations in what can be accepted as EBMT representations, from unstructured strings with no variables at one end of the spectrum to dependency trees of input and example sentences at the other end of the spectrum. The 'simple' matching of input strings (after segmentation) against unstructured example SL sentences (strings of 'surface' forms) would be obviously accepted as true EBMT (e.g. Somers et al. 1994). Generalizations of strings in the form of sequences of words with variables (e.g. templates such as "I do not care for the X", "X gave the Y his particulars", "Do you want a room costing X dollars?") are seen as reasonable and natural developments designed to improve the recall of suitable examples from the database.

What is 'problematic' in EBMT (as far as defining the framework is concerned) is the analysis of sentences (clauses) as dependency and phrase structure tree representations, whether applied just to input sentences or also to example sentences in the database, or to both (e.g. Watanabe et al. 2003, Menezes and Richardson 2001). It would seem to be acceptable that systems are included within the EBMT framework if parsing is restricted to only one side of SL-TL correspondences, e.g. only to SL sentences in the database or only to their corresponding TL sentences, and if otherwise the system deals with 'surface' strings (with variables).

However, if *all* the processes of a system (pre-processing, input decomposition, matching, extraction, recombination) are based on parses as dependency trees and on comparisons of sub-trees, then what is the difference from tree transduction processes in RBMT systems (e.g. in the Eurotra architecture)? Although these systems stand at the edge of the EBMT spectrum – i.e. by taking generalisation of examples to the extreme – they are still not categorizable as being in effect RBMT systems. The reason is that the processes of tree transduction in these types of EBMT systems are based on comparisons and selections of tree (and subtree) representations which are comprised of lexical items and which are derived from bilingual corpora of SL and TL example sentences. That is to say that the 'transfer' processes are example based because they are performed *with reference to* databases of paired SL-TL sentences and phrases.

By contrast RBMT trees comprise both lexical items and grammatical categories (N, NP, PP, etc.) and trees are converted by rules operating on both lexical and grammatical nodes of trees (and subtrees). In RBMT systems tree transduction is based on rules applied to abstract representations consisting of categories as well as lexical items. RBMT systems may derive (all or some of) their rules from bilingual databases – whether manually or (semi)automatically – but the use of such resources does not make them EBMT systems.

Consequently, even though the processes of decomposition and recombination in such types of EBMT systems are identical to the processes of analysis and synthesis in RBMT systems, there remains a clear dividing line in principle with respect to the core process of 'transfer' – rule-based versus example-based. However, it can be argued that "there is no essential difference between translation examples and translation rules... they can be handled in a uniform way; that is, a translation example is a special case of translation rules, whose nodes are lexical entries rather than categories" (Maruyama and Watanabe 1992: 183). In this view, those EBMT systems with RBMT-like representations and RBMT-like tree processing appear to be variants of traditional RBMT. The uncertainty remains, and perhaps it would be better to refer to such systems as 'hybrids' of EBMT and RBMT.

8 EBMT and SMT

Initially, differences between SMT and EBMT were distinct: SMT input was decomposed into individual SL words and TL words were extracted by frequency data (in the 'translation model'), while in EBMT input was decomposed into SL fragments and TL examples (in the form of corresponding fragments) were extracted from the database. More recent developments of 'phrase-based' and 'syntax-based' SMT models have blurred these distinctions.

In phrase-based and syntax-based SMT systems parsing (i.e. statistical parsing) is performed for a variety of reasons: to improve alignments (e.g. Watanabe et al. 2002), or to facilitate the matching of input strings (rather than just individual words, e.g. Koen and Knight 2003), or to allow for the analysis of input sentences as phrase structures (e.g. Charniak et al. 2003) and matching against parsed sentences in the database.[8] There is thus a similar divergence as in EBMT between systems where parsing is part of the pre-processing stage and where it is (also) part of the analysis (decomposition) and matching stages. However, the SMT systems retain the distinctive use of 'translation models' and 'language models', and most processes remain word- and string-based.

This use in SMT of models based (partially or wholly) on dependency trees rather than surface strings represents a 'convergence' towards those EBMT systems which also operate with parsed representations. As far as phrase-based SMT and EBMT are concerned, it seems that both may be regarded as variants of a single framework. The only residual differences are that while SMT works mainly on the basis of statistical methods, EBMT works mainly on the basis of linguistic (symbolic) fragments and text examples.

9 Conclusion

The need for a definition of EBMT is motivated by the confusing variety of techniques which have been discussed as 'example-based' and the difficulty of locating the essential 'architecture' of EBMT from the great variety of descriptions of EBMT systems. As the last two sections demonstrate also there are some cases where EBMT approaches appear to differ little from those of RBMT and SMT approaches. The attempt to define EBMT is to provide researchers and observers with an 'archetype' (comparable to definitions of RBMT systems which distinguished transfer-based and interlingua-based systems, while in practice few operational systems conformed to the archetype in all details.)

