"New paradigms" in MT: the state of play now that the dust has settled

Harold L. SOMERS
Centre for Computational Linguistics
UMIST, PO Box 88,
Manchester M60 1QD, England
harold@ccl.umist.ac.uk

1. Background

In 1988, at the Second TMI conference at Carnegie Mellon University, IBM's Peter Brown shocked the audience by presenting an approach to Machine Translation (MT) which was quite unlike anything that most of the audience had ever seen or even dreamed of before. IBM's "purely statistical" approach, inspired by successes in speech processing, and characterised by the infamous statement "Every time I fire a linguist, my system's performance improves" flew in the face of all the received wisdom about how to do MT at that time, eschewing the rationalist linguistic approach in favour of an empirical corpus-based one.

There followed something of a flood of "new" approaches to MT, few as overtly statistical as the IBM approach, but all having in common the use of a corpus of translation examples rather than linguistic rules as a significant component. This apparent difference was often seen as a confrontation, especially for example at the 1992 TMI conference in Montreal, which had the explicit theme "Empiricist vs. Rationalist Methods in MT" (Isabelle, 1992), though already by that date most researchers were developing hybrid solutions using both corpus-based and theory-based techniques.

The heat has largely evaporated from the debate, so that now the "new" approaches are considered mainstream, in contrast though not in conflict with the older rule-based approaches.

In this paper, we will review the achievements of a range of approaches to corpus-based MT which we will consider variants of "example-based MT" (EBMT), although individual authors have used alternative names, perhaps wanting to bring out some key difference that distinguishes their own approach: "analogy-based", "memory-based", and "case-based" are all terms that have been used. These approaches all have in common the use of a corpus or database of already translated examples, and involve a process of matching a new input against this database to extract suitable examples which are then recombined in an analogical manner to determine the correct translation.

Two variants of the corpus-based approach stand somewhat apart from the scenario suggested here. One, which we will not discuss at all in this paper, is the Connectionist or Neural network approach. So far, only a little work with not very promising results has been done in this area (see Waibel et al., 1991; McLean, 1992; Castaño et al. 1997; Koncar & Guthrie, 1997).

The other major "new paradigm" is the purely statistical approach already mentioned, and usually identified with the IBM group's Candide system (Brown et al. 1990, 1993), though the approach has also been taken up by a number of other researchers. 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; more significant is that the important issues for the statistical approach are somewhat different, focusing, as one might expect, on the mathematical aspects of estimation of statistical parameters for the language models. Nevertheless, we will try to include these approaches in our overview.

2. EBMT and Translation Memory

EBMT is sometimes confused with the related technique of "Translation Memory" (TM). This problem is exacerbated by the fact that the two gained wide publicity at roughly the same time, and also by the (thankfully short-lived) use of the term "memory-based translation" as a synonym for EBMT. Although they have in common the idea of reuse of examples of already existing translations, they differ in that TM is an

interactive tool for the human translator, while EBMT is an essentially automatic translation technique or methodology. They share the common problems of storing and accessing a large corpus of examples, and of matching an input phrase or sentence against this corpus; but having located a (set of) relevant example(s), the TM leaves it to the human to decide what, if anything, to do next, whereas for EBMT the hard work has only just begun!

One other thing that EBMT and TM have in common is the long period of time which elapsed between the first mention of the underlying idea and the development of systems exploiting the ideas. It is interesting, briefly, to consider this historical perspective. The original idea for TM is usually attributed to Martin Kay's well-known "Proper Place" paper (1980), although the details are only hinted at obliquely:

> ... the translator might start by issuing a command causing the system to display anything in the store that might be relevant to [the text to be translated].... Before going on, he can examine past and future fragments of text that contain similar material. (Kay, 1980:19)

Interestingly, Kay was pessimistic about any of his ideas for what he called a "Translator's Amanuensis" ever actually being implemented. But Kay's observations are predated by the suggestion by Peter Arthern (1978) that translators can benefit from on-line access to similar, already translated documents, and in a follow-up article, Arthern's proposals quite clearly describe what we now call TMs:

> It must in fact be possible to produce a programme [*sic*] which would enable the word processor to 'remember' whether any part of a new text typed into it had already been translated, and to fetch this part, together with the translation which had already been translated, ...
>
> Any new text would be typed into a word processing station, and as it was being typed, the system would check this text against the earlier texts stored in its memory, together with its translation into all the other official languages [of the European Community]. ...
>
> One advantage over machine translation proper would be that all the passages so retrieved would be grammatically correct. In effect, we should be operating an electronic 'cut and stick' process which would, according to my calculations, save at least 15 per cent of the time which translators now employ in effectively producing translations. (Arthern, 1981:318).

Alan Melby (1995:225f) suggests that the idea might have originated with his group at Brigham Young University (BYU) in the 1970s. What is certain is that the idea was incorporated, in a very limited way, from about 1981 in ALPS, one of the first commercially available MT systems, developed by personnel from BYU. This tool was called "Repetitions Processing", and was limited to finding exact matches and *modulo* alphanumeric strings. The much more inventive name of "translation memory" does not seem to have come into use until much later.

