AN INTRODUCTION TO EXAMPLE-BASED MACHINE TRANSLATION

Nadia Luiza DINCĂ
Research Institute for Artificial Intelligence, Bucharest

Abstract: *The essence of EBMT, called machine translation by example-guided inference, or machine translation by the analogy principle by Makoto Nagao (1984), 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 languages phrases, and finally by properly composing these fragmental translations into one long sentence."*

The ideal translation unit for EBMT is the sentence. Only if the translation of an identical sentence is not available in the bilingual corpus, do EBMT systems make use of some of similarity metric to find the best matching translation examples. Suitable sub-sequences are iteratively replaced, substituted, modified or adapted in order to generate the translation.

The main components that EBMT uses are: matching fragments against a database of real examples; identifying the corresponding translation fragments; recombining these to give the target text.

In this paper I will introduce the main Example-Based Machine Translation systems and I will compare their translation results to the human translation, by considering, for English and Romanian languages, the same source texts.

Key words: *Example-Based Machine Translation, translation unit, source and target languages*

I. Introduction

The main process of the example-based machine translation is divided into three phases. First, find the most similar examples as the input sentence. Then, recombine the translation of the input sentence according to most similar example and bilingual dictionary. Lastly, produce the translation of the input sentence. The main resources are bilingual dictionary, thesaurus, the standard template system and bilingual sentence aligned corpora. The bilingual dictionary, thesaurus, and the standard template system are used to calculate the similarities between two words, two chunks, and two sentences. At the same time, bilingual dictionary and bilingual sentence aligned corpora are used to adjust and produce the translation.

In its very first moment, the example-based machine translation was defined as a translation by analogy which was using an unannotated example data base, created, usually, from a bilingual dictionary- (NAGAO, 1984: 173-180). The equivalents were represented as word pairs, except the verb equivalents, formalised as case frames.

Later, the structural translation conceives the representation of translation examples as dependency trees with explicited links established between sub-trees (including the leaf nodes, corresponding to the lexical units). These links allow using parts of the translation example or sub-trees in order to recognise, for the source language, the exact match between input segments and structures, and for the target language, to select and to combine the equivalent translation units.

The translation example is a lexical phrase, sometimes having a different meaning than the one composed by the meanings of its every word, and to whom is assigned, for the target language, a translation and an exact meaning.

A translation example is composed by three parts:

- an English dependency tree (in this paper, English will be the source language);

- a French dependency tree (the target language, in the paper);
- correspondence links.

These three parts are shown in the next verb phrase, extracted from G. Orwell's novel, "1984", subject of a very extended linguistic project, *Multext-East*:

had imagined everything ↔ avait tout imaginé
en_e ([en1, [had, aux]],
[en2, [imagine, v],
[en3, [everything, pron]]]])
fr_e ([fr1, [avait]],
[fr2, [imaginer, v],
[fr3, [tout, pron]]]])
clinks ([[fr1, en1], [fr2, en2], [fr3, en3]]).

Each number with prefix 'en' or 'fr' in the word-dependency trees represents the ID of the sub-tree. Each node in a tree contains a word (in root form) and its syntactic category. A correspondence link is represented as a pair of IDs: *clinks ([[fr1, en1], [fr2, en2], [fr3, en3]]).* A word-dependency (sub)tree which has a correspondence link is translatable; e.g.: el, e2, e3, fr l, fr2, fr3. A translatable tree in which some translatable sub-trees are removed is also translatable; e.g.: e l - e 2 , e 2 - e 3 , e l - e 2 - e 3 , fr l - fr 2, fr2 - fr3, fr1 - fr2- fr3 .

The translation process consists of three steps: decomposition, transfer, and composition. In decomposition, the system decomposes a source word-dependency tree into translation units, and makes a source matching expression. In the transfer step, the system replaces every ID in the source matching expression with its corresponding ID. In the composition step, the system composes the target word-dependency tree according to the target matching expression.

II. Underlying Problems

Since EBMT is corpus-based MT, the first thing that is needed is a parallel aligned corpus. Once a suitable corpus has been located, there remain the problems of aligning it, i.e. identifying at a finer granularity which segments (typically sentences) correspond to each other.

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

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" (CARL and HANSEN, 1999: 619).

