[MT Summit IV, July 20-22, 1993, Kobe, Japan]

After Linguistics-based MT

Jun-Ichi TSUJII

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

1. Linguistics-based MT

The 80s can be characterized as the era of Linguistics-based MT (LBMT) and of its failure in the history of MT, in which (computational) linguists have initiated the first serious attempt at constructing scientific or computational theories of MT. Partly because of a large discrepancy between scientific interests and engineering practices, this work has little influence on the performance of commercial MT systems in the market. Because of this failure of commercial exploitation of LBMT, the pendulum has swung suddenly to the other extreme of the spectrum and researchers are now interested in developing, **without any theory,** practical systems which can readily be used in actual translation environments. Some have even abandoned the idea of machine translation altogether and switched to development of tools for translators, including those of retrieving translation examples from data bases.

However, it is my contention that a computational theory of MT (or translation in general) remain essential in all engineering attempts of easing translation loads by using computers, which include not only high quality MT but also more intelligent tools for translators, and therefore that the programme of research initiated by linguists should not be abandoned.

2. Deficiencies of LBMT

LBMT has made a set of interesting proposals which future research into the computational theory of MT should take seriously, such as **Compositionality of Translation, Distinction of Possible Translation and Plausible Translation, Modularity of Mono-lingual and Bi-lingual knowledge, Expressivity of Formalisms,** etc. However, while it has established a set of these goals or pre-theoretical criteria by which the adequacy of individual MT formalisms or theories have to be assessed, LBMT has failed in proposing actual theories or formalisms which satisfy them and at the same time are viable as engineering frameworks in actual translation environments.

The most crucial of all is that linguists in LBMT have placed excessive importance on mono-lingual theories and largely ignored bi-lingual counterparts. As a result, their theories of MT become mere parasites of mono-lingual theories, while ideal theories of MT, to my mind, should center around a bi-lingual theory and reconstruct mono-lingual theories accordingly. Furthermore, the mono-lingual theories they have adopted are those of the generative paradigm in a broad sense, which have exclusively focused on syntax and largely ignored other linguistic phenomena by claiming that they are performance-related. Because translation highly depends on semantic and pragmatic issues, the theories in LBMT which heavily reply on such mono-lingual theories fail to address a large portion of translation problems.

3. Directions of Future Research

These deficiencies of LBMT, however, do not mean in any sense that their attempt is futile. It may only mean that the theories they have developed are not **sufficient** as a theory of MT. They are most likely to constitute **necessary** parts of a computational theory of MT. What we have to do is to construct a more comprehensive theory based on their achievements, but not cancel them to re-start from scratch.

The following are the issues which I think are particularly important in future research:

A. Empirical Study of Translation: Through the long history of theoretical linguistics in the generative paradigm, linguists have crystallized a set of problems which they (syntacticians) address in their theories, such as long-distance dependency, control phenomena, etc. However, we have not yet succeeded in formulating such a set of problems in translation, except for a set of patchy problems like **head-switching,** etc. We have not even accumulated enough empirical data in our field which shows what happen in actual translation and which constitutes the basis for formulation of problems to be solved by theories. Recent initiatives in bi-lingual corpus collection will alleviate the difficulties. First observation, then comes theory construction.

B. Architecture of MT: It has become clearer that the traditional distinction between **Transfer** and **Interlingua** does not grasp the essence of differences in actual MT systems. The same term **interlingua,** for example, has been used for two, essentially different, types of representation schemes. One is a representation of understanding results which is in its nature extra-linguistic and which often depends on individual subject domains. The other is a result of linguistic abstraction and independent of individual domains. It is also clear that the traditional distinction confuses the level of representation through which translation is performed, with the level to which the translation process is able to refer. In order to cope with diverse architectures like Statistics-based MT, Example-based MT, Knowledge-based MT, etc., we have to have a set of theoretical criteria by which we can see essential differences of proposed architectures. This will be the first step to discuss merits of proposed architectures and to integrate them, if possible, into a coherent system.

C. Context and Knowledge: In LBMT, effects of context and knowledge on translation have scarcely been studied. While these issues have been addressed in KBMT, it seems that KBMT tends to focus on their dynamic effects on translation and emphasize the importance of understanding of texts themselves. However, as corpus-based research has revealed in various fields of NLP, much simpler mechanisms than those for understanding and reasoning, such as statistics, can be used even in MT. We need a proper classification of what we call context and extra-linguistic knowledge. Some context effects, for example, can and should be treated in a knowledge acquisition phase, but not in the translation phase. What we call extra-linguistic knowledge also has a broad spectrum. On the one end, there are types of knowledge like knowledge about space and time which are generic and subject domain independent. One the other end, there are types of knowledge which are subject domain specific. These differences should be treated appropriately.