The idea for EBMT dates from about the same time, though the paper presented by Makoto Nagao at a 1981 conference was not published until three years later (Nagao, 1984). The essence of EBMT, called "machine translation by example-guided inference, or machine translation by the analogy principle" by Nagao, is succinctly captured by his much quoted statement:

> Man does not translate a simple sentence 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. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference. (Nagao, 1984:178f)

Nagao correctly identified the three main components of EBMT: matching fragments against a database of real examples, identifying the corresponding translation fragments, and then recombining these to give the target text. Clearly EBMT involves two important and difficult steps beyond the matching task which it shares with TM.

Mention should also be made at this point of the work of the DLT group in Utrecht, often ignored in discussions of EBMT, but dating from about the same time as (and probably without knowledge of) Nagao's work,. The matching technique suggested by Nagao involves measuring the semantic proximity of the words, using a thesaurus. A similar idea is found in DLT's "Linguistic Knowledge Bank" of example phrases described in Pappegaaij et al. (1986) and Schubert (1986).

3. Underlying problems

In this section we will review some of the general problems underlying example-based approaches to MT. Starting with the need for a database of examples, i.e. parallel corpora, we then discuss how to choose appropriate examples for the database, how they should be stored, various methods for matching new

inputs against this database, what to do with the examples once they have been selected, and finally, some general computational problems regarding speed and efficiency.

3.1 Parallel corpora

Since EBMT is corpus-based MT, the first thing that is needed is a parallel aligned corpus.[1] Machine-readable parallel corpora in this sense are quite easy to come by: EBMT systems are often felt to be best suited to a sublanguage approach, and an existing corpus of translations can often serve to define implicitly the sublanguage which the system can handle. Researchers may build up their own parallel corpus or may locate such corpora in the public domain. The Canadian and Hong Kong parliaments both provide huge bilingual corpora in the form of their parliamentary proceedings, the European Union is a good source of multilingual documents, while of course many World Wide Web pages are available in two or more languages. Not all these resources necessarily meet the sublanguage criterion, of course.

Once a suitable corpus has been located, there remains the problem of aligning it, i.e. identifying at a finer granularity which segments (typically sentences) correspond to each other. There is a rapidly growing literature on this problem (Fung & McKeown, 1997, includes a reasonable overview and bibliography; see also Somers, 1998) which can range from relatively straightforward for "well behaved" parallel corpora, to quite difficult, especially for typologically different languages and/or those which do not share the same writing system.

The alignment problem can of course be circumvented by building the example database manually, as is sometimes done for TMs, when sentences and their translations are added to the memory as they are typed in by the translator.

[1] By 'parallel' we mean a text together with its translation. By 'aligned', we mean that the two texts have been analysed into corresponding segments; the size of these segments may vary, but typically corresponds to sentences. It is of interest to note that for some corpus linguists, the term 'translation corpus' is used to indicate that the texts are mutual translations, while 'parallel corpus' refers to any collection of multilingual texts of a similar genre. Other researchers prefer the term 'comparable corpus' (cf. McEnery & Wilson, 1996:60n).

3.2 Suitability of examples

The assumption that an aligned parallel corpus can serve as an example database is not universally made. Several EBMT systems work from a manually constructed database of examples, or from a carefully filtered set of "real" examples.

There are several reasons for this. A large corpus of naturally occurring text will contain overlapping examples of two sorts: some examples will mutually reinforce each other, either by being identical, or by exemplifying the same translation phenomenon. But other examples will be in conflict: the same or similar phrase in one language may have two different translations for no other reason than inconsistency.

Where the examples reinforce each other, this may or may not be useful. Some systems involve a similarity metric (see below) which is sensitive to frequency, so that a large number of similar examples will increase the score given to certain matches. But if no such weighting is used, then multiple similar or identical examples are just extra baggage, and in the worst case may present the system with a choice — a kind of "ambiguity" — which is simply not relevant: in such systems, the examples can be seen as surrogate "rules", so that, just as in a traditional rule-based MT system, having multiple examples (rules) covering the same pheno-menon leads to over-generation.

Nomiyama (1992) introduces the notion of "exceptional examples", while Watanabe (1994) goes further in proposing an algorithm for identifying examples such as the sentences in (1) and (2).[2]

(1) a. *Watashi wa kompyuutaa o kyooyoosuru.*
I (subj) COMPUTER (obj) SHARE-USE.
I share the use of a computer.
b. *Watashi wa kuruma o tsukau.*
I (subj) CAR (obj) USE.
I use a car.

(2) *Watashi wa dentaku o shiyoosuru.*
I (subj) CALCULATOR (obj) USE.
a. I share the use of a calculator.
b. I use a calculator.

Given the input in (2), the system might choose (2a) as the translation because of the closer similarity of 'calculator' to 'computer' than to 'car' (the three words for 'use' being considered synonyms). So (1a) is an exceptional example because it introduces the

[2] I have adapted Watanabe's transcription, and corrected an obvious misprint in (2a).

unrepresentative element of 'share'. The situation can be rectified by removing example (1a) and/or by supple-menting it with an unexceptional example.

Distinguishing exceptional and general examples is one of a number of means by which the example-based approach is made to behave more like the traditional rule-based approach. Although it means that "example interference" can be minimised, EBMT purists might object that this undermines the empirical nature of the example-based method.

