Example-Based Machine Translation (EBMT)

SEMINAR REPORT
2009-2011

In partial fulfillment of Requirements in
Degree of Master of Technology
In
COMPUTER & INFORMATION SCIENCE

SUBMITTED BY

JESSY K

DEPARTMENT OF COMPUTER SCIENCE
COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY
KOCHI – 682 022
This is to certify that the seminar report entitled “Example-Based Machine Translation (EBMT)” is being submitted by Jessy K in partial fulfillment of the requirements for the award of M.Tech in Computer & Information Science is a bonafide record of the seminar presented by her during the academic year 2010.

Mr. G.Santhosh Kumar
Lecturer
Dept. of Computer Science

Prof. Dr.K.Poulse Jacob
Director
Dept. of Computer Science
ACKNOWLEDGEMENT

First of all let me thank our Director Prof: Dr. K. Poulose Jacob, Dept. of Computer Science, CUSAT who provided with the necessary facilities and advice. I am also thankful to Mr. G. Santhosh Kumar, Lecturer, Dept of Computer Science, CUSAT for his valuable suggestions and support for the completion of this seminar. With great pleasure I remember Dr. Sumam Mary Idicula, Reader, Dept. of Computer Science, CUSAT for her sincere guidance. Also I am thankful to all of my teaching and non-teaching staff in the department and my friends for extending their warm kindness and help.

I would like to thank my parents without their blessings and support I would not have been able to accomplish my goal. I also extend my thanks to all my well wishers. Finally, I thank the almighty for giving the guidance and blessings.
ABSTRACT

Example-based translation is essentially translation by analogy. An Example-Based Machine Translation (EBMT) system is given a set of sentences in the source language (from which one is translating) and their corresponding translations in the target language, and uses those examples to translate other, similar source-language sentences into the target language. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again.

At the foundation of example-based machine translation is the idea of translation by analogy. When applied to the process of human translation, the idea that translation takes place by analogy is a rejection of the idea that people translate sentences by doing deep linguistic analysis. Instead it is founded on the belief that people translate firstly by decomposing a sentence into certain phrases, then by translating these phrases, and finally by properly composing these fragments into one long sentence. Phrasal translations are translated by analogy to previous translations. The principle of translation by analogy is encoded to example-based machine translation through the example translations that are used to train such a system.

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”

Example-based machine translation systems are trained from bilingual parallel corpora, which contain sentence pairs like the example shown in the table. Sentence pairs contain sentences in one language with their translations into another. The particular example shows an example of a minimal pair, meaning that the sentences vary by just one element. These sentences make it simple to learn translations of subsentential units. For example, an example-based machine translation system would learn three units of translation:

**KeyWords:** Machine Translation, Adaptability, Alignment, Recombination, Character based, Structure based, Word based, matching
# CONTENTS

<table>
<thead>
<tr>
<th>Slno</th>
<th>Title</th>
<th>Page no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>History</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>EBMT Pyramid</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>EBMT Paradigm</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Underlying problems</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>5.1 Parallel Corpora</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>5.2 Granularity Of Examples</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>5.3 How Many Examples</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>5.4 Suitability Of Examples</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>5.5 How Are Examples Stored?</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>5.5.1 Annotated Tree Structures</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>5.5.2 Generalized Examples</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>5.5.3 Statistical Approaches</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>5.6 Matching</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>5.6.1 Character-based Matching</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>5.6.2 Word-based Matching</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>5.6.3 Structure-based Matching</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>5.7 Adaptability And Recombination</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>Applications of the Component Technologies</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>6.1 Speech Translation</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>6.2 Cross-Language Retrieval</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td>Conclusion</td>
<td>19</td>
</tr>
<tr>
<td>8</td>
<td>References</td>
<td>20</td>
</tr>
</tbody>
</table>
1. Introduction

Example-based translation is essentially translation by analogy. An Example-Based Machine Translation (EBMT) system is given a set of sentences in the source language (from which one is translating) and their corresponding translations in the target language, and uses those examples to translate other, similar source-language sentences into the target language. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again.

