MT FROM THE RESEARCH PERSPECTIVE

Zaharin Yusoff
Computer-Aided Translation Unit
School of Computer Sciences
Universiti Sains Malaysia
11800 PENANG, Malaysia

Abstract

There has been a wide range of research and development work on machine translation in terms of goals and objectives as well as emphasis. Some are fundamental, some are engineering based and some are product-driven.
Different researchers may be motivated by different reasons, some by funding, some by potential commercial returns, some by challenges on the application of various technologies and some by sheer search of knowledge.
Trends also tend to emerge in terms of the 'best' approach at a given point in time. This short discussion advocates the possibility of synergising among all types of research while working towards different goals as opposed to looking for the best direction(s) to follow.

Introduction

Work on machine translation (MT) has been evolving continuously. What started as attempts at automating dictionaries and reshuffling words led to the development of language analyzers and generators backed by deep research in linguistics and formalisms. Some moved away from fully automated translation into the development of machine-aided human translation (MAHT) systems, and others abandon linguistic based approaches to take on corpus based methods. Artificial intelligence (AI) techniques are incorporated and much work has been put into engineering combinations of various approaches. Delivery systems vary from simple dictionary look-ups to sophisticated knowledge based systems with voice recognition capabilities, while many systems have been remodeled to capitalize on the internet service market. Whatever the case may be, one cannot help but sense that trends do emerge and that those who do not follow the current trend at a given time tend to be cast aside.

It is indeed very timely that researchers should now take a step back and look closely at what they have been doing and in what direction(s) they are progressing. More so, one should be asking various questions such as whether they are indeed doing research or simply engineering products for the market. On one hand one may be asking whether there is really a need for fundamental research in MT while on the other hand one may ask whether one should succumb to pressures to produce a full blown system. Perhaps as opposed to looking for the direction(s) for all to follow, one should be looking for ways and means to synergise the results from all possible directions. This way there will not be such a thing as a trend and thus researchers would be free to move in any direction in search of knowledge without having to feel obliged to fit in with a current movement. These are precisely the questions that will be explored in this paper, without of course having any pretensions to providing the right answers.

What is research?

This is perhaps the first question to ask before embarking on the main discussion on MT from the research perspective . Many ways have been put forward for classifying research, but this is certainly not the forum to discuss the matter. Of the many, the following three major categories are usually mentioned and are probably the most relevant to work in MT:

Fundamental research
This is where the main objective is to further knowledge. In MT terms, work that fall under this category would be on the various theories, such as linguistic, translation and knowledge representation, as well as on the supporting models, hence also grammar formalisms. The underlying theme here is the understanding of how and why certain things work the way they do.

Engineering research
The main objective here is to perfect methodologies and their combination within a particular goal. Corpus based MT approaches tend to fall under this category and so do work on certain MAHT systems. Parsing and generation would also be here and so would many AI techniques introduced to MT. The underlying theme is to make each module in a system work as best as it can in terms of speed, accuracy and reliability.

Product-based research

The ultimate aim is to produce complete systems that work and are also marketable. Ideally the systems should result from fundamental and engineering research but it is not a major condition. User acceptance and contribution to productivity are the main concerns. The major part of work on MAHT systems falls under this category as well as the usage of various supporting technologies to provide translation services.

It is not the aim here to argue for the best type of research for MT because they all have their roles to play. In fact, the history of MT records various phases with each phase emphasizing a particular type, but then again that may have been due to certain beliefs at that point in time rather than a conscious effort in choosing the approach. The direct approach in MT was an engineering approach but that was probably the best approach given the technologies at the time. The indirect approaches (transfer and interlingua) were well thought out approaches supported by many theories and hence provide good examples of fundamental research but the end results were rather disappointing as MT systems. AI based systems are further examples of fundamental research work while corpus based approaches are more engineering based, but arguably neither produced much better results. Product research type systems have always been around taking whatever results offered by the other two types and adding practical functionalities to form usable systems to be tried out in the market. Although there have been a few claims to the effect, it remains very difficult to attribute major success to any MT system produced thus far.

Must there be a product?

Given the situation described above, this is a major question to ask. One recognizes the fact that all MT work needs to be financed and that most sponsors tend to insist on complete MT systems as the end result. This is despite the knowledge that the same scenario has been played over and over again since the beginning of MT history, where ideas were put forward for the development of a new MT system based on newer theories and technologies but the end results (the translation quality) never really improved from the previous one.

Before coming to a premature conclusion that MT may not be feasible or viable, it has to be first recognized that MT is probably the longest running application domain in the history of the computer. Indeed the domain has produced as by-products many other technologies, including (among others) compilers, logic programming and lexical databases within the computational domain as well as various theories and models that led to applications (commercial or otherwise) in many other domains. MT is perhaps the most encompassing and the most challenging amongst all application domains, albeit rather expensive as an investment. To some extent, comparisons can be made with research on space travel where the by-products total up more contributions than those obtained from the main goal, and that the main goal is very long term, one that may not even be attained in a satisfactory and economical way on a large scale.

MT should be viewed in this light, namely as an application domain where theories, methodologies and technologies may be tested with some long term hope of producing a MT system but certainly not short of by-products. This also means that sponsors should also view MT in the same manner and that investments are to be made for long term objectives with by-products as short term benefits.