3.3 How are examples stored?

EBMT systems differ quite widely in how the translation examples themselves are actually stored. Obviously, the storage issue is closely related to the problem of searching for matches.

In the simplest case, the examples may be stored as pairs of strings, with no additional information associated with them. Sometimes, indexing techniques borrowed from Information Retrieval (IR) can be used: this is often necessary where the example database is very large, but there is an added advantage that it may be possible to make use of a wider context in judging the suitability of an example. Imagine, for instance, an example-based dialogue translation system, wishing to translate the simple utterance *OK*. The Japanese translation for this might be *wakarimashita* 'I understand', *iidesu yo* 'I agree', or *ijoo desu* 'let's change the subject', depending on the context.[3] It may be necessary to consider the immediately preceding utterance both in the input and in the example database. So the system could broaden the context of its search until it found enough evidence to make the decision about the correct translation.

Of course if this kind of information was expected to be relevant on a regular basis, the examples might actually be stored with some kind of contextual marker already attached. This was the approach taken in the proposed MEG system (Somers & Jones, 1992).

Early attempts at EBMT — where the technique was often integrated into a more conventional rule-based system — stored the examples as fully annotated tree structures with explicit links (e.g. Sato & Nagao, 1990). Figure 1 (from Watanabe, 1992) shows how the Japanese example in (3) and its English translation is represented.

(3) *Kanojo wa kami ga nagai.*
SHE (topic) HAIR (subj) IS-LONG.
She has long hair.

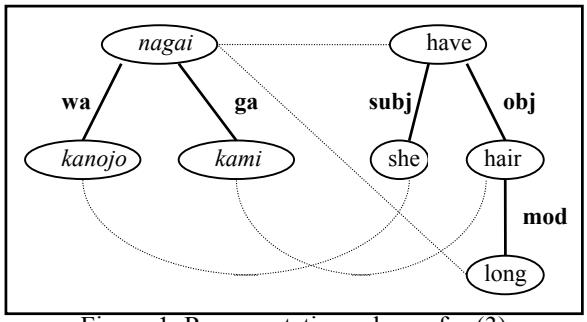

Figure 1. Representation scheme for (3).
(Watanabe, 1992:771)

More recent systems have adapted this somewhat inefficient approach, and annotate the examples more superficially. In Jones (1996) the examples are POS-tagged, carry a Functional Grammar predicate frame and an indication of the sample's rhetorical function. In Collins & Cunningham's (1995) system, the examples are tagged, and carry information about syntactic function. Somers et al. (1994) use only tags.

In Furuse & Iida's (1992) proposal, examples are stored in one of three ways: (a) literal examples, (b) "pattern examples" with variables instead of words, and (c) "grammar examples" expressed as context-sensitive rewrite rules, using semantic features. Each type is exemplified in (4–6), respectively.

(4) *Sochira ni okeru* ⇒ We will send it to you
Sochira wa jimukyoku desu ⇒ This is the office

(5) *X o onegai shimasu* ⇒ may I speak to X′
(X = *jimukyoku* 'office')
X o onegai shimasu ⇒ please give me the X′
(X = *bangoo* 'number)

(6) *N1 N2 N3* ⇒ N3′ of N1′ (N1 = *kaigi* 'meeting', N2 = *kaisai* 'opening', N3 = *kikan* 'time')
N1 N2 N3 ⇒ N2′ N3′ for N1′ (N1 = *sanka* 'participation', N2 = *mooshikomi* 'application', N3 = *kyooshi* 'form')

We will see how these different example types are matched in the next section; but what is clear is the hybrid nature of this approach, where the type (a) examples are pure strings, type (c) are effectively "transfer rules" of the traditional kind, with type (b) half-way between the two.

At this point we might also mention the way examples are "stored" in the statistical approaches. In fact, in these systems, the examples are not stored at all, except inasmuch as they occur in the corpus on which the

[3] Examples are from Somers et al. (1990:274).

system is based. What *is* stored is the precomputed statistical parameters which give the proba-bilities for bilingual word pairings, the "translation model". The "language model" which gives the probabilites of target word strings being well-formed is also precomputed, and the translation process consists of a search for the target-language string which optimises the product of the two sets of probabilities, given the source-language string.

3.4 Matching

The first task in an EBMT system is to take the source-language string to be translated and to find the example (or set of examples) which most closely match it. This is also the essential task facing a TM system too. This search problem depends of course on the way the examples are stored. In the case of the statistical approach, the problem is the essentially mathematical one of maximising a huge number of statistical probabilites. In more conventional EBMT systems the matching process may be more or less linguistically motivated.

All matching processes necessarily involve a distance or similarity measure. In the most simple case, where the examples are stored as strings, the measure may be a traditional character-based pattern-matching one. In the earliest TM systems as mentioned above (ALPS' "Repetitions Processing", cf. Weaver, 1988), only exact matches, *modulo* alphanumeric strings, were possible: (7a) would be matched with (7b), but the match in (8) would be missed.

(7) a. This is shown as A in the diagram.
b. This is shown as B in the diagram.

(8) a. The large paper tray holds up to 400 sheets of A3 paper.
b. The small paper tray holds up to 400 sheets of A4 paper.

