Machine Translation - How do we succeed?

K. Narayana Murthy
Department of Computer and Information Sciences,
University of Hyderabad,
Hyderabad, 500 046,
email: knmcs@uohyd.ernet.in

Abstract

In this paper we explore the feasibility of developing High Quality Machine Translation systems in the Indian context. We take a look at the current state of the art, the challenges ahead and outline a road-map to achieve usable machine translation systems.

Keywords: Machine Translation, Automatic Translation

1 Introduction

Automatic Translation was one of the first major application domains in modern computer science and linguistics and it continues to be one of the most talked about areas. Several centres have been working on Automatic Translation systems in India, for more than a decade and a half now, and demonstration level prototype systems have been developed. Yet we are far from achieving a break-through. There are hardly any systems that are really being used. The purpose of this paper to sketch the current state of the art, discuss the challenges ahead and then outline a plan of action to achieve usable machine translation systems within a reasonable time frame.

2 Why Machine Translation is Hard

The first question that is invariably asked in any discussion on machine translation is why is it a hard problem. We know it is hard and we wish to know exactly why. There are several aspects that make automatic translation inherently difficult, if not impossible. Here are some:

- *Lexical Ambiguities*: Is the word 'like' a verb, an adjective or a preposition in a give sentence
- *Structural Ambiguities*: Prepositional Phrase Attachment, Subordinate Clause Attachment: 'I saw a man on the hill with a telescope', 'I gave the book to the boy who had come home after taking bath', 'mothers with babies 6 months old' versus 'mothers with babies more than 40 years old'
- *Parsing Limitations*: The performance of parser-based translation system is limited by the performance of the current syntactic parsing systems. The best available parsers are not good enough.

  There are no wide coverage computational grammars for any Indian language yet. There are no syntactic parsers. How long can we keep wishing away the necessary technological foundations?
- *Difficulties in lexical substitution*: Non-availability of suitable substitutable equivalents in the target language, difficulties in extending the meaning of existing words, difficulties in coining new words
- *Identifying Correct Sense of words*: 'follow' (Vivekananda's ideals in your life, a thief who is running away with your purse, a lecture in a classroom). Is capital a capital city, the financial capital or an upper case letter of the alphabet? Word Sense Disambiguation has remained a hard nut to crack. People use world knowledge and commonsense. Machine are poor at these.
- *Anaphoric References, Discourse Coherence* It is not sufficient to work with one sentence at a time. Discourse level analysis is essential to ensure good translation.

- *Differences between Source Language and Target Language Structure* Non-availability of suitable mappings, less information encoded in source language than what is needed, what to do with extra information available in source language that cannot be encoded in the target language, etc.

It is clear that fully automatic high quality translation is difficult to realize in practice, if not completely impossible. Despite the best brains working hard for 15 years or so, there is no system in regular use today in our country. Inherent difficulty of machine translation task is not the only reason. There are other very important reasons as well:

- *The Need*: NLP researchers in India have always been so obsessed with the idea of Machine Translation that one starts wondering why Machine Translation? Who wants it? Who are the potential users, potential beneficiaries? We often find people starting off by saying literary translation is very difficult. Of course it is. But why should one even try to automate literary translation? Don't we human beings enjoy doing literary translation ourselves? Is it something so routine, boring and tedious that we want machines to take over? Ordinary people are a lot more likely to find information retrieval, automatic summarization, text categorization applications such as email filtering or even spell checking tools much more useful on a day to day basis than machine translation. How often have you felt the need for a tool to automatically translate some documents? Has any systematic survey of MT applications and potential beneficiaries been conducted?

  The purpose here is not to belittle the importance or usefulness of machine translation. It is just to state that researchers somehow do not seem to know very clearly why they are doing what they are doing. Any technology developed without identifying potential users, involving them from day one and keeping their requirements and expectations in mind while designing the system is not very likely to succeed.