Underlying the definition of EBMT attempted here is that the characteristic feature of EBMT remains the assumption (or hypothesis) that translation involves the finding of 'analogues' (similar in meaning and form) of SL sentences in existing TL texts. By contrast, neither SMT nor RBMT work with analogues: SMT uses statistically established word and phrase correspondences, and RBMT works with representations (of sentences, clauses, words, etc.) of 'equivalent' meanings. Since EBMT occupies an intermediary position between RBMT and SMT and it makes use of both statistical (SMT-like) and symbolic or linguistic (RBMT-like) methods, it is open to a wider variety of methodologies, and it is consequently less easy to characterise and define.

10 Acknowledgements

My thanks to the anonymous reviewers who contributed to improvements in this paper and to Michael Carl and Andy Way who encouraged me to write it.

[8] For a general model for parsing aligned bilingual texts see the work of Dekai Wu (e.g. Wu 2000, and references therein).

11 References

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. M. 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, Pennsylvania: Carnegie-Mellon University.

J. Carbonell, K. Probst, E. Peterson, C. Monson, A. Lavie, R. Brown, and L. Levin. 2002. Automatic rule learning for resource-limited MT. In: S. D. Richardson (ed.) *Machine translation: from research to real users, 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002*, Tiburon, CA, USA, October 2002 (Berlin: Springer), 1-10.

M. Carl, and A. Way, eds. 2003. *Recent advances in example-based machine translation.* Dordrecht: Kluwer Academic Publishers.

E. Charniak, 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; 40-46.

N. Gough and A. Way. 2004. Robust large-scale EBMT with marker-based segmentation. *TMI-2004: proceedings of the Tenth International Conference on Theoretical and Methodological Issues in Machine Translation,* October 4-6, 2004, Baltimore, Maryland, USA, 95-104.

P. Koen, and K. Knight. 2003. Feature-rich statistical translation of noun phrases. *ACL-2003*: 41st Annual Meeting of the Association for Computational Linguistics, July 7-12, 2003, Sapporo, Japan

T. S. Kuhn. 1962. *The structure of scientific revolutions*. Chicago: University of Chicago. (International Encyclopedia of Unified Science, vol.2, no.2)

B. Lavoie, M. White, and T. Korelsky. 2001. Inducing lexico-structural transfer rules from parsed bi-texts. *ACL-EACL 2001 workshop "Data-driven Machine Translation",* July 7, 2001, Toulouse, France, 17-24.

K. McTait. 2003. Translation patterns, linguistic knowledge and complexity in an approach to EBMT. In: Carl and Way (2003), 307-338.

H. Maruyama and H. Watanabe. 1992. Tree cover search algorithm for example-based translation. *Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI-90). Proceedings,* 11-13 June 1990, Austin, Texas, 173-184.

A. Menezes and S. D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. *ACL-EACL 2001 workshop "Data-driven Machine Translation",* July 7, 2001, Toulouse, France, 39-46. Repr. in: Carl and Way (2003), 421-442.

M. Nagao. 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), 173-180.

S. Sato and M. Nagao. 1990. Towards memory-based translation. In: H. Karlgren (ed.) *Coling-90: papers presented to the 13th International Conference on Computational Linguistics*, Helsinki, August 1990; vol.3, 247-252.

H. Somers. 1999. Review article: example-based machine translation. *Machine Translation* 14(2), 113-157. Revised as: An overview of EBMT. In Carl and Way (2003), 3-57.

H. Somers, I. McLean and D. Jones. 1994. Experiments in multilingual example-based generation. *CSNLP 1994: 3rd Conference on the Cognitive Science of Natural Language Processing*, Dublin City University, 6-8 July 1994.

E. Sumita, H. Iida, and H. Kohyama. 1990. Translating with examples: a new approach to machine translation. *Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages (TMI-90). Proceedings,* 11-13 June 1990, Austin, Texas, 203-212.

E. Sumita. 2003. An example-based machine translation system using DP-matching between word sequences. In Carl and Way (2003), 189-209.

D. Turcato and F. Popowich. 2003. What is example-based machine translation? In: Carl and Way (2003), 59-81.

H. Watanabe, S. Kurohashi, and E. Aramaki. 2003. Finding translation patterns from paired source and target dependency structures. In Carl and Way (2003), 397-420.

T. Watanabe and E. Sumita. 2002. Statistical machine translation based on hierarchical phrase alignment. *Ninth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002),* March 13-17, 2002, Keihanna, Japan; 188-198.

D. Wu. 2000. Bracketing and aligning words and constituents in parallel text using stochastic inversion transduction grammars. In: J. Véronis (ed.) *Parallel text processing: alignment and use of translation corpora* (Dordrecht: Kluwer Academic Publishers), 139-167.