In the case of Japanese–English translation, which many EBMT systems focus on, the notion of character-matching can be modified to take account of the fact that certain "characters" (in the orthographic sense: each Japanese character is represented by two bytes) are more discriminatory than others (e.g. Sato, 1991). This introduces a simple linguistic dimension to the matching process, and is akin to the well-know device in IR, where only keywords are considered.

Perhaps the "classical" similarity measure, suggested by Nagao (1984) and used in many systems, is the use of a thesaurus. Here, matches are permitted when words in the input string are replaced by near synonyms (as measured by relative distance in a hierarchically structured vocabulary) in the example sentences. This measure is particularly effective in choosing between competing examples, as in Nagao's examples, where, given (9a,b) as models, we choose the correct translation of *eat* in (10a,b) as *taberu* 'eat (food)' or *okasu* 'erode', on the basis of the relative distance from *he* to *man* and *acid*, and from *potatoes* to *vegetables* and *metal*.

(9) a. A man eats vegetables. *Hito wa yasai o taberu*.
b. Acid eats metal. *San wa kinzoku o okasu*.

(10) a. He eats potatoes. *Kare wa jagaimo o taberu*.
b. Sulphuric acid eats iron. *Ryuusan wa tetsu o okasu*.

In a little-known research report, Carroll (1990) suggests a trigonometric similarity measure based on both the relative length and relative contents of the strings to be matched: the relevance of particular mismatches is reflected as a "penalty", and the measure can be adjusted to take account of linguistic general-isations, e.g. a missing comma may incur a lesser penalty than a missing adjective or noun. Carroll hoped that his system would be able, given (11) as input, to offer both (12a,b) as suitable matches.

(11) When the paper tray is empty, remove it and refill it with appropriate size paper.

(12) a. When the bulb remains unlit, remove it and replace it with a new bulb.
b. If the tray is empty, refill it with paper.

The availability to the similarity measure of information about syntactic classes implies some sort of analysis of both the input and the examples. Cranias et al. (1994) describe a measure that takes function words into account, and makes use of POS tags. Furuse & Iida's (1994) "constituent boundary parsing" idea is not dissimilar. Veale & Way (1997) users sets of closed-class words to segment the examples.

Earlier proposals for EBMT, and proposals where EBMT is integrated within a more traditional approach, assumed that the examples would be stored as tree structures, so the process involves a rather more complex tree-matching (e.g. Watanabe, 1995; Matsumoto et al. 1993).

In the multi-engine Pangloss system (see below), the matching process successively "relaxes" its requirements, until a match is found (Nirenburg et al., 1993, 1994): the process begins by looking for exact matches,

then allows some deletions or insertions, then word-order differences, then morphological variants, and finally POS-tag differences, each relaxation incurring an ever-increasing penalty.

3.5 Adaptability and recombination

Having matched and retrieved a set of examples, with associated translations, the next step is to extract from the translations the appropriate fragments, and to combine these so as to produce a grammatical target output. This is arguably the most difficult step in the EBMT process: its difficulty can be gauged by imagining a source-language monolingual trying to use a TM system to compose a target text. The problem is twofold: (a) identifying which portion of the associated translation corresponds to the matched portions of the source text, and (b) recombining these portions in an appropriate manner. Compared to the other issues in EBMT, this one has received considerably less attention.

Sato's approach, as detailed in his 1995 paper, takes advantage of the fact that the examples are stored as tree structures, with the correspondences between the fragments explicitly labelled. So problem (a) effectively disappears. The recombination stage is a kind of tree unification, familiar in computational linguistics. Watanabe (1992, 1995) adapts a process called "gluing" from Graph Grammars.

The problem is further eased, in the case of languages like Japanese and English, by the fact that there is little or no grammatical inflection to indicate syntactic function. So for example the translation associated with *the handsome boy* extracted, say, from (13), is equally reusable in either of the sentences in (14). This however is not the case for a language like German (and of course many others), where the form of the determiner, adjective and noun can all carry inflections to indicate grammatical case, as in (15).

(13) The handsome boy entered the room.
(14) a. The handsome boy ate his breakfast.
b. I saw the handsome boy.
(15) a. *Der schöne Junge aß seinen Frühstück.*
b. *Ich sah den schönen Jungen.*

Collins & Cunningham (1997) stress this question of whether all examples are equally reusable with their notion of "adaptability". Their example retrieval process includes a measure of adaptability which indicates the similarity of the example not only in its internal structure, but also in its external context.

In Somers et al. (1994), the recombination process considers the left and right context of each fragment in the original corpus, which gives as "hooks" tagged words which can then be compared with the context into which the system proposes to fit the fragment. As a further measure, the system attempts to compare the target texts composed by the recombination process with the target-language side of the original corpus, reusing the matching algorithm as if the proposed output were in fact an input to be translated: the ease with which the generated text can be matched against the corpus is a measure of the verisimilitude of the constructed sentence.

One other approach to recombination is that taken in the purely statistical system: like the matching problem, recombination is expressed as a statistical modelling problem, the para-meters having been precomputed. This time, it is the "language model" that is invoked, with which the system tries to maximise the product of the word-sequence probabilities.