- *User Expectations*: People find it perfectly fine if an Information Retrieval system or a Search Engine gives only 40% performance but when it comes to machine translation, even a small deviation from the expected is completely unacceptable. Nothing less than perfect is OK.

Before we start working on a project we must have a clear idea of what is expected and whether that is achievable.

- *Non-availability of adequate lexical resources*: parallel corpora, suitable electronic dictionaries, thesauri, word nets, annotated corpora (part-of-speech tagged, parsed, sense-tagged)

- *Bad Planning*: Knowing very well that high quality translation is not feasible in the current situation, the focus should be placed on developing the enabling technologies and data resources before we take major initiatives in machine translation per se. Our time, effort and money should be channelized based on a realistic plan of action. There are no computational grammars or parsers for any of the Indian languages. Lexical resources are grossly inadequate. Machine translation is clearly a pre-mature initiative in our country.

- *Who should be doing it?* Hardly any translator is using (or trying to use) any machine translation system and the experts developing machine translation system rarely if ever have done any translation themselves.

- *Living in our own world* We researchers often live in our own dream worlds imagining all kinds of things that users need, what they like and what they don't. Sometimes we have assumed that the users are willing to do pre-editing. Users usually do not like to do any pre-processing of the text. If they have to manually read and take some actions, why not do the translation also manually? Often we expect users to understand a great deal of our technology. Who cares? Expectations placed on the users have often been completely unrealistic.

- *Choosing the wrong problem* We spend loads of time, effort and money to develop a system for translating a certain kind of documents from a certain source language to a target certain, only to discover that there is no need for that and there are no takers at all.

- *Ad-hoc work culture* Lack of systematic approach to development: In most cases formal specifications, design documents, test plans, test data, performance evaluation strategies, independent and unbiased performance evaluation etc. are either not there at all or poor and ad-hoc at best. What do you mean one says a machine translation system gives

85% performance? Are 85 chapters out of 100 in the book perfectly translated? Are 85% of the senetnces OK or 85% of the words OK? Anything less than near 100% may be worse than useless in some practical situations. Evaluating the performance of MT systems is itself a very difficult task.

- *Failing to understand the difference between science and engineering* Researchers wish to pursue scientific inquiry. Developing a product in another thing. Good engineering judgment is essential. A well engineered product will be a bigger success even if it rests on an imperfect theory.

3 MT or not?

The question that arises then is whether machine translation should be attempted at all or not. The benefits of a usable machine translation system are well known. While human costs are going up, machine costs are coming down. Human translators are not always available where and when you need them. You can make multiple copies of an MT system and use it simultaneously at many places whereas one human expert can do only one thing at a time. The machine does not get bored or tired, and it does not complain if asked to work 24 hours a day. The translation load is increasing every day and we will never be able to meet the demands only through human translators. MT can thus save a lot of time, effort and money. But to realize these benefits, the performance of the system must be very high. Poor quality systems can be worse than useless - we may be better off doing the translation by hand. Therefore the question is not whether we should work on automatic translation or not, the question is how do we identify the right application domains and how do we go about developing high quality automatic translation systems for those domains. One must also keep in mind embedded applications, not just stand alone MT systems.

4 A plan of Action

The task at hand is large and complex. It must be done in phases. There are many subtasks. Different groups can and should work on various subtasks

in a coordinated fashion. The tendency to say our group will do everything should be avoided at all costs. A task as large and complex as machine translation can only be a national initiative.

In the first phase, spread over say 2 to 5 years, several different kinds of activities can be done in parallel. One stream will focus on the development of data resources - parallel corpora, alignment of parallel corpora, large electronic dictionaries, morphological analyzers and generators, computational grammars etc. A thorough investigation of how words, phrases and expression are mapped from one language into another needs to be carried out. Linguistic inquiries into issues like how negation, relativization and participial constructions, conjunction and ellipses need to be undertaken. During this phase, tendency to jump into building machine translation applications must be avoided. Success depends upon not how much we do but how well we do. Appropriate engineering practices must be strictly followed.

