Recent Advances in Example-Based Machine Translation

Michael Carl and Andy Way (editors)
(Universität des Saarlandes and Dublin City University)

Dordrecht: Kluwer Academic Publishers (Text, speech and language technology series, edited by Nancy Ide and Jean Véronis, volume 21), 2003, xxxi+482 pp; hardbound, ISBN 1-4020-1400-7, $173.00, £115.00, €180.00

Reviewed by
Walter Daelemans
University of Antwerp

This book, an outcome of a 2001 workshop on example-based machine translation (EBMT) in Santiago de Compostela, very appropriately starts with a preface by professor Makoto Nagao in which he explains how the limits of rule-based machine translation (MT) led him to propose his principle of translation by analogy in 1981 (published as Nagao 1984). His idea, inspired by second-language learning methodology, is convincing and elegant. We learn a language not by deep linguistic analysis and rule application, but by rote learning and analogy-based generalization. Starting from simple sentences and their translations, variations of these sentences can also be translated by using similarity-based reasoning. If the Japanese translation of a sentence such as *A man eats vegetables* is memorized (implicitly including changes in word order, morphology, translation selection of *eat*, etc.), then a sentence such as *He eats potatoes* can be translated correctly by analogy to this memorized case if the semantic relation between *he* and *a man* and between *vegetables* and *potatoes* can be determined (e.g., using a thesaurus). Such a method bypasses deep linguistic analysis by leaving transfer rules implicit in an aligned bilingual corpus.

Michael Carl and Andy Way have done the field of MT and that of natural language processing in general a big favor by producing this collection. This is the first (fat) book-sized collection of current research in this fascinating area of MT, and it contains a number of chapters that will bring anyone interested quickly up to speed in this research area. In order to differentiate EBMT from statistical MT and rule-based MT, it is tempting to describe it as a hybrid technique taking a middle ground between rule-based methods (linguistic knowledge used in the representation of translation units) and data-oriented methods (learning from bilingual parallel corpora, similarity-based reasoning). But in reality, EBMT appears in this book as a widely varying bunch of somehow related approaches: a concept that can best be described in an example-based way. In their introduction, Carl and Way acknowledge this lack of an analytical definition but see no harm in it and compare the situation with other "scientific realities" such as artificial intelligence. The volume is divided into four parts, the first being about historical and technical foundations, and the other three describing current research.

Part I contains a wealth of background information. Chapter 1, by Harold Somers ("An Overview of EBMT") traces the history of the idea and dives into the underlying problems to be solved in applying it: collecting and aligning parallel corpora, deciding on the grain size and linguistic representation of the examples (sentence level or below, superficial or deep?), the number and selection of examples needed, storage, indexing

and matching of examples, and issues involved in using the best matching examples to produce grammatical output (adaptation of the retrieved knowledge). The remainder of the chapter discusses different flavors, extensions, and uses of EBMT and describes how EBMT has been incorporated in multiengine systems, in parallel with rule-based and statistical systems. The anecdotal, impressionistic, and limited evaluation of EBMT available in the literature according to Somers is surprising. Especially given the trend toward multiengine systems, it is to be hoped that a thorough comparative evaluation of EBMT will become available soon. Somers mentions corpus coverage, simple extension, linguistic theory-freeness, and easy development from aligned corpora as the main advantages of EBMT and scalability and dealing with some translation problems as its main problems. He sees it not as a rival to rule-based MT but as an enhancement of it. Given the results of, for example, Sumita in a later chapter of the book, I think he gives up on "pure" EBMT too easily.

Chapter 2 by Davide Torcato and Fred Popowich ("What is Example-Based Machine Translation?") attempts to single out what sets EBMT apart from other approaches. The authors argue that the linguistic information being used by an MT system (rather than how it is represented or acquired) should be the basis for classification. On this basis, they go to great length contending that most of the properties of EBMT (use of parallel corpus, holistic approach, etc.) are not really distinguishing properties and that only the translation-by-analogy principle stands out as truly example-based.