3.6 Computational problems

All the approaches mentioned so far of course have to be implemented as computer programs, and significant computational factors influence many of them. One criticism to be made of the approaches such as Sato & Nagao (1990), Watanabe (1992) and even Jones (1996), which store the examples as fully annotated structures, is the huge computational cost in terms of creation, storage and complex retrieval algorithms. Sumita & Iida (1995) is one of the few papers to address this issue explicitly, turning to parallel processing for help, a solution also adopted by Kitano (1994) and Sato (1995).

One important computational issue is speed, especially for those of the EBMT systems that are used for real-time speech translation. The size of the example database will obviously affect this and it is thus understandable that some researchers are looking at ways of maximising the effect of the examples by identifying and making explicit significant generalisations. In this way the hybrid system has emerged, assuming the advantages of both the example-based and rule-based approaches.

4. Flavours of EBMT

So far we have looked at various solutions to the individual problems which make up EBMT. In this section, we prefer to take a wider view, to consider the various different contexts in which EBMT has been proposed. In many cases, EBMT is used as a component in an MT system which also has more traditional elements: EBMT may be used in parallel with these other "engines", or just for certain classes of problems, or when some other component cannot deliver a result. Also, EBMT methods may be better suited to some kinds of applications than others. And finally, it may not be obvious any more what exactly is the dividing line between EBMT and so-called "traditional" rule-based approaches. As the title and first paragraph of this paper suggest, EBMT was once seen as a bitter rival to the existing paradigm, but there now seems to be a much more comfortable coexistence.

4.1 Suitable translation problems

Let us consider first the range of translation problems for which EBMT is best suited. Certainly, EBMT is closely allied to *sublanguage* translation, not least because of EBMT's reliance on a real corpus of real examples: at least implicitly, a corpus can go a long way towards defining a sublanguage. On the other hand, nearly all research nowadays in MT is focused on a specific domain or task, so perhaps all MT is sublanguage MT.

More significant is that EBMT is often proposed as an antidote to the problem of "structure-preserving translation as first choice" (cf. Somers, 1987:84) inherent in MT systems which proceed on the basis of structural analysis. Because many EBMT systems do not compute structure, the source-language structure cannot by definition be imposed on the target language. Indeed, some of the early systems in which EBMT is integrated into a more traditional approach explicitly use EBMT for such cases:

> When one of the following conditions holds true for a linguistic phenomenon, RBMT [rule-based MT] is less suitable than EBMT.
> (a) Translation rule formation is difficult.
> (b) The general rule cannot accurately describe [the] phenomen[on] because it represents a special case.
> (c) Translation cannot be made in a compositional way from target words. (Sumita & Iida, 1991:186)

4.2 Pure EBMT

Very few research efforts have taken an explicitly "purist" approach to EBMT. One exception is our own effort (Somers et al., 1994), where we wanted to push to the limits a "purely non-symbolic approach" in the face of, we felt, a premature acceptance that hybrids were the best solution. Not incorporating any linguistic information that could not be derived automatically from the corpus became a kind of dogma.

The other non-linguistic approach is of course the purely statistical one of Brown et al. (1990, 1993). In fact, their aspirations were much less dogmatic, and in the face of mediocre results, they were soon resorting to linguistic knowledge (Brown et al., 1992); not long afterwards the group broke up, though other groups have taken up the mantle of statistics-based MT (Vogel et al., 1986; Wang & Waibel, 1997).

Other approaches, as we have seen above, while remaining more or less true to the case-based (rather than theory-based) approach of EBMT, accept the necessity to incorporate linguistic knowledge either in the representation of the examples, and/or in the matching and recombination processes. This represents one kind of hybridity of approach; but in this section we will look at hybrids in another dimension, where the EBMT approach is integrated into a more conventional system.

4.3 EBMT for special cases

As the quotation from Sumita & Iida above shows, one of the first uses envisaged for the EBMT approach was where the rule-based approach was too difficult. The classical case of this, as described in one of the earliest EBMT papers (Sumita et al., 1990; Sumita & Iida, 1991), was the translation of Japanese adnominal particle constructions (*A no B*), where the default or structure-preserving translation (*B of A*) is wrong 80% of the time. In Sumita & Iida's traditional rule-based system, the EBMT module was invoked just for this kind of example (and a number of other similarly difficult cases). In a similar way, Katoh & Aizawa (1994) describe how only "parameterizable fixed phrases" in economics news stories are translated on the basis of examples, in a way very reminiscent of TM systems.

4.4 Example-based transfer

Because their examples are stored as tree structures, one can describe the systems of Sato & Nagao (1990) and Sato (1991) as "example-based transfer": source-language

inputs are analysed into dependency representations in a conventional manner, only transfer is on the basis of examples rather than rules, and then generation of the target-language output is again done in a traditional way.

4.5 Deriving transfer rules from examples

Some researchers take this scenario a step further, using EBMT as a research technique to build the rule base rather than a translation technique *per se*. We can see this in the case of Furuse & Iida's (1992) distinction of three types of "example" (4)–(6) in Section 3.3. above: they refer to "string-level", "pattern-level" and "grammar-level" transfer knowledge, and it seems that the more abstract representations are *derived* from examples by a process of generalisation.

