Machine Translation

A Gentle Introduction

Durgesh D Rao

Machine translation is the study of designing systems that translate from one human language into another. This is a hard problem, since processing natural language requires work at several levels, and complexities and ambiguities arise at each of those levels. Some pragmatic approaches can be used to tackle these issues, leading to extremely useful systems. This article introduces the main concepts, issues and techniques involved in machine translation, and looks at some applications.

Durgesh D Rao is a senior staff scientist with the knowledge based computing systems (KBCS) division at the National Centre for Software Technology, Mumbai. His areas of research interest are natural language processing, information retrieval and machine translation.

Introduction

Imagine this scenario. You dial your colleague in Tokyo. You do not speak Japanese, and he does not speak English. Yet, you are able to converse! You speak into the phone in English, which automatically gets translated into Japanese for him. He replies in Japanese, and you hear it in English. Such a scenario is not yet a complete reality[1], but it is a possibility that has excited and engaged researchers in a field known as *machine translation (MT)* for a number of years.

Machine translation is an important sub-discipline of the wider field of *artificial intelligence (AI)*. AI (among other things) deals with getting machines to exhibit intelligent behaviour. As you might imagine, both AI and MT are interesting and challenging fields. In this article, we will look at the important concepts, issues and techniques in MT.

A machine translation system essentially takes a text[2] in one language (called the *source language*), and translates it into another language (called the *target language*). The source and target languages are natural languages such as English and Hindi, as opposed to man-made languages such as C or SQL.

[1] The 'translating telephone' technology is expected to mature in the 2010s. The Janus project at the Carnegie Mellon University is an example of current efforts in speech-to-speech translation (see *Box 1*).

[2] In this article, we assume text-to-text translation. Some applications, such as the translating phone, may need speech input or speech output, or both, but those can be handled by speech recognition and synthesis modules respectively, with a text-to-text translation system in between.

Box 1. Research MT System Example: The 'Janus' Translating Phone Project

The Janus project at the Interactive Systems Lab, Carnegie Mellon University, is working on a set of translation projects. One important prototype is the 'Translating Videophone Station'.

This prototype system allows two users to communicate in a given domain via a videoconferencing connection. Each party sees the other conversant, hears his/her original voice and sees/hears translation of what he/she says as subtitles, captions and/or synthetic speech. The situation is cooperative, that is, both users want to understand each other and collaborate via the system to achieve understanding.

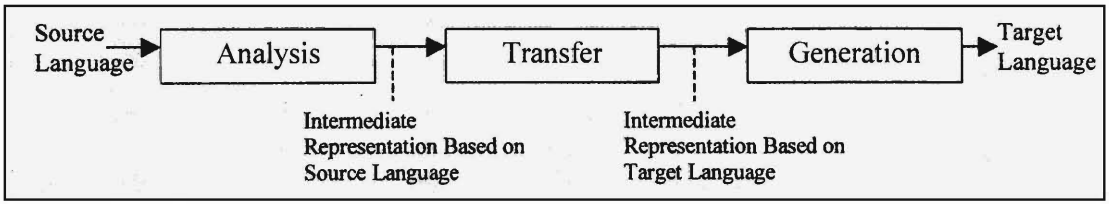
After the record button is activated, the station accepts spoken input and produces a paraphrase of the input sentence first. Once the user has verified that the system properly understood the intended meaning, he/she activates the 'send' button to send a translation of this intended meaning to the other site in the desired language. Various interactive correction mechanisms facilitate quick recovery, should possible processing errors and miscommunications have altered the intended meaning.

Using the technology of this prototype, ISL is also developing prototypes for a portable translation system based on laptops, and simultaneous translation of two speakers in a dialogue.

For more information, visit the ISL homepage at http://www.is.cs.cmu.edu/

Hence an MT system can be said to be doing *natural language processing (NLP)*. In fact as we shall see later, most machine translation applications require some degree of natural language understanding to do the translation.