In chapter 3 ("EMBT in a Controlled Environment"), Reinhard Schäler, Andy Way, and Michael Carl see EBMT (in the guise of phrasal lexicons) as a way of enhancing translation memories (TM). TM is arguably the commercially most successful approach to MT. The idea is probably as old as EBMT (I heard it being mentioned in the very first workshop I ever attended in 1984 in Davos). In TM, a translator has access to previous translations of sentences that are retrieved by (fuzzy) matching. Such an approach will have high precision but very low recall. Schäler et al. demonstrate that by extending the approach to the subsentential level (phrases), recall can be improved. The main contributions of this paper are pointing to the use EBMT might have in improving the usefulness and coverage of TM systems (although that point has been made by many others as well), analyzing how this could be done in the context of controlled language TM, and what the benefits and limitations could be. Unfortunately it remains a position paper more than a proof of concept.

The foundational part of the book ends with an interesting chapter by Bróna Collins and Harold Somers ("EBMT Seen as Case-Based Reasoning") in which MT is analyzed in the framework of case-based reasoning (CBR), a well-known reasoning approach in AI. The goal is to investigate whether other applications of CBR can contribute to its application to MT. Unfortunately, the chapter doesn't go very deep into analyzing the relations and differences between different kinds of "lazy learning," of which CBR is an instance (memory-based, instance-based, and exemplar-based reasoning and learning methods have all been proposed in the AI literature). I would refer people interested in that topic rather to Aha (1997) or Kasif et al. (1998). Nevertheless, the explicit analogy between the two analogical approaches in this chapter is extensive and insightful and may inspire new research on extensions and variations of EBMT.

EBMT systems differ in whether they acquire the knowledge implicit in the bilingual corpus at runtime or off-line in advance. Runtime approaches, of which a few are presented in part II of the book, extract their translation units during translation from a sentence-aligned bilingual corpus. Chapter 5 ("Formalizing Translation Memory") by Emmanuel Planas and Osamu Furuse addresses in a sense the same issues as chapter 3, trying to extend translation memory approaches to EBMT. The way

this is done in their "shallow translation" approach is to add lemmas, POS tags, and string-edit distance on multiple levels of sentences for matching in TM. Preliminary evaluation comparing the matching approach to a standard TM-matching algorithm shows that especially at high similarity thresholds, their approach retrieves more useful cases. With the contribution of Eiichiro Sumita ("An Example-Based Machine Translation System using DP-Matching between Word Sequences"), we get near to the pure EBMT approach, generalizing examples on the fly and not using any syntactic parsing or bilingual treebanks. On a reasonably sized corpus (200,000 sentences), the approach covers about 90% of the 500 test sentences, with about 80% acceptable translations, taking 10 seconds per sentence on average. This is hopeful news for the pure approach, although many problems remain, especially regarding handling of sparse data, long sentences, and context dependency in the approach. Francis Bond and Satoshi Shirai ("A Hybrid Rule and Example-Based Method for Machine Translation") contribute an approach to combine the strengths of rule-based and example-based approaches, and Tantely Andriamanankasina, Kenji Araki, and Koji Tochinai ("EBMT of POS-Tagged Sentences by Recursive Division via Inductive Learning") apply the CBR approach to subsententially aligned examples. Both approaches claim promising results.

Parts III and IV of the book concern approaches to and systems of EBMT that do not acquire their knowledge from the bilingual corpus dynamically and on-line but rather do it off-line, either as extracted translation templates (part III) or as something that starts to resemble structural transfer rules (part IV). Notice that from the point of view of a machine-learning interpretation of example-based reasoning and learning, something bizarre is happening here. The techniques described in these last two parts of the book become increasingly less example-based. It is not because rules are learned from examples that an approach becomes example-based; the crucial aspect is that the examples themselves are used in reasoning, not generalizations extracted on the basis of them. The fundamental difference between rule-based and example-based approaches is that the former, because of the nature of rules, have to abstract from low-frequency and untypical examples in order to formulate compact rules, whereas the latter keep all information, exceptions, and noise included. This does not imply that the work described in these papers is any less interesting, of course, but I would not necessarily call it EBMT. Ilyas Cicekli and H. Altay Güvenir ("Learning Translation Templates from Bilingual Translation Examples") show how translation templates can be learned by means of a language-independent method for generalizing exemplars based on similarities and differences. Ralf Brown ("Clustered Transfer Rule Induction for Example-Based Translation"), meanwhile, adopts a similar approach to that used in the previous chapter to learn translation templates and adds to that a bottom-up agglomerative clustering method for both words and replacement rules. He shows that clustering and rule induction each outperform simple string matching and that the combination outperforms both. In chapter 11 ("Translation Patterns, Linguistic Knowledge and Complexity in an Approach to EBMT"), Kevin McTait discusses a template extraction based on similarity in distributions of strings in source and target language sentences and fails to improve its accuracy significantly by adding morphological analysis and POS tagging. Finishing Part III, Michael Carl ("Inducing Translation Grammars from Bracketed Alignments") presents a system that extracts lexical transfer rules and translation templates from a tagged and bracketed corpus, thereby effectively moving from example-based reasoning to grammar induction. Interestingly, the induced grammar has the desirable properties of homomorphy, invertibility, and compositionality.