Kaji et al. (1992) describe their "two phase" EBMT methodology, the first phase involving "learning" of templates (i.e. transfer rules) from a corpus. Each template is a "bilingual pair of pseudo sentences", i.e. example sentences containing variables. The translation templates are generated from the corpus first by parsing the translation pairs and then aligning the syntactic units with the help of a bilingual dictionary, resulting in a translation template as in Figure 2. This can then be generalised by replacing the coupled units with variables marked for syntactic category, also shown in Figure 2.

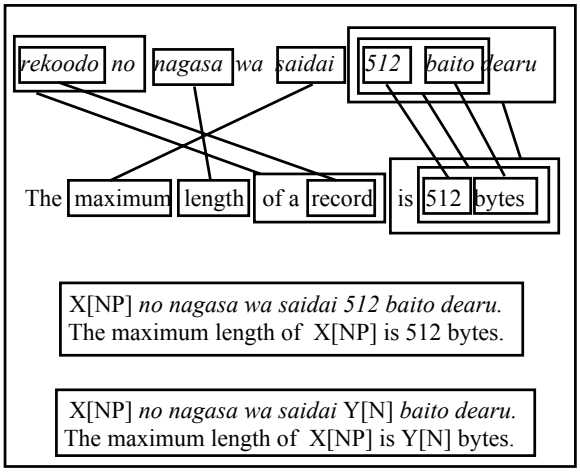

Figure 2. Generation of translation templates from aligned example. (Kaji et al., 1992:673)

Kaji et al. do not make explicit the criteria for choosing the units, though they do discuss the need to refine templates which give rise to a conflict, as in (16–17).

(16) a. play baseball → *yakyu o suru*
b. play tennis → *tenisu o suru*
c. play X[NP] → X[NP] *o suru*
(17) a. play the piano → *piano o hiku*
b. play the violin → *baiorin o hiku*
c. play X[NP] → X[NP] *o hiku*

Nomiyama (1992) similarly describes how examples ("cases") can be generalised into rules by combining them when similar segments occur in similar environments, this similarity being based on semantic proximity as given by a hierarchical thesaurus.

Almuallim et al. (1994) and Akiba et al. (1995) report much the same idea, though they are more formal in their description of how the process is implemented, citing the use of two algorithms from Machine Learning. Interestingly, these authors make no claim that their system is therefore "example-based". Also, may of the examples that they use to induce the transfer rules are artificially constructed.

To end this section we could mention briefly the huge amount of work that has been done in the area of extracting linguistic knowledge from corpora for various purposes, including MT. The literature on this topic is vast, including much of the parallel corpus alignment literature, where vocabulary extraction is one of the major goals (se, e.g. Somers, 1998). Some researchers have also addressed various aspects of grammatical knowledge acquisition from corpora.

4.6 EBMT as one of a multi-engine system

One other scenario for EBMT is exemplified by the Pangloss system, where EBMT operates in parallel with two other techniques: knowledge-based MT and a simpler lexical transfer engine (Frederking & Nirenburg, 1994; Frederking et al. 1994). Nirenberg et al. (1994) and Brown (1996) describe the EBMT aspect of this work in most detail. What is most interesting is the extent to which the different approaches often mutually confirm each other's proposed translations, and the comparative evidence that the multi-engine approach offers.

5. Conclusions

In this review article, we have seen a range of applications all of which might claim to "be" EBMT systems. So one outstanding question might be, What counts as EBMT? Certainly, the use of a bilingual corpus is part of the definition, but this is not sufficient. Almost all

research on MT nowadays makes use at least of a "reference" corpus to help to define the range of vocabulary and structures that the system will cover. It must be something more, then.

EBMT means that the main knowledge-base stems from examples. However, as we have seen, examples may be used as a device to shortcut the knowledge-acquisition bottleneck in rule-based MT, the aim being to generalise the examples as much as possible. So part of the criterion might be whether the examples are used at run-time or not: but by this measure, the statistical approach would be ruled out; although the examples are not used to derive rules in the traditional sense, still at run-time there is no consultation of the database of examples.

The original idea for EBMT seems to have been couched firmly in the rule-based paradigm: examples were to be stored as tree structures, so rules must be used to analyse them: only transfer was to be done on the basis of examples, and then only for special, difficult cases. After the comparative success of this approach, and also as a reaction to the apparent stagnation in research in the conventional paradigm, the idea grew that EBMT might be a "new" paradigm altogether, in competition with the old, even. As we have seen, and as the title of this paper suggests, this confrontational aspect has quickly died away, and in particular EBMT has been integrated into more traditional approaches (and *vice versa*, one could say) in many different ways.

We will end this article by mentioning, for the first time, some of the advantages that have been claimed for EBMT. Not all the advantages that were claimed in the early days of polemic are obviously true. But it seems that at least the following do hold, inasmuch as the system design is primarily example-based (e.g. the examples may be "generalised", but corpus data is still the main source of linguistic knowledge):

- Examples are real language data, so their use leads to systems which cover the constructions which really occur, and ignore the ones that don't, so over-generation is reduced.
- The linguistic knowledge of the system can be more easily enriched, simply by adding more examples.
- EBMT systems are data-driven, rather than theory-driven: since there are therefore no complex grammars devised by a team of individual linguists, the problem of rule conflict and the need to have an overview of the "theory", and how the rules interact, is lessened. (On the other hand, as we have seen, there is the opposite problem of conflicting examples.)
- The example-based approach seems to offer some relief from the constraints of "structure-preserving" translation.