Machine translation as a discipline dates back to the early nineteen-fifties. The complexity of the problem was originally underestimated, and some early successful demonstrations of experimental systems lead to unrealistic expectations which were hard to fulfil. This led to some skepticism, and funding on MT work almost ceased. In the early eighties, the Japanese Fifth Generation Computing Project revived interest in this work. The current approach to MT is more pragmatic and realistic. It is now widely accepted that *fully automatic, general-purpose, high quality* machine translation is a very difficult problem, but very useful and practical systems can nevertheless be developed by relaxing one or more of these criteria, and several useful systems have been built by doing so, and are in use today. Such systems are being used to translate public announcements, weather bulletins, technical documents, and web pages. Some machine



translation services are starting to become available on the World Wide Web. For example, the web page of the Altavista search engine (http://www.altavista.digital.com) also provides a translation service that can translate simple sentences among a handful of languages.

Components of an MT System

We can divide the machine translation task into two or three main phases – the system has to first *analyse* the source language input to create some internal representation. It then typically manipulates this internal representation to *transfer* it to a form suitable for the target language. Finally, it *generates* the output in the target language.

A typical MT system contains components for analysis, transfer and generation as shown in the diagram. These components incorporate a lot of knowledge about words (*lexical* knowledge), and about the language (*linguistic* knowledge). Such knowledge is stored in one or more *lexicons*, and possibly other sources of linguistic knowledge, such as *grammar*. The user interface is invariably a crucial part of most MT systems. The interface allows users to verify, disambiguate and if necessary correct the output of the system. Another common feature of NLP work is the use of large '*corpora*' (plural for 'corpus'). A *corpus* is a large collection of text which has been appropriately tagged, and is used for acquiring the required lexical and linguistic knowledge.

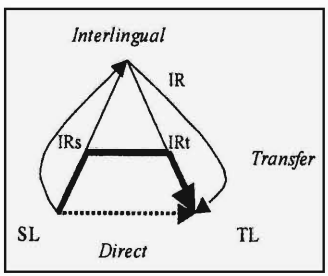
A *corpus* is a large collection of text which has been appropriately tagged, and is used for acquiring the required lexical and linguistic knowledge.

The *lexicon* is an important component of any MT system. A lexicon contains all the relevant information about words and phrases that is required for the various levels of analysis and generation. A typical lexicon entry for a word would contain the following information about the word: the part of speech, the

An MT lexicon typically needs to be much more formal, precise and elaborate than a typical human dictionary.

morphological variants, the expectations of the word (or the typical words, phrases or constructs that this word typically goes with), some kind of semantic or sense information about the word, and information about the equivalent of the word in the target language. Some systems prefer to split the lexicon into a *source* lexicon, a *target* lexicon, and a *transfer* lexicon that maps between the two. The exact format of the lexicons is a matter of engineering design, and would take into account the system designer's policy about issues like how to handle morphological variations, multiple word senses, synonyms and so on. An MT lexicon typically needs to be much more formal, precise and elaborate than a typical human dictionary, since it is meant for mechanical processing, and not for reading by humans. The lexicon plays a central role in modern MT systems.

Approaches to Machine Translation



Based on how closely the internal representation depends on the source and target languages, approaches to MT can be divided into three major classes – *direct*, *transfer-based* and *inter-lingual*, as illustrated in the diagram.

A *direct* MT system tries to directly map the source language to the target language, and is therefore highly dependent on both the source and target languages. A *transfer-based* approach first converts the source language into an internal representation (IRs) which is dependent on the source, but not the target language. The system then transforms IRs into a form (IRt) which is independent of the source language and depends only on the target language, and finally generates the target language output from IRt. The *inter-lingual* approach converts the input into a single internal representation (IR) that is independent of both source and target languages, and then converts from this into the output.

The difference between these approaches becomes clear if we consider translation between more than one pair of languages. Consider a case where we want to translate among N different

languages. The direct approach allows one to exploit knowledge about the particular language pair while developing the systems, but has the disadvantage that we need to create a new system for every new language pair, or $N(N-1)$ such systems. The inter-lingual system is theoretically the best, since it requires only N analysers and N generators. However it requires the creation of a very general inter-lingua, which may be very difficult. The transfer approach lies between these extremes, and is the most widely used approach in practice.

The lexical level deals with looking at the input string of characters and separating them into *tokens*, which may be words, space or punctuation.

Levels of Natural Language Processing

Dealing with natural language typically requires processing at various levels. In increasing order of difficulty, they are:

- ***The lexical level (or the word level)***: This level deals with looking at the input string of characters and separating them into *tokens*, which may be words, space or punctuation. This level also deals with issues like hyphenated words, and misspelt words. It is the lexical level which tells us that the input *"He joined the parti"* consists of four words, of which the last is incorrect. This level is sometimes called *'tokenisation'* or *'lexical analysis'*.

- ***The syntactic level (or the sentence level):*** This level deals with identifying the *structure* of a sentence, and verifying whether a sentence is grammatically correct. This level typically consists of a *'parser'*, which looks at a *grammar* of the language, and the input sentence, and tries to form a *'parse tree'*. If it can form a parse tree, the sentence is syntactically correct, and the parse tree gives us the structure and function of the various components. For example, a typical English sentence would consist of a subject and a predicate. The subject is normally a noun phrase, and the predicate is a verb phrase, and so on. The syntactic level tells us that the sentence *"He the party joined"* is (syntactically) incorrect, even though each word in it is (lexically) correct.

The syntactic level deals with identifying the *structure* of a sentence, and verifying whether a sentence is grammatically correct.

- ***The semantic level (or the meaning level):*** This level deals with the meaning of the input and its components. It is the semantic level which tells us that the sentence *"He ate the*

The semantic level deals with the meaning of the input and its components.

party" is semantically incorrect, though it is lexically and syntactically well-formed. In general, semantic analysis involves knowledge about the world, or at least the relevant aspects of the world.

- *The discourse and pragmatic level (or the conversation context level):* This level deals with information carried across multiple sentences, and with information that is not explicit in the input, but is implicit in the socio-cultural context of the input passage or conversation. For example, the expected answer to the question *"Do you know what the time is?"* is something like *"4 p.m"*, and not just *"Yes"*, though the latter is lexically, syntactically and semantically accurate.

A machine translation system needs to transfer across all these levels.

Issues in Machine Translation

Machine translation (and natural language processing in general), is a difficult problem. There are two main reasons, which are related. The first reason is that natural language is *highly ambiguous*. The ambiguity occurs at all levels – lexical, syntactic, semantic and pragmatic. A given word or sentence can have more than one meaning. For example, the word 'party' could mean a political entity, or a social event, and deciding the suitable one in a particular case is crucial to getting the right analysis, and therefore the right translation. The second reason is that when humans use natural language, they use an enormous amount of common sense, and knowledge about the world, which helps to resolve the ambiguity. For example, in *"He went to the bank, but it was closed for lunch"*, we can infer that *'bank'* refers to a financial institution, and not a river bank, because we know from our knowledge of the world that only the former type of bank can be closed for lunch. To get MT systems to exhibit the same kind of world knowledge in an unrestricted context requires a lot of effort.

When humans use natural language, they use an enormous amount of common sense, and knowledge about the world, which helps to resolve the ambiguity.

Contemporary Machine Translation Systems

It is impractical to design completely general-purpose, high-quality, fully-automatic machine translation systems.

The factors mentioned above make it impractical to design completely general-purpose, high-quality, fully-automatic machine translation systems, as already mentioned in the introduction. Contemporary practical machine translation systems therefore adopt one or both of the following strategies:

1. ***Restrict the domain of application, or the complexity of the language***: This is known as the *sublanguage* approach, which relaxes the criterion of general-purpose translation. One example of a restricted-domain approach is a Canadian system known as TAUM/METEO, which has been translating weather bulletins from English to French, completely automatically, for a number of years. Another example of a sub-language approach is the system in use at the Caterpillar Tractor Company in the US. This company sells tractors in all parts of the world, and needs to maintain manuals in more than 15 languages. They have a system that ensures that the original manuals are written in a subset of English that is sufficiently well-defined to allow automatic translation into all the other languages.

2. ***Involve the human in the loop***: Such systems relax the criterion of full automation, and rely on human editors for pre-editing the input, or post-editing the output, or disambiguating during translation. Depending on who does more work, the systems can be called 'Human Assisted Machine Translation (HAMT)', or 'Machine Assisted Human Translation (MAHT)', also known as *translation tools*. One example of a translation tool is a technical dictionary, which is particularly useful for translation in technical domains, to help ensure correct and consistent translations of technical terms. For example, the word for 'Dialog-box' in Italian is 'finestra', which means 'window', and in French is 'boite', which means 'box'. A translation tool would allow a user who does not know too much French or Italian to translate such terms.

Typical commercial systems (See *Box 2*) use one or more of these methods.

Box 2. Commercial MT System/Tool Examples: Systran and METAL

Systran is an automatic translation system based on the transfer approach. It works on a sentence by sentence basis. The three main module classes are the Dictionary, Systems Software and Linguistic Software. The Dictionary module contains over 2.5 million terms across all covered language pairs, and includes specialised phrase dictionaries and an ability to have user-specific dictionaries. The Linguistic modules contain the analysis, transfer and generation components for the various languages. The Systems module deals with the user interface, file manipulation and integration of the other components.

Systran began as a research project at the California Institute of Technology in 1957, and was later funded by the US government to do Russian to English translation. Over the years, the technology matured and came to be used by corporates like Xerox and Ford. The Altavista Search Engine site uses Systran to offer a translation service. Today, Systran is a set of translation systems ranging from stand-alone PC versions to client-server models, covering 14 language pairs.

For more information, visit the Systran homepage at http://www.systransoft.com/

METAL was a translation system initiated in the late seventies by Siemens-Nixdorf with the University of Texas. It uses the concept of a controlled language to achieve high quality translation in various technical domains. It can also produce indicative translation for general texts, which needs to be post-edited for style. METAL is now called LANT-MARK, and marketed by LANT, a Belgian company. LANT has a suite of related Language Technology products: LANT-Master, a language checker, integrates into existing word processors like MS-Word and allows the vocabulary and style of texts to be in a controlled language which can then be automatically translated; Pangaea is an electronic dictionary that allows the creation of customised phrase lexicons; Eurolang Optimiser, is a translation tool that uses the concept of translation memory.

For more information, visit the LANT homepage at http://www.lant.be/

Current Work

The current focus in MT research is on using machine learning techniques to automatically acquire the lexicon and grammar. This involves using large corpora and applying statistical techniques such as symbolic induction or neural networks to capture correlations in the corpus. The corpora used can be either for a single language, or can be 'aligned corpora' which means a bilingual corpus of translated text in the source and target languages, containing information about which part of the source language text corresponds to which part of the target language text. Another related approach, called as translation memory, is to automatically 'remember' the entire translations

Box 3. Indian Machine Translation Projects

India is a relatively new player in the machine translation field. With its large number of languages, it is a major potential beneficiary of the technology, so research in this area needs to be intensified and coordinated.

Currently, there are at least four major machine translation efforts in India. They are currently being funded by the Technology Development in Indian Languages (TDIL) project of the Department of Electronics (DoE), Government of India.

- The Anusaaraka group with researchers from the Indian Institute of Technology, Kanpur and the Centre for Applied Linguistics and Translation Studies (CALTS) and University of Hyderabad has been working on using a Paninian Grammar approach to deal with Indian languages. The Paninian Grammar approach uses an internal representation that is based on the kaaraka theory developed by Panini to describe Sanskrit grammar, and exploits the similarity among Indian languages.
- The Anglabharati project, also originating from IIT Kanpur, deals with translation from English to Indian languages. It uses a rule-based transfer approach, and has been tested on technical domains, such as product manuals, and medical descriptions.
- The Knowledge Based Computing Systems (KBCS) group at the National Centre for Software Technology (NCST) in Mumbai is currently working on MaTra, a prototype system for human-assisted translation of news sentences from English to Hindi. This group is exploiting the use of human-computer interaction to alleviate some of the traditional problems of machine translation.
- The Centre for Development of Advanced Computing (CDAC), Pune is developing Mantra, a translation system for translating office documents from English to Hindi. It uses a formalism known as XTAG, developed at University of Pennsylvania, and views the translation process as a mapping from one syntactic tree to another.

of frequently occurring phrases or sentences in order to avoid processing them repeatedly.

There is significant activity on machine translation in Japan and Europe, and to a lesser extent, in the US. India