Part IV moves even further away from example-based approaches in requiring the extraction of translation grammars from structured representations (bilingual treebanks). Kaoru Yamamoto and Yuji Matsumoto, in chapter 13 ("Extracting Translation

Knowledge from Parallel Corpora"), report on two successful studies on extracting translation knowledge from parallel automatically annotated corpora, with robust results even with the unavoidable annotation errors. They also show that chunk boundaries, especially, provide useful information for translation and that dependency relations are crucial for longer phrasal translation pairs. Hideo Watanabe, Sadao Kurohashi, and Eiji Aramaki ("Finding Translation Patterns from Paired Source and Target Dependency Structures") follow a similar approach but add a method for extracting translation patterns by comparing correct and wrong translations, as a means of enhancing a database of translation patterns. Arul Menezes and Stephen Richardson ("A Best-First Alignment Algorithm for Automatic Extraction of Transfer Mappings from Bilingual Corpora") describe a logical-form alignment algorithm for the Microsoft MSR-MT architecture. Andy Way, in the final chapter ("Translating with Examples: The LFG-DOT Models of Translation") describes various models based on adapting data-oriented parsing (DOP, a memory-based parsing method) to translation using an LFG-parsed bilingual treebank; he shows how this approach solves the problem of boundary friction (retrieved translations that do not fit the syntactic context).

Although it is very clear that the book consists of a number of independent papers and keeps some of the repetitive and overlapping flavor of a workshop proceedings volume, the editors and authors have done an excellent job in making this a coherent and self-contained book by adding cross-references and inviting additional papers. My main, not very vital regret regarding the book is that although there is some reference to relevant "lazy learning" techniques such as case-based, memory-based, and instance-based learning in AI, few links are made to the application of these ideas in areas of computational linguistics other than MT (ranging from case-based or memory-based phonology to pragmatics). As an example, see the special issue of *Journal of Experimental and Theoretical Artificial Intelligence* that I edited several years ago (Daelemans 1999). This work starts from the same inspiration as EBMT: Language-processing behavior is best modeled as similarity-based reasoning on the basis of stored experiences rather than as based on explicit rules extracted from these experiences. The motivation for this assumption, which has considerable empirical support, is that language data contain so many subregularities and exceptions that rules abstracting away from these exceptional or infrequent cases are at a disadvantage. It is my (undoubtedly biased) impression that this work sometimes goes a lot further than EBMT work in analyzing deeper questions such as why these techniques are better suited for NLP tasks and which task-independent similarity metrics and algorithms are best suited for solving NLP tasks in this paradigm. There may be a chance for mutually beneficial interaction here. In summary, I would recommend this book to everyone active or interested in MT, and especially the papers of the foundational part I to computational linguistics researchers in general.

References

Aha, David W., editor. 1997. Lazy learning. Special quintuple issue, *Artificial Intelligence Review*, 11(1–5):1–423.

Daelemans, Walter, editor. 1999. Memory-based language processing. Special issue, *Journal of Experimental and Theoretical Artificial Intelligence*, 11(3):287–471.

Kasif, Simon, Steven Salzberg, David Waltz, John Rachlin, and David Aha. 1998. A probabilistic framework for memory-based reasoning. *Artificial Intelligence*, 104(1–2):287–311.

Nagao, Makoto A. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In Alick Elithorn and Ranan Banerji, editors, *Artificial and Human Intelligence*, pages 173–180. Amsterdam: North-Holland.

Walter Daelemans is a professor of Computational Linguistics at the University of Antwerp (and part-time at Tilburg University). His research area is machine learning of language. Daelemans can be reached at CNTS, Department of Linguistics, Universiteitsplein 1 (A), B-2610 Antwerpen, Belgium; e-mail: walter.daelemans@ua.ac.be.