In parallel, enabling technologies such as part-of-speech tagging, robust parsing and word sense disambiguation should be taken up with automatic translation as a major application in mind.

Another stream could focus on a theoretical and empirical front. Translation theories and strategies can be explored. Experimental and exploratory work can be taken up.

Yet another stream may take up market surveys and feasibility studies. Business models can be developed, killer applications identified, users trained.

Only then will we be ready for taking up any major initiatives in machine translation. Potential application areas must be identified, potential users must be included in the team and specifications discussed and written down. Overall and detailed design documents must be prepared, discussed and approved. Expected performance, performance criteria, evaluation strategies and methodologies, test plans, test data generation etc. must be worked out in detail. Only after that should implementation start. Regular monitoring should be carried out and mid-course corrections effected if and where required. Including potential users is a key ingredient in ensuring the success of the system in terms of user acceptability. Suitable standards must be fol-

lowed at every step. Where required, standards must be developed.

5 How do we go about doing all this?

- MT should be taken as a major national initiative. The big picture must be drawn carefully so that all the small parts nicely fit in. One group should take the overall responsibility for integration, testing, evaluation, documentation, training and manpower development, deployment and technology transfer, maintenance etc. Full cooperation of all the groups is essential. A chain is as strong as its weakest link. We cannot afford any weak links.

- Do just one small thing at a time. There are only a few active groups and everybody wants to do everything. Let one group develop a good bilingual dictionary between a specified pair of languages. The dictionary must be developed in accordance with the specifications for content, structure, organization, formats and standards. No work should start before a detailed, formal written specification is evolved with the active involvement of all relevant experts. A whole lot of work done so far in our country is gravely wanting in many of these counts. Whatever we do, we must do it well, very well.

6 The MAT2 Initiative

There are certain tasks that are best done by the man and there are others that are best done by the machine. We often fail by asking the wrong person to do a given thing. Fully automatic high quality translation is not feasible until and unless in depth understanding of natural languages in particular and the whole of human cognition in general become possible. The trick therefore is to get the best of both worlds. High quality translation can only be achieved by establishing a clever synergy between man and the machine.

The MAT2 initiative [Murthy1999, Murthy2002] at University of Hyderabad was taken up based on the experiences in developing an English-Kannada machine aided translation system for the Government of Karnataka

- a real life application funded directly by the user agency. MAT2 has several objectives. Firstly, it aims at developing an architecture for high quality translation between English and Indian Languages. High quality translation cannot be a fully automatic process. Translation in MAT2 is going to be interactive. It is not intended as a product or a system for routine use - instead it is a research environment for translation. Secondly, the byproducts of MAT2 will be very useful resources - high quality, POS tagged, sense-tagged, parsed and aligned parallel corpora. Thirdly, MAT2 architecture will enable exploration of combining linguistic and statistical evidence, of man-machine synergy, of learning and adaption in translation. Along the way, lexical resources are created and those available are tested and refined. Hopefully, this initiative would finally lead to the design and development of real Machine Translation systems.

7 Conclusions

In this paper we have attempted to explore the reasons for our failure in achieving success in automatic translation in our country. Some suggestions are made for successfully developing and deploying translation systems for practical use. It is hoped that at least some of the issues brought out here will be discussed and debated widely leading to a clearer picture of the machine translation field in our country.

References

[Murthy1999] K Narayana Murthy. MAT: A Machine Assisted Translation System. In *Proceedings of the NLPRS-99 Fifth Natural Language Processing Pacific Rim Symposium, Beijing, China, Nov 5-7*, 1999.

[Murthy2002] K Narayana Murthy. MAT2: Enhanced Machine Aidedd Translation System. In *Proceedings of the STRANS-2002 Symposium on Translation Support Systems, 15-17 March 2002, Indian Institute of Technology, Kanpur, India*, 2002.