25. Strategies for Development of Machine Translation Systems

Ch. Ram Anirudh
Prof. Kavi Narayana Murthy
School of Computer and Information Sciences, University of Hyderabad, Hyderabad.
<ramanirudh28@gmail.com, knmuh@yahoo.com>

Abstract

In this paper we describe the state of the art in Machine Translation. We include a critical review of the challenges faced in MT and the possible reasons behind the failures. We give a number of suggestions for overcoming these limitations and challenges. We support our claims by describing several MT systems we are currently developing. Our approaches hold promise and raise hopes. We believe that the strategies suggested here are applicable to other Indian languages as well. We can develop usable MT system within a very short time and gradually improve the systems as we keep using them.

1 Introduction

Translation is a meaning preserving transformation from one language to another. Machine Translation (MT) or Automatic Translation deals with the design of computational models for translation between human languages. MT systems usually do not attempt to directly capture and preserve meanings, instead, they try to capture and transfer structure, in the hope that structure captures meaning. Thus a natural model for MT is to do analysis of source language (SL) text, and generate the corresponding target language (TL) text, preserving the structure. MT systems normally work sentence by sentence. Morphological analysis and generation to take care of word-internal structure and syntactic analysis and generation to take care of sentence-internal structure are thus natural considerations. A transfer module may be incorporated between the analysis and generation phases to take care of divergences between the two languages.

An alternative view that has become popular in recent times is the Statistical Machine Translation (SMT) view point [5]. Translation involves finding TL words corresponding to the SL words in a given sentence, and re-ordering the words if required to take care of syntactic divergences. SMT models these aspects in a probabilistic fashion. During the Training phase, these probabilities are estimated from a Training Corpus. Then the system can apply the learned models to translate given texts. A large and relatively high quality training corpus is essential for SMT. There is no need for any dictionary, morph analyzer or a parser, nor do we need to explicitly model any divergences.

Of late highly successful MT systems have been built using Deep Learning Neural Networks. These are similar to SMT systems in that they are trained on a large training data set and linguistics is not explicitly used.

In the next few sections, we shall review the state of the art in MT, identify the weak points and suggest strategies to overcome these weaknesses. We shall then mention our own work in MT where we have fruitfully applied some of these suggested strategies.

2 Machine Translation: The State of the Art

Machine Translation is at least 60 years old. Until about 1990, most of research and development effort in MT was in the Rule-Based approach. Included in the Rule-Based Approach are the Direct, Transfer and Interlingua approaches. In hind sight, it would not be wrong to say that MT in general has not been very successful, for any language pair, anywhere in the world. Truely successful systems, which are able to capture and preserve meanings and produce publishable quality translations, are very few and limited in domain and applicability.

End-user expectations are very high and it looks next to impossible to achieve the expected degree of perfection by fully automatic means. In the initial stages, it was felt that the translations produced by the machine can be post-edited by humans to produce high quality translations. However, unless the quality of outputs produced by machines are very good, humans will prefer to translate on their own rather than checking and correcting all the mistakes the machine has made. Psychologically, post-editing is not a very pleasant task [5]. As a result, MT systems rarely came to fully usable level. MT could only be used in those situations where rough translations suffice, and publishable quality translations are not required.

There are many reasons for this failure. Firstly, adequate linguistic resources are not available in many languages. Large, representative, and clean corpora, good dictionaries, good grammar books etc. are not easy to find in all languages of the world. A dictionary meant for human use is very different from the dictionary we need for automatic use by machines and many ’good’ dictionaries may still not be good enough for use in MT. Similarly, many works on grammar are not sufficiently detailed and sufficiently precise for implementation on a computing framework. As a result, MT developers end up also doing a good deal of original lexicographic work and grammar discovery. Needless to say, these are hard tasks requiring decades of research work. Also, good linguists are not always available for consultation. More importantly, linguistics is not a finished science, we cannot expect ready-made answers to all the questions we get while developing MT systems. There are many competing theories, theories are constantly undergoing developments and refinements, many problems of critical importance in MT have still not been solved fully and satisfactorily. Divergences between different languages or language families have not been fully worked out, making the transfer stage a big challenge. MT developers start with great hopes, assuming that languages are rule-governed and the rules can be easily discovered. They encounter harder and harder problems down the line, which can be quite frustrating. A point of saturation is reached and further improvements become very difficult.

Statistical approaches to MT, on the other hand, require large, high quality parallel corpora. Such corpora are available for some language pairs in the world today - English-French, English-German, English-Chinese, for example [4]. A lot of progress has been made in SMT taking such language pairs for study. However, for languages where such large scale parallel corpora are not yet available, these theoretical results are only of academic curiosity, not useful in actually developing usable MT systems. The same observations are also valid for the Deep Learning approaches.

Unless the quality of translations produced by a machine is very good, the outputs cannot be post-edited to produce large scale, high quality parallel corpora. Thus we are left in a chicken-and-egg kind of a situation.

3 MT in India

Interest in MT started in India as early as 1986. [6] MAT [8], MATRA and MANTRA are some of the noteworthy MT systems developed during the 1990s. [3] During the year 1990-91 DIT (Department of IT), Govt. of India initiated the TDIL (Technology Development for Indian Languages) programme. [3] Several major projects have been funded by TDIL since then in MT and related areas. IL-IL-MT, E-IL-MT, Anglabharti-E-IL MT [1] are some of the major projects funded by TDIL in recent times in consortium mode. All these projects generally use rule-based approaches. Shatanuvadak [5] by IIT-Mumbai in 2014 is a notable deviation, it uses SMT in a big way for Indian languages.

Third party evaluations have shown that none of these MT systems have reached a level of performance adequate for deployment and large scale regular use. People are hesitating to come forward to post-edit the translations produced by these MT systems. We do not hear of any big success story.

It appears that everybody is interested in cooking and nobody wants to eat what is cooked. In the next section, we propose a set of ideas to overcome this unpleasant situation.

4 Strategies for Development of MT systems

In this section we analyze the main reasons for not being able to reach the final goal of high quality translations, and our own suggestions for overcoming these challenges. In the next section, we shall describe ’The saara Translator’, a set of MT systems being developed by us, that actually try to put these ideas into practice.

Two main tasks an MT system has to do is lexical substitution and re-ordering. SL words need to be substituted with equivalent TL words. We may then need to reorder the words as dictated by the syntactic divergences between the SL and the TL.

Since parallel corpora are not available for many language pairs of interest, we shall restrict our attention to rule-based approaches here. A common belief is that morphology is absolutely essential, especially in Indian languages, which are considered to exhibit rich morphology. That is, a single root word can give rise to a very large number of word forms, through processes such as inflection, derivation, sandhi and compounding. However, developing high performance morph systems has not been easy. Can we bypass morphology?

Generally speaking, it is not an intelligent decision to throw away morphology and try to keep all forms of all words directly in a dictionary. Morphology has its due share of relevance and importance. However, developing proper systems for morphological analysis and generation is a hard task, requiring many years of labour. There are theoretical problems, there is the challenge of finding good linguists, even good books on grammar may not be available, and practical experience shows that approaches to MT which critically depend on morphological analysis, transfer and generation have not worked very well in India. Therefore, while development of complete morphology in a computational framework should remain an important goal, we should keep that on the back burner as a long term strategy and think of developing MT systems that bypass morphology. Our own experiments show that equal or higher performance in translation can be obtained much faster by bypassing morphology. MT systems can be built within a matter of months, taking advantage of ordinary people who are bilinguals, instead of waiting for linguists or try on our own to develop computational systems of morphology. We find that bypassing morphology actually works even for Dravidian languages, which are considered to exhibit exceptional levels of morphological complexity.

Of course we must take good advantage of a system for morphological analysis and generation if we already have one. We can build hybrid MT systems which combine the best of rule-based and statistical approaches. Performance of a hybrid MT system will improve not only with larger and better training corpora that may become available over time, but also as morphology improves over time.

In the simplest statistical MT model, we start with the assumption that all lexical substitutions are equally likely. To give an example, any word in English can map to any word in Hindi. The English word ’table’ may mean ’maa’, ’khaaya’, ’us’, ’mej’, ’idhar’, ’kutta’ or any other word in Hindi, all of these are equally probable. SMT systems then try to adjust these probabilities based on a large training corpus of English-Hindi sentence pairs. In other words, pure SMT systems do not take any advantage of the linguistic knowledge we may have. Pure SMT systems do not use dictionaries, morph, syntax or any other aspect of language and linguistics, even if we have access to such knowledge. That is why they need a very large training corpus. Instead of waiting for a large training corpus of parallel sentences, we can get started off if we make good use of available resources such as bilingual dictionaries. Lexical substitution possibilities are greatly reduced and so we can start seeing good MT performance even before we have any training corpora.

Thus the use of a word-for-word substitution dictionary makes sense both from the rule-based view-point and the SMT view-point. Development of such dictionaries should therefore be given the highest priority.

Traditional SMT systems combine lexical substitution and re-ordering, both of which are learned together from the training corpus. This makes SMT a lot more complex than it needs be. By separating the lexical substitution task from the reordering task, we can greatly simplify the system, both in terms of training corpus requirements and overall simplicity and efficiency. Syntactic divergences can be handled through a transfer grammar if the divergences are already clearly known. Otherwise, while such divergence studies can go in the background, we can develop and use pattern-matching ideas to discover and apply rules of re-ordering. Purely statistical methods will become feasible only after large scale parallel corpora have been developed.

We need to collect large, representative and clean corpora. We need to perform quantitative analysis at each stage, always trying to exploit the exponential nature of distributions we find in the statistics of linguistic material. We need to adapt sound engineering principles. We must base our work on solid theoretical foundations and avoid the tendency to take short cuts. See [7] for more on the theoretical foundations of language engineering.

To build an MT system quickly, we just need to develop a large database of word-for-word substitutions. We do not need any morph, nor do we need a POS tagger, local word grouper or a chunker to start with. We can start getting promising performance in translation very quickly. We can translate at great speed too. We need to carefully observe the outputs and enhance and improve the databases. Once we have sufficient data and sufficient experience, we can then start addressing the remaining research issues one by one.

One of the most widely held beliefs not only in MT but also in the whole field of Natural Language Processing (NLP) is that human languages are highly ambiguous. In fact, disambiguation is considered to be the main focus and priority. Words have multiple senses, they even belong to multiple grammatical categories, there are multiple ways of grouping words to form higher level structures. The degree of ambiguity is claimed to be very high, researchers have talked of hundreds, thousands, Millions, even Trillions of possibilities. The claim is that simply using a dictionary of equivalents will not work.

Upon careful analysis, we find that this is not true. Human languages cannot be so very ambiguous, otherwise, seven and a half Billion people would not be doing their daily business in natural languages. The exponential nature of ambiguities in natural languages is a result of lack of proper understanding of what a word is. We take the written form of language too seriously and we simply go by what we see in a piece of text. Words are taken to be sequences of characters (letters of the alphabet, punctuation marks, special symbols etc.) separated by white spaces. This is not right.

Of critical importance in our approach is the notion of a word. What exactly is a word? We have shown [9] that there is a much better way of defining words, starting from meanings, rather than from spellings. By re-defining words in a proper way, a great deal of ambiguities in languages melt away automatically. Computational complexity is also reduced significantly. In our own work, we find that most words are not ambiguous at all. In the case of ambiguous words, the degree of ambiguity is small. There are also simple ways of disambiguating the real cases of ambiguity.

MT is often projected as a product. This is the big problem. We may never be able to bring MT to a level where end users can directly use it as a ready-to-use product. We have not been able to reach that stage in the last 30 years in India. Instead, MT should be considered as a service. The user submits his translation requirements to a service provider. The service provider runs an MT system and looks at the machine generated output carefully. He may add more entries to the MT databases, he may correct the errors found in the databases if any, he may even edit and clean the input SL texts for the purposes of translation. He may run the MT system several times, each time improving the quality of translations produced as also the MT system itself. Finally, he may proof-read and manually post-edit the machine generated outputs to produce publishable quality translations. We can take good advantage of a synergy between the man and the machine, and semi-automatically produce high quality translations. Our experiments show that this is economically viable and practically feasible too.

Some people argue that high quality is not always essential in translation. Instead of taking this ’sour grape’ attitude, we can actually start producing high quality translations by giving up the claim of making it fully automatic. Most successful engineering systems in the world use the best of both the man and the machine. People are ready to do their bit, if only we can guarantee high quality and economic viability. Reasonably good translation performance is possible by using a dictionary of word-to-word mappings, and re-ordering the TL words as required. After reaching this milestone, further improvements can be made through disambiguation rules etc.

Reasonably good translation performance is possible by using a dictionary of word-to-word mappings, and re-ordering the TL words as required. After reaching this milestone, further improvements can be made through disambiguation rules etc.

5 The saara Translator

Based on the saara theory [7] and the ideas and strategies discussed above, three different MT systems are being developed at the School of Computer and Information Sciences, University of Hyderabad, known as MT1, MT2 and MT3 respectively. A brief description of each, along with the current levels of performance, is given below. All the three MT systems are for translating Modern Kannada Prose, written in the so called Standard Dialect, into Modern Telugu Prose.

Quality of translation is measured in terms of Comprehensibility[2], that is, whether the meaning of the sentence can be understood by the reader. Comprehensibility is measured by manual evaluation, on a scale of 0-4, 4 being perfect and 0 indicating total failure. A score of 3 indicates almost perfect translation and a score of 2 indicates that the meaning of the sentence can be fully comprehended, albeit with some difficulty. A sentence is considered to be successfully translated if the score is 2 or more. The performance of the MT system is indicated in terms of the percentage of sentences that are successfully translated. This method of evaluation and scoring has become the defacto standard in India in recent times. We report here the translation performances obtained on a corpus of 4.6 lakh sentences, based on sample studies on several sets of 100 sentences each.

The MT2 system is a dictionary based SMT system. It does not use morphology, syntax or any other linguistic modules, nor does it use a parallel corpus. Its main focus as of now is lexical substitution. This system is very fast - it can translate 1,00,000 sentences per second on an ordinary Desktop PC or a Laptop. The database has about 1.2 lakh entries. Translation performance varies from about 45% to about 55% on first run. Once the outputs are checked and the databases updated as required, translation performance jumps to the range of 85% to 95%. Experiments have shown that the quality of translations so produced is acceptable for the purposes of final proof-reading and post-editing. Also, the time and cost of the entire process is comparable to that of manual translation, actually somewhat more economical as on date. The system improves with time and we believe that it will become a strong competitor to manual translation very soon.

The MT1 system, is an Analysis-Transfer-Generation based MT system. A comprehensive computational grammar of Kannada is used to perform morph analysis of Kannada words. Inflection, derivation and sandhi are all handled. Spelling variations are normalized automatically to a large extent. More than 90% of words are analyzed, with less than 10% error. The morph analyzer produces a fine-grained hierarchical tag for each input word. This tag includes all the necessary lexical, morphological, syntactic and semantic information required for further processing. The MT1 system currently uses a simple tag-for-tag transfer grammar. Telugu word forms are generated using a morph generator. Translation performance varies from about 35% to about 55%. It may be recalled that this same or better performance is achieved by the MT2 system without using any morphology. This shows that morphology can be bypassed, even for morphologically rich languages, as a practical strategy to start with.

The MT3 system combines the best of the above two MT systems and achieves a performance of 55% to 72% on the first run. As already indicated, we can cross 90% performance using the man-machine synergy.

These MT systems are already much better than the Google Translator. We are in fact ready to take up small translation jobs. Further work is on to improve the databases in the MT2 system and to improve the linguistic modules in the MT1 system. Automatic post-editing modules are being designed to automatically improve the quality of translations, thereby reducing the time and effort required in final proof-reading.

6 Conclusions

In this paper we have described the state of the art in Machine Translation, we have included a critical review of the challenges faced in MT and the possible reasons behind the failures. We have given a number of suggestions for overcoming these limitations and challenges.

We have supported our claims by describing several MT systems we are currently developing. Our approaches, generally labelled ’The saara Translator’, hold promise and raise hopes. We believe that the strategies followed in the development of ’The saara Translator’ are applicable to other Indian languages as well. We can develop usable MT system within a very short time and gradually improve the systems as we keep using them.

References

[1] Paradigm shift of language technology initiatives under tdil programme. Vishwabharat, Apr - jan 2007.
[2] Akshar Bharati, Rajni Moona, Smriti Singh, Rajeev Sangal, and Dipti Mishra Sharma. Mteval: An evaluation methodology for machine translation systems. In *Proceedings of SIMPLE-Symposium on Indian Morphology, Phonology and Language Engineering*, IIT Kharagpur, India, 2004.

[3] Hemant Darbari, Anuradha Lele, Aparupa Dasgupta, Ranjan Das, Debasri Dubey, Shraddha Kalele, Shahzad Alam, Priyanka Jain, and Pavan Kurariya. Enabling linguistic idiosyncrasy in anuvadaksh. Vishwabharat, July - Dec 2013.
[4] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. MT Summit, 2005.
[5] Philipp Koehn. *Statistical Machine Translation*. Cambridge Univeristy Press, 2009.
[6] R.Mahesh K.Sinha. Man-machine integration in translation processes: an Indian scenario. In Bernadette Sharp, Michael Zock, Michael Carl, and Arnt Lykke Jakobsen, editors, *Proceedings of the 8th international NLPSC workshop, Special theme: Human-machine interaction in translation*, Copenhagen Studies in Language 41, pages 9–20, Copenhagen Business School, 20-21 August 2011. Samfundslitteratur.
[7] Kavi Narayana Murthy. The saara approach to language engineering. Forthcoming.
[8] Kavi Narayana Murthy. Mat: A machine assisted translation system. In *Proc. of 5th Natural Language Pacific Rim Symposium*, Beijing, China, 5-7 Nov 1999.
[9] Kavi Narayana Murthy. On defining word. *IJDL*, XLIV(1):129–161, Jan 2015.

*****