A restricted form of example-based translation is available commercially, known as a translation memory. In a translation memory, as the user translates text, the translations are added to a database, and when the same sentence occurs again, the previous translation is inserted into the translated document. This saves the user the effort of re-translating that sentence, and is particularly effective when translating a new revision of a previously-translated document (especially if the revision is fairly minor).

The Example-based machine translation (EBMT) approach to machine translation is often characterized by its use of a bilingual corpus with parallel texts as its main knowledge base, at run-time. It is essentially a translation by analogy and can be viewed as an implementation of case-based reasoning approach of machine learning.

At the foundation of example-based machine translation is the idea of translation by analogy. When applied to the process of human translation, the idea that translation takes place by analogy is a rejection of the idea that people translate sentences by doing deep linguistic analysis. Instead it is founded on the belief that people translate firstly by decomposing a sentence into certain phrases, then by translating these phrases, and finally by properly composing these fragments into one long sentence. Phrasal translations are translated by analogy to previous translations. The principle of translation by analogy is encoded to example-based machine translation through the example translations that are used to train such a system.
Example-Based Machine Translation

Example of bilingual corpus

<table>
<thead>
<tr>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>How much is that red umbrella?</td>
<td>Ano akai kasa wa ikura desu ka.</td>
</tr>
<tr>
<td>How much is that small camera?</td>
<td>Ano chiisai kamera wa ikura desu ka.</td>
</tr>
</tbody>
</table>

Example-based machine translation systems are trained from bilingual parallel corpora, which contain sentence pairs like the example shown in the table. Sentence pairs contain sentences in one language with their translations into another. The particular example shows an example of a minimal pair, meaning that the sentences vary by just one element. These sentences make it simple to learn translations of subsentential units. For example, an example-based machine translation system would learn three units of translation:

1. *How much is that X?* corresponds to *Ano X wa ikura desu ka.*
2. *red umbrella* corresponds to *akai kasa*
3. *small camera* corresponds to *chiisai kamera*

Composing these units can be used to produce novel translations in the future. For example, if we have been trained using some text containing the sentences:

> President Kennedy was shot dead during the parade.
> and The convict escaped on July 15th.

We could translate the sentence *The convict was shot dead during the parade.* by substituting the appropriate parts of the sentences.

2. History

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
Example Based Machine Translation

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:
1. Matching fragments against a database of real examples
2. Identifying the corresponding translation fragments
3. Recombining these to give the target text.

Input:

(1) He buys a book on international politics.
(2) a. He buys a notebook.
Kare wa n¯oto o kau.
HE topic NOTEBOOK obj BUY.
b. I read a book on international politics.
Watashi wa kokusai seiji nitsuite kakareta hon o yomu.
I topic INTERNATIONAL POLITICS ABOUT CONCERNED BOOK obj

Output:

(3) Kare wa kokusai seiji nitsuite kakareta hon o kau.

To illustrate, we can take Sato & Nagao’s (1990) example in which the translation of (1) can be arrived at by taking the appropriate fragments – underlined – from (2a, b) to give us (3).

3. EBMT Pyramid

It is perhaps instructive to take the familiar pyramid diagram, probably first used by Vauquois (1968), and superimpose the tasks of EBMT. The source text analysis in conventional MT is replaced by the matching of the input against the example set. Once the relevant example
or examples have been selected, the corresponding fragments in the target text must be selected. This has been termed “alignment” or “adaptation” and, like transfer in conventional MT, involves contrastive comparison of both languages. Once the appropriate fragments have been selected, they must be combined to form a legal target text, just as the generation stage of conventional MT puts the finishing touches to the output. The parallel with conventional MT is reinforced by the fact that both the matching and recombination stages can, in some implementations, use techniques very similar to analysis and generation in conventional MT.

One aspect in which the pyramid diagram does not really work for EBMT is in relating “direct translation” to “exact match”. In one sense, the two are alike in that they entail the least analysis; but in another sense, since the exact match represents a perfect representation, requiring no adaptation at all, one could locate it at the top of the pyramid instead.

