Machine Translation **14:** 113–157, 1999.

Review Article: Example-based Machine Translation

HAROLD SOMERS
Centre for Computational Linguistics, UMIST, PO Box 88, Manchester M60 1QD, England
(E-mail: harold@fs1.ccl.umist.ac.uk)

Abstract. In the last ten years there has been a significant amount of research in Machine Translation within a "new" paradigm of empirical approaches, often labelled collectively as "Example-based" approaches. The first manifestation of this approach caused some surprise and hostility among observers more used to different ways of working, but the techniques were quickly adopted and adapted by many researchers, often creating hybrid systems. This paper reviews the various research efforts within this paradigm reported to date, and attempts a categorisation of different manifestations of the general approach.

Key words: example-based MT, hybrid methods, corpora, translation memory

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 (Brown et al. 1988). IBM's "purely statistical" approach, inspired by successes in speech processing, and characterized 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" (TMI 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", "case-based" and "experience-guided" 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.

There is an obvious affinity between EBMT and Machine Learning techniques such as Exemplar-Based Learning (Medin & Schaffer 1978), Memory-Based Reasoning (Stanfill & Waltz 1986), Derivational Analogy (Carbonell 1986), Case-Based Reasoning (Riesbeck & Schank 1989), Analogical Modelling (Skousen 1989), and so on, though interestingly this connection is only rarely made in EBMT articles, and there has been no explicit attempt to relate the extensive literature on this approach to Machine Learning to the specific task of translation, a notable exception being Collins' (1998) PhD thesis.

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; Wang & Waibel 1995; 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 (e.g. Vogel et al. 1986; Chen & Chen 1995; Wang & Waibel 1997; etc.). 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 often linked with the related technique of "Translation Memory" (TM). This link is strengthened 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. Some commentators regard EBMT and TM as basically the same thing, while others – the present author included – believe there is an essential difference between the two, rather like the difference between computer-aided (human) translation and MT proper. Although they have