EBMT is certainly here to stay, not as a rival to rule-based methods but as an alternative, available to enhance and, sometimes, replace it. Nor is research in the purely rule-based paradigm finished. As I mentioned in Somers (1997:116), the problem of scaling up remains, as do a large number of interesting translation problems, especially with new uses for MT (e.g. web-page and e-mail translation) emerge. The dust has settled, and the road ahead is all clear.

Acknowledgements

I am very grateful to Nick In 't Veen for useful discussions about EBMT.

References

Akiba, Y., M. Ishii, H. Almuallim and S. Kaneda. 1995. Learning English Verb Selection Rules from Hand-made Rules and Translation Examples. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, 206–220.

Almuallim, H., Y. Akiba, T. Yamazaki, A. Yokoo and S. Kaneda. 1994. Two methods for Learning ALT-J/E Translation Rules from Examples and a Semantic Hierarchy. *COLING 94, The 15th International Conference on Computational Linguistics*, Kyoto, 57–63.

Arthern, P. J. 1978. Machine Translation and Computerized Terminology Systems: A Translator's Viewpoint. In B.M. Snell (ed.) *Translating and the Computer: Proceedings of a Seminar, London, 14th November 1978*, pp. 77–108. Amsterdam (1979): North-Holland.

Arthern, P. J. 1981. Aids Unlimited: The Scope for Machine Aids in a Large Organization. *Aslib Proceedings* 33, 309–319.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer and P. S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16, 79–85.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, J. D. Lafferty and R. L. Mercer. 1992. Analysis, Statistical Transfer, and Synthesis in Machine Translation. In Isabelle (1992), 83–100.

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, 263–311.

Brown, R. D. 1996. Example-Based Machine Trans-lation in the Pangloss System. *COLING-96: The 16th International Conference on Computational Linguistics*, Copenhagen, 169–174.

Carroll, J. J. 1990. Repetitions Processing Using a Metric Space and the Angle of Similarity. Report No. 90/3, Centre for Computational Linguistics, UMIST, Manchester.

Castaño, M. A., F. Casacuberta and E. Vidal. 1997. Machine Translation using Neural Networks and Finite-State Models. *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*, Santa Fe, New Mexico, 160–167.

Collins, B. and P. Cunningham. 1995. A Methodology for Example Based Machine Translation. *CSNLP 1995: 4th Conference on the Cognitive Science of Natural Language Processing*, Dublin.

Collins, B. and P. Cunningham. 1997. Adaptation Guided Retrieval: Approaching EBMT with Caution. *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*, Santa Fe, New Mexico, 119–126.

Cranias, L., H. Papageorgiou and S. Piperidis. 1994. A Matching Technique in Example-based Machine Translation. *COLING 94, The 15th International Conference on Computational Linguistics*, Kyoto, 100–104.

Frederking, R. and S. Nirenburg. 1994. Three Heads are Better than One. *4th Conference on Applied Natural Language* Processing, Stuttgart. 95–100.

Frederking, R., S. Nirenburg, D. Farwell, S. Helmreich, E. Hovy, K. Knight, S. Beale, C. Domashnev, D. Attardo, D. Grannes and R. Brown. 1994. Integrating Translations from Multiple Sources within the Pangloss Mark III Machine Translation System. *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland. 73–80.

Fung, P. and K. McKeown. 1997. A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups. *Machine Translation* 12, 53–87.

Furuse, O. and H. Iida. 1992. An Example-Based Method for Transfer-Driven Machine Translation. In Isabelle (1992), 139–150.

Furuse, O. and H. Iida. 1994. Constituent Boundary Parsing for Example-based Machine Translation. *COLING 94, The 15th International Conference on Computational Linguistics*, Kyoto, 105–111.

Isabelle, P. (ed.) 1992. *Quatrième colloque international sur les aspects théoriques et méthodologiques de la traduction automatique; Fourth International Conference on Theoretical and Methodological Issues in Machine Translation. Méthodes empiristes versus méthodes rationalistes en TA; Empiricist vs. Rationalist Methods in MT. TMI-92*. Montréal: CCRIT-CWARC.

Jones, D. 1996. *Analogical Natural Language Processing.* London: UCL Press.

Kaji, H., Y. Kida and Y. Morimoto. 1992. Learning Translation Templates from Bilingual Text. *Proceedings of the fifteenth* [sic] *International Conference on Computational Linguistics, COLING-92*, Nantes, 672–678.

Katoh, N. and T. Aizawa. 1994. Machine Translation of Sentences with Fixed Expressions. *4th Con-ference on Applied Natural Language Processing*, Stuttgart. 28–33.

Kay, M. 1980. The Proper Place of Men and Machines in Language Translation. Research Report CSL-80-11, Xerox PARC, Palo Alto, Calif. Reprinted in *Machine Translation* 12 (1997), 3–23.

Kitano, H. 1994. *Speech-to-Speech Translation: A Massively Parallel Memory-Based Approach.* Boston: Kluwer.