### 4. EBMT Paradigm

The basic EBMT system that we are generalizing performs partial exact matches against the examples it has been given, and relies on the multi-engine architecture to assemble the partial translations. Its training consists of indexing the source-language half of each translation.
Example Based Machine Translation

example in an *inverted index* -- a listing, for every possible word, of all the locations at which the word occurs. When asked to perform a translation, it first finds every exactly-matching phrase in its database, regardless of any overlap with any other matches (it always selects the longest match in any particular translation example). To find the matching phrases using an inverted index, one simply retrieves the occurrence lists for the first two words, and determines which of the occurrences of the first word are adjacent to an occurrence of the second word. Then one retrieves the occurrence list for the third word, and either extends the match or creates a new one where the third word appears adjacent to the second one in the example base. Repeat until the entire input text has been processed.
New Sentence (Source)
Yesterday, 200 delegates met with President Clinton

Matches Found

Yesterday, 200 delegates met behind closed doors to discuss the new tax code.

Generiher Flowers is said to have had an affair with President Clinton for many years.

Alignment

Yesterday, 200 delegates met behind closed doors to discuss the new tax code.

Generiher Flowers has allegedly had an affair with President Clinton.

Translated Sentence (Target)
System trafen sich 200 Abgeordnete mit President Clinton

Maximal Length match of source substrings and concatenation of intra-sentence aligned text
5. 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.

5.1. Parallel Corpora

Since EBMT is corpus-based MT, the first thing that is needed is a parallel aligned corpus. 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.

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 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.
5.2. Granularity Of Examples

As Nirenburg et al. (1993) point out, the task of locating appropriate matches as the first step in EBMT involves a trade-off between length and similarity. As they put it: The longer the matched passages, the lower the probability of a complete match and (...) The shorter the passages, the greater the probability of ambiguity (one and the same S can correspond to more than one passage T and the greater the danger that the resulting translation will be of low quality, due to passage boundary friction and incorrect chunking.

5.3. How Many Examples

There is also the question of the size of the example database: how many examples are needed? Not all reports give any details of this important aspect. Table I shows the size of the database of those EBMT systems for which the information is available.

When considering the vast range of example database sizes in Table I, it should be remembered that some of the systems are more experimental than others. One should also bear in mind that the way the examples are stored and used may significantly effect the number needed. Some of the systems listed in the table are not MT systems as such, but may use examples as part of a translation process, e.g. to create transfer rules.

One experiment, reported by Mima et al. (1998) showed how the quality of translation improved as more examples were added to the database: testing cases of the Japanese adnominal particle construction (A no B), they loaded the database with 774 examples in increments of 100. Translation accuracy rose steadily from about 30% with 100 examples to about 65% with the full set. A similar, though less striking result was found with another construction, rising from about 75% with 100 examples to nearly 100% with all 689 examples. Although in both cases the improvement was more or less linear, it is assumed that there is some limit after which further examples do not improve the quality. Indeed, as we discuss in the next section, there may be cases where performance starts to decrease as examples are added.
5.4 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
Example Based Machine Translation

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 system involve a similarity metric 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 phenomenon 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 (4) and (5a).

(4) a. *Watashi wa kompyūtā o kyōsuru.*
I topic COMPUTER obj SHARE-USE.
‘I share the use of a computer.’
b. *Watashi wa kuruma o tsukau.*
I topic CAR obj USE.
‘I use a car.’

(5) *Watashi wa dentaku o shiyō osuru.*
I topic CALCULATOR obj USE.
a. ‘I share the use of a calculator.’
b. ‘I use a calculator.’

Given the input in (5), the system might incorrectly choose (5a) as the translation because of the closer similarity of *dentaku* ‘calculator’ to *kompyūtā* ‘computer’ than to *kuruma* ‘car’ (the three words for ‘use’ being considered synonyms; whereas (5b) is the correct translation. So (4a) is an exceptional example because it introduces the unrepresentative element of ‘share’. The situation can be rectified by removing example (4a) and/or by supplementing 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.

5.5. 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 ij¯o desu ‘let’s change the subject’, depending on the context. 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.

5.5.1. Annotated Tree Structures

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. Figure 2 shows how the Japanese example in (6) and its English translation is represented.

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

*Figure 2.* Representation scheme for (6). (Watanabe 1992: 771).

More recently a similar approach has been used by Poutsma (1998) and Way (1999): here, the source text is parsed using Bod’s (1992) DOP (data-oriented parsing) technique, which is itself a kind of example-based approach, then matching subtrees are combined in a compositional manner. In the system of Al-Adhaileh & Kong (1999), examples are represented as dependency structures with links at the structural and lexical level expressed by indexes.
Figure 3 shows the representation for the English–Malay pair in (7).

(7) a. He picks the ball up.

b. *Dia kutip bola itu.*

The nodes in the trees are indexed to show the lexical head and the span of the tree of which that item is the head: so for example the node labelled “ball(1)[n](3-4/2-4)” indicates that the subtree headed by *ball*, which is the word spanning nodes 3 to 4 (i.e. the fourth word) is the head of the subtree spanning nodes 2 to 4, i.e. *the ball*. The box labelled “Translation units” gives the links between the two trees, divided into “Stree” links, identifying subtree correspondences (e.g. the English subtree 2-4 *the ball* corresponds to the Malay subtree 2-4 *bola itu*) and “Snod” links, identifying lexical correspondences (e.g. English word 3-4 *ball* corresponds to Malay word 2-3 *bola*).
5.5.2. Generalized Examples

In some systems, similar examples are combined and stored as a single “generalized” example. Brown (1999) for instance tokenizes the examples to show equivalence classes such as “person’s name”, “date”, “city name”, and also linguistic information such as gender and number. In this approach, phrases in the examples are replaced by these tokens, thereby making the examples more general.

This idea is adopted in a number of other systems where general rules are derived from examples show how examples can be generalized for the purposes of retrieval, but with a corresponding precision–recall trade-off.

The idea is taken to its extreme in Furuse & Iida’s (1992a, b) proposal, where 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 sets of words which are concrete instances of each category. Each type is exemplified in (8–10), respectively.

(8) Sochira ni okeru => We will send it to you.
Sochira wa jimukyoku desu => This is the office.

(9) X o onegai shimasu => may I speak to the X
(X = jimukyoku ‘office’, . . . )
X o onegai shimasu => please give me the X
(X = bang¯o ‘number’, . . . )

(10) N1 N2 N3 => the N30 of the N10
(N1 = kaigi ‘meeting’, N2 = kaisai ‘opening’, N3 = kikan ‘time’)
N1 N2 N3 N20 N30 for N10
(N1 = sanka ‘participation’, N2 = m¯oshikomi ‘application’, N3 = y¯oshi ‘oshi”)
Example Based Machine Translation

‘form’)

As in previous systems, the appropriate template is chosen on the basis of distance in a thesaurus, so the more appropriate translation is chosen as 

(11)

a. *jinjika o onegai shimasu* (*jinjika* = ‘personnel section’) => may I speak to the personnel section

b. *kenkyukai kaisai kikan* (*kenkyukai* = ‘workshop’) => the time of the workshop

c. *happy¯o m¯oshikomi y¯oshi* (*happy¯o* = ‘presentation’) => application form for presentation

5.5.3. Statistical Approaches

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 system is based. What is stored is the precomputed statistical parameters which give the probabilities for bilingual word pairings, the “translation model”. The “language model” which gives the probabilities 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.

5.6. 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. 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 probabilities. In more conventional EBMT systems the matching process may be more or less linguistically motivated.
5.6.1. Character-based Matching

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, \textit{modulo} alphanumeric strings, were possible: (12a) would be matched with (12b), but the match in (13) would be missed because the system has no way of knowing that \textit{small} and \textit{large} are similar.

(12) a. This is shown as A in the diagram.
b. This is shown as B in the diagram.
(13) a. The large paper tray holds up to 400 sheets of A3 paper.
b. The small paper tray holds up to 300 sheets of A4 paper.