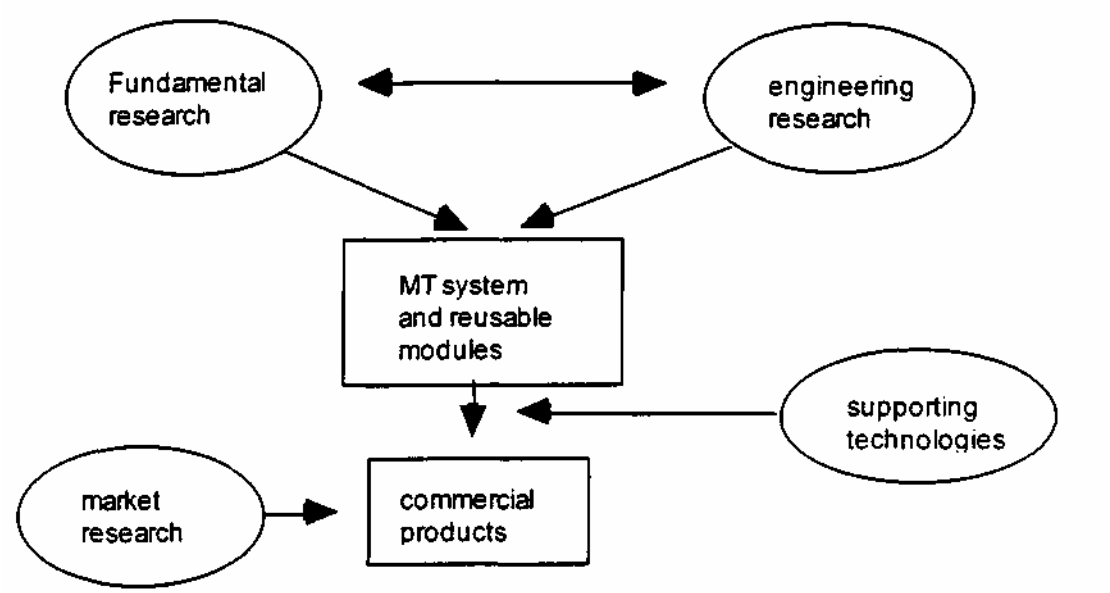
An ideal situation

At the risk of being branded archaic, one gets the feeling that fundamental research in MT may have not been explored fully yet. Linguistic models are still to be perfected at the syntactic and semantic level and extended to include discourse and pragmatics. Knowledge representation and manipulation do have a lot of potential and all these may still lead to the elusive interlingua. Grammar formalisms can still be improved and so do their parsing and generation techniques. Research in all these areas may not lead to viable MT systems in the short term but it does provide a better understanding of language and translation and hence may be benefited by the engineering approaches. Besides, such a wealth of knowledge is bound to lead to many by-products.

Engineering approaches do have their merits especially in view of obtaining a certain level of results without having to mount a strong team of linguists and lingware programmers required by the more fundamental approaches. There is no denying that engineering research also furthers knowledge especially in the domain of techniques (statistical methods, neural networks, etc.). Data is widely available and should be fully utilized.

Along the same lines, products need to be developed and relevant supporting technologies should be made use of. After all, MT is an applied domain and hence the end result must ultimately be a MT system. Efforts in building independent modules to fit into a plug-and-play mode should be very much encouraged as they do lead to many by-products. Morphological analyzers/generators, syntactic and semantic analyzers/generators and transfer/interlingua modules as well as the lexicon are completely reusable in the development of products other than within a MT system.

An ideal situation is one where all the types of work mentioned above can be continued by experts in the respective areas. The thing to look for is some form of synergy where the results, be it in terms of modules/products or knowledge, can be utilized by each other. Such a synergy may be depicted by the diagram below.



The diagram above also includes market research, which is one area that is rather neglected within the MT community. Researchers tend to make assumptions on the needs of users without conducting proper research to substantiate their claims. Users may be translators/service providers or translation clients and each may have differing requirements and these may vary form country to country and from industry to industry. There is indeed a need to understand the market better in order to set the proper directions in MT research.

The synergy that is sought here should be of tight coupling. Surely there must be ways that corpus based MT research can take advantage of linguistic knowledge, and similarly linguistic models need to be substantiated by large scale corpus studies. An automatic generation of linguistic rules from a corpus is an interesting possibility and so is an automatic generation of parallel text matching algorithms based on linguistic rules. One possibly interesting direction is the development of libraries of independent modules with highly specialised tasks and declared input-output specifications that can be put together in different combinations to obtain different products and applications. This will fit very well with advances in network and internet technologies. It will also allow different theories and techniques to be explored within the specified specialised tasks and hence contribute to the advancement of knowledge in general, which is after all the raison d'etre of research.

Concluding Remarks

Having said all the above, it has to be noted here that there is little wrong with the happenings in the MT community today. One is simply hoping that some emphasis could be placed on fundamental R&D, albeit a small segment, and that researchers should be encouraged to explore all other avenues in search of knowledge, this being as opposed to full concentration on product development. One way of attaining this is to encourage the development of independent modules with highly specialised tasks that can be combined in different ways to produce different application systems. Such a scenario will help advance knowledge in MT and related domains as well as result in many by-products beside satisfying the needs of all types of researchers. Sponsors should take note of this and learn to capitalise on the situation.