Koncar, N. and G. Guthrie. 1997. A Natural-Language-Translation Neural Network. In D. Jones & H. Somers (eds) *New Methods in Language Processing*, 219–228. London: UCL Press.

Matsumoto, Y., H. Ishimoto and T. Utsuro. 1993. Structural Matching of Parallel Texts. *31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 23–30.

McEnery, T. and A. Wilson. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McLean, I. J. 1992. Example-based Machine Trans-lation Using Connectionist Matching. In Isabelle (1992), 35–43.

Melby, A. K. 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., S. Beale and C. Domashnev. 1994. A Full-Text Experiment in Example-Based Machine Translation. *International Conference on New Methods in Language Processing (NeMLaP),* Man-chester, 78–87.

Nirenburg, S., C. Domashnev and D.J. Grannes. 1993. Two Approaches to Matching in Example-Based Machine Translation. *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation: MT in the Next Generation*, Kyoto, 47–57.

Nomiyama, H. 1992. Machine Translation by Case Generalization. *Proceedings of the fifteenth* [sic] *International Conference on Computational Linguistics, COLING-92*, Nantes, 714–720.

Pappegaaij, B. C., V. Sadler and A. P. M. Witkam (eds) 1986. *Word Expert Semantics: An Inter-lingual Knowledge-Based Approach.* Dordrecht: Reidel.

Sato, S. 1991. Example-Based Translation Approach. *Proceedings of International Workshop on Fundamental Research for the Future Generation of Natural Language Processing (FGNLP)*, Kyoto, 1–16.

Sato, S. 1995. MBT2: A Method for Combining Fragments of Examples in Example-based Machine Translation. *Artificial Intelligence* 75, 31–49.

Sato, S. and M. Nagao. 1990. Toward Memory-based Translation. *COLING-90, Papers Presented to the 13th International Conference on Computational Linguistics*, Helsinki, Vol. 3, 247–252.

Schubert, K. 1986. Linguistic and Extra-Linguistic Knowledge: A Catalogue of Language-related Rules and their Computational Application in Machine Translation. *Computers and Translation* 1, 125–152.

Somers, H. 1987. Some Thoughts on Interface Structure(s). In W. Wilss and K.-D. Schmitz (eds) *Maschinelle Übersetzung — Methoden und Werk-zeuge,* 81–99. Tübingen: Niemeyer.

Somers, H. L. 1997. The Current State of Machine Translation. *MT Summit VI: Machine Translation Past Present Future*, San Diego, California. 115–124.

Somers, H. 1998. Further Experiments in Bilingual Text Alignment. *International Journal of Corpus Linguistics* 3, 1–36.

Somers, H. L. and D. Jones. 1992. Machine Translation Seen as Interactive Multilingual text Generation. *Translating and the Computer 13: The Theory and Practice of Machine Translation — A Marriage of Convenience?*, 153–165, London: Aslib.

Somers, H., 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.

Somers, H. L., J. Tsujii and D. Jones. 1990. Machine Translation without a Source Text. *COLING-90, Papers presented to the 13th International Con-ference on Computational Linguistics*, Helsinki, Vol. 3, 271–276.

Sumita, E. and H. Iida. 1991. Experiments and Prospects of Example-Based Machine Translation. *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, 185–192.

Sumita, E. and H. Iida. 1995. Heterogeneous Computing for Example-Based Translation of Spoken Language. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, 273–286.

Sumita, E., H. Iida and H. Kohyama. 1990. Translating with Examples: A New Approach to Machine Translation. *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of natural Language*, Austin, Texas. 203–212.

Veale, T. and A. Way. 1997. *Gaijin*: A Bootstrapping Approach to Example-Based Machine Translation. *International Conference, Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, 239–244.

Vogel, S., H. Ney and C. Tillmann. 1996. HMM-Based Word Alignment in Statistical Translation. *COLING-96: The 16th International Conference on Computational Linguistics*, Copenhagen, 836–841.

Waibel, A., A. N. Jain, A. E. McNair, H. Saito, A. G. Hauptmann and J. Tebelskis. 1991. JANUS: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies. *Proceedings of the 1991 IEEE International Conference on Acoustics, Speech, and Signal processing*, Toronto, 793–796.

Wang, Y-Y. and A. Waibel. 1997. Decoding Algorithm in Statistical Machine Translation. *35th Annual Meeting of the Association for Compu-tational Linguistics and 8th Conference of the European Chapter of the Association for Compu-tational Linguistics*, Madrid. 366–372.

Watanabe, H. 1992. A Similarity-Driven Transfer System. *Proceedings of the fifteenth* [sic] *International Conference on Computational Linguistics, COLING-92*, Nantes, 770–776.

Watanabe, H. 1994. A Method for Distinguishing Exceptional and General Examples in Example-based Transfer Systems. *COLING 94, The 15th International Conference on Computational Ling-uistics*, Kyoto, 39–44.

Watanabe, H. 1995. A Model of a Bi-Directional Transfer Mechanism Using Rule Combinations. *Machine Translation* 10, 269–291.

Weaver, A. 1988. Two Aspects of Interactive Machine Translation. In M. Vasconcellos (ed.) *Technology as Translation Strategy*, 116–123, State University of New York at Binghamton (SUNY).