5.6.2. Word-based Matching

Perhaps the “classical” similarity measure, suggested by Nagao (1984) and used in many early EBMT systems, is the use of a thesaurus or similar means of identifying word similarity on the basis of meaning or usage. 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, or by collocation scores such as mutual information) in the example sentences. This measure is particularly effective in choosing between competing examples, as in Nagao’s examples, where, given (14a, b) as models, we choose the correct translation of \textit{eat} in (15a) as \textit{taberu} ‘eat (food)’, in (15b) as \textit{okasu} ‘erode’, on the basis of the relative distance from \textit{he} to \textit{man} and \textit{acid}, and from \textit{potatoes} to \textit{vegetables} and \textit{metal}.

(14)
a. A man eats vegetables. \textit{Hito wa yasai o taberu}.
b. Acid eats metal. \textit{San wa kinzoku o okasu}.
Example Based Machine Translation

(15)
a. He eats potatoes. *Kare wa jagaimo o taberu.*
b. Sulphuric acid eats iron. *Ry¯usan wa tetsu o okasu.*

5.6.3. *Structure-based Matching*

Earlier proposals for EBMT, and proposals where EBMT is integrated within a more traditional approach, assumed that the examples would be stored as structured objects, so the process involves a rather more complex tree-matching (e.g. Maruyama & Watanabe 1992; Matsumoto et al. 1993; Watanabe 1995; Al-N Adhaileh & Tang 1999) though there is generally not much discussion of how to do this (cf. Maruyama & Watanabe 1992; Al-Adhaileh & Tang 1998), and there is certainly a considerable computational cost involved. Indeed, there is a not insignificant literature on tree comparison, the “tree edit distance” which would obviously be of relevance.

5.7. *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 (“alignment” or “adaptation”), and to combine these so as to produce a grammatical target output (“recombination”). 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, they have received considerably less attention.

(20)
a. He buys a notebook ⇒ *Kare wa n¯oto o kau*
b. I read a book on politics ⇒ *Watashi wa seiji nitsuite kakareta hon o yomu*
6. Applications of the Component Technologies

Generalized EBMT is not just useful as a stand-alone system for translating text, but it (and in some cases one of the underlying components) is also useful for other applications. Two which have already been implemented are speech-to-speech translation and cross-language information retrieval.

6.1. Speech Translation

EBMT system to translate a conversation in real time, and that is precisely how EBMT (as part of a multi-engine translation system) is used by the DIPLOMAT project. DIPLOMAT has used EBMT for translation between English and Croatian, Haitian Creole, Korean, and Spanish.

In the DIPLOMAT system, the Sphinx-II continuous speech recognizer is used to transcribe the user's spoken utterance into text. This text, after an opportunity to correct recognition errors, is translated and then synthesized in the other language using the Phonebox concatenative speech synthesizer developed at CMU.

DIPLOMAT is a bi-directional system for translating a conversation, so it uses two copies of the translator software. This not only permits the translation of each of the conversants' speech, but also gives us the opportunity of providing a back-translation. If the system's back-translation of its output correctly conveys the meaning of the original input, we have much greater confidence that it actually translated things correctly.

6.2. Cross-Language Retrieval

One increasingly important area of research in recent years has been cross-language information retrieval, where a query in one language is used to find documents in another language. Perhaps the most common way of crossing the language barrier is to translate the
Example Based Machine Translation

query (translating the entire document collection is usually impractical). But since queries tend to consist of isolated words with an occasional short phrase, rather than a full sentence or paragraph, full-blown machine translation isn't applicable to translating queries.

A common method of translating the query is to look up each word in a bilingual dictionary, and replacing it with every possible translation listed for that word. This produces a new query in the other language, which can then be used with standard monolingual retrieval systems. Statistically-generated dictionaries can be used with this method simply by replacing the general-purpose dictionary with the statistical one, but it is possible to do even better. In addition to being attuned to the actual usage of words in the training corpus (and thus sometimes listing translations which would not appear in a general-purpose dictionary), a statistically-generated dictionary also contains frequency information -- which we can use to give greater importance to the more common translations.

7. Conclusion

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.

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 generalize the examples as much as possible. So part of the criterion might be whether the examples are used.

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.
8. References

Journal article

Web sites
http://www.springerlink.com/content/k00lw822j783503t/
www.cs.cmu.edu/~ralf/ebmt/ebmt.html