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.

Somers (1999) discusses about three increasingly specific criteria for defining EBMT:

1. EBMT uses a bilingual corpus.

2. EBMT uses a bilingual corpus as its main knowledge base.
3. EBMT uses a bilingual corpus as its main knowledge base, at run-time.

Somers states that the first two criteria are two broad, but he argues that the third criterion may be too strict, as it rules out, for instance, statistical MT, where all the corpus-driven probabilities are computed in advance.

There are two reasons for which a corpus is used at run-time in an MT system:
1. the system uses knowledge that can only be dynamically acquired at run-time by accessing an entire corpus, or sections of it whose extent cannot be determined in advance.
2. the system uses knowledge that could be extracted in advance, but is instead left implicit in the corpus, and extracted as needed at run-time.

In fact, most EBMT systems assume the existence of a bilingual lexicon to perform substitutions in examples. Work on semantic database like WordNet has shown that much of their information can be misleading in specific domains. For example, a MT system dealing with weather reports would have serious problems using a thesaurus where very frequent words like *snow* and *C* (for *Celsius*) were considered semantically similar because they are both synonyms for cocaine (TURCATO et al., 2000)

III. Linear Example-Based Machine Translation Systems

Traditional linear or non-structural Example-Based Machine Translation systems that do not extract a rule base nor model themselves on transfer-based systems typically extract target language equivalents of overlapping partial exact matches of the source language input dynamically and recombine them in an appropriate manner to produce target language translations. There is frequently no or very little pre-processing translation examples rather than against abstract representations of them. Furthermore, the majority of, but by no means all, bilingual relationships are computed at run-time. This compares with rule or pattern-based systems which compute all bilingual knowledge in a pre-matching extraction phase. Moreover, recombination represents more of a challenge as there are no sentential patterns of translation or translation rules to determine the order of items in the target language.

An example system of this strategy is the MEG system (SOMERS et al., 1994). This approach is claimed to be a *pure* EBMT system in that no external linguistic knowledge, no matter how minimal, is used. Only information gleaned from the corpus itself is used.

First, the corpus is POS tagged, even this is undertaken using a tag-set derived entirely from the corpus to maintain maximum portability. Subsequently, word-level alignment is carried out. Matching a source language input against the corpus is carried out at run-time. The tagged source language input is matched against each relevant source language sentence in the corpus to produce a possibly non-continuguous fragment which the two sentences have in common. Strong (word and tag) or weak (tag only) correspondences are computed. By a similar method, the target language equivalent of each source language fragment is computed.

The Pangloss Example-Based Machine Translation engine (PanEBMT) is a translation system requiring essentially no knowledge of the structure of a language, merely a large parallel corpus of example sentences and a bilingual dictionary. Input texts are segmented into sequences of words occurring in the corpus, for which translations are determined by subsentential alignment of the sentence pairs containing those sequences. These partial translations are then combined with the results of other translation engines to form the final translation produced by the Pangloss system. In an

internal evaluation, PanEBMT achieved 70.2% coverage of unrestricted Spanish news-wire text, despite a simplistic subsentential alignment algorithm, a suboptimal dictionary, and a corpus from a different domain than the evaluation texts.

IV. Structured Example-Based Machine Translation Systems

Some of the first approaches to Example-Based Machine Translation involve the storage of the translation examples as fully annotated tree structures with alignments at the lexical and structural level. These aligned tree structures served as the rule base against which parsed source language input sentences were matched. Typically, the closest matching source language structure to the parsed source language input is retrieved. The Alignments enable the retrieval of translations of segments of the source language input from other translation examples in the corpus. The corresponding target language tree is then constructed from these fragments. The target language sentence is subsequently generated.

Matching against a set of tree structures is a more complex task than matching against a set of raw translation examples and involves a considerable computational cost. EBMT based on the correspondence of tree structures also requires a significant amount of external linguistic knowledge in the form of parsers and perhaps bilingual lexicons. This detracts from portability. However, the more linguistic information that a system is given, in theory, the more accurate its translation is. One advantage of including structural information in translation examples is the ability to represent explicitly alignments between languages that indicate a structural divergence.

MBT2 is the second prototype system in S. Sato and M. Nagao's Memory-based Translation Project. The two researchers introduced the representation called *matching expression*, which represents the combination of fragments of translation examples. The translation process consists of three steps: (i) make the source matching expression from the source sentence. (ii) transfer the source matching expression into the target matching expression. (iii) construct the target sentence from the target matching expression.

The concept *matching expression* considers three basic operations applied on dependency sub-trees which are already in database: delete the identifier of a certain sub-tree; replace the identifier with a matching expression; add a matching expression as a child of root node of the identifier.

This mechanism generates some candidates of translation. To select the best translation out of them, a score of a translation was defined, so that it should reflect the correctness of the translation unit. The last is a fragment of a source (or target) word-dependency tree, and also a fragment of a translation example. The more similar these two environments are, the better.

The system proposed by H. Kaji in 1992 is a two-phase example-based machine translation methodology which develops translation templates from examples and then translates using template matching.

A translation template is a bilingual pair of sentences in which corresponding units (words and phrases) are coupled and replaced with variables. Conditions concerning syntactic categories, semantic categories, etc. are attached to each variable. A word or phrase satisfying the conditions can be substituted for a variable. The two pseudo-sentences constituting a template include the same set of variables.

The learning procedure is divided into two steps. In a first step, a series of translation templates is generated from each pair of sentences in the corpus. The first step is subdivided into coupling of corresponding units (words and phrases) and

generation of translation templates. In the second step, translation templates are refined to resolve conflicts among them.

Translation based on templates consists of (i) source language template matching, (ii) translation of words and phrases and (iii) target language sentence generation. First, a translation template is retrieved. Words and phrases in the source language sentence are then bound to each variable in the template. Second, the words and phrases which are bound to variables are translated by a conventional machine translation method. Finally, a target language sentence is generated by substituting the translated words and phrases for the variables in the target language part of the translation template.

V. Language-Neutral Generalisation Techniques

An approach to translation pattern extraction by the Department of Computer Engineering and Information Sciences at the University of Bilkent, Ankara, Turkey is based on analogical reasong between pairs of translation examples in a sentence-aligned bilingual corpus. Their attempts at extracting translation patterns involved the correlation of syntactic structures between English and Turkish (GÜVENIR and TUNÇ, 1998). However, the authors consider that they could not find reliable parsers for both languages. This led to the development of language-neutral techniques for extracting correspondences, or translation templates as they term them, between languages by analogical methods.

Their method is based on the next assumption: given two sentence-pairs in a bilingual corpus, the orthographically similar parts of the two source language sentences correspond to the orthographically similar parts of the two target sentences. In a similar way, the different parts of the two source sentences correspond to different parts of the two target sentences. The differences are replaced by variables, in order to produce general examples.

As an example, the next two sentence pairs (1) may be generalized to produce the translation pattern (2). The similar text shows similar parts of the two sentence pairs:

(1) I gave the ticket to Mary <-> Mary' e bileti verdim.
I gave the pen to Mary <-> Mary' e kalemi verdim.
(2) I gave the X_S to Mary <-> Mary'e X_T+i verdim.

Note that due to the correspondence of the variables X_S and X_T in the translation pattern, it is realized the bilingual relationship between the items *ticket* and *pen*.

It is obvious that more translation rules are generated from the translation rules extracted by recursively applying the same induction process to the translation rules extracted. This ability to create more refined translation rules adds to the flexibility of the system.

This section shows how translation patterns are formed by generalisation of translation examples. However, this is not the only method by which they are created. A template-based EBMT system called Gaijin (VEALE and WAY, 1997) fits loosely into the categorisation schema of EBMT systems that operate by extracting translation patterns to be used as a rule base.

The translation templates extracted represent mappings between source and target chunks for each sentence pair in a bilingual corpus. A single translation template provides the sentential context for the translation of a given source language input, but the translation is performed by using aligning source and target languages chunks from

other translation templates. At this point, the Gaijin system marks as departure from the approaches described above: when the aligned chunks within the templates do not match the source input exactly, it is possible to adapt the example chunks to match the source input exactly. Minor differences between a source chunk in a translation template and a chunk of the source input can be rectified by adaptation. Any changes made to the source example chunk are reflected in the target chunk. Templates and chunks are retrieved in the matching process based on the level of ease of adaptability to the source input.

VI. Conclusions

One of the most important aspects of the EBMT is the evaluation. In fact, the declaration evaluation has as purpose to measure the ability of a MT system to handle texts representative of an actual end-user.

As with feasibility and internal evaluation, we look at coverage of linguistic phenomena and handling of samples of real text. Declarative evaluations generally test for the functionality attributes of intelligibility (how fluent or understandable it appears to be) and fidelity (the accurateness and completeness of the information conveyed).

Readability or fluency means the extent to which a sentence reads naturally, the ease with which a translation can be understood, i.e. clarity to the reader.

The comprehensibility is the extent to which the text as a whole is easy to understand. That is, the extent to which valid information and inferences can be drawn from different parts of the same document.

The coherence refers to the degree to which the reader can describe the role of each individual sentence (or group of sentences) with respect to the text as a whole. Theories such as Rhetorical Structure Theory attempt to formalize coherence using a set of inter-segment relations (such as Cause, Solutionhood, Elaboration) that express the internal document structure.

In the following lines we introduce the main aspects of the coverage of corpus-specific phenomena. Coverage refers to the ability of the system to deal satisfactorily with linguistic phenomena, both generally addressing known cross-language phenomena and specifically addressing phenomena in a corpus of interest:

a. Style. This is the subjective evaluation of the correctness of the style of each sentence. This quality is also commonly referred to as *register* and includes degree of formality, forcefulness and bias as exhibited through both lexical and morpho-syntactic choices.

b. Accuracy. This refers to the capability of the software product to provide the right or agreed results or effects with the needed degree of precision.

c. Consistency. This is the capability of the system to produce from a given input, and at a given point in time, the same output.

d. Terminology. Its metrics is responsible for the percentage of domain terms correctly translated. Names should be transliterated or translated (e.g., London -> fr. Londres) as appropriate.

e. Wellformedness. It is the degree to which the output respects the reference rules of the target language at the specified linguistic level. Systran, for example, uses at least seven types of errors to rank the quality of the output: segmentation/ tokenization; morphological analysis; homograph analysis; syntactic analysis; target language word selection; target language morphology; target language word order; target language grammar.

In conclusion, an example-based machine translation system to exploit and integrate a number of knowledge resources, such as linguistics and statistics, and symbolic and numerical techniques, for integration into one framework. In this way, rule-based morphological, syntactic and/or semantic information is combined with knowledge extracted from bilingual texts which is then re-used in the translation process.

However, it is unclear how one might combine the different knowledge resources and techniques in an optimal way. In EBMT, therefore, the question is asked: what can be learned from a bilingual corpus and what needs to be manually provided? Furthermore, we remain uncertain as to how far the EBMT methodology can be pushed with respect to translation quality and/or translation purpose. Finally, one wonders what the implications and consequences are for size and quality of the reference translations, (computational) complexity of the system, size ability and transportability, if such an approach is taken.

BIBLIOGRAPHY

Carl, M., Hansen, S., *Linking Translation Memories with Example-Based Machine Translation*. in Streiter, O., M. Carl & J. Haller (eds.) *Hybrid Approaches to Machine Translation*, 1999.

Halil Altay Güvenir and Ilyas Cicekli, *Learning Translation Templates from Exaples*. In Information Systems, 23(6):353-363, 1998.

Nagao, M., *A framework of a mechanical translation between Japanese and English by analogy principle*, in Alick Elithorn & Ranan Banerji, eds., *Artificial and Human Intelligence*: edited review papers at the International NATO Symposium on Artificial and Human Intelligence, Amsterdam, North-Holland: Elsevier Science Publishers, 1984, chap. 11, pp. 173-180.

Somers, H., McLean, I., Jones, D., *MT seen as multilingual example-based generation*. IWMT '94, Limerick, 1994

Turcato, D., Popowich, F., Toole, J., Fass, D., *Adapting a Synonym Database to Specofic Domains*. in *Proceedings of ACL'2000 Workshop on Information Retrieval and Natural Language Processing*, Hong Kong, October, 2000

Veale, T., Way, A., *Gaijin: A Template-Driven Bootstrapping Approach to Example-Based Machine Translation*, in Proceedings of the NeMNLP'97, *New Methods in Natural Language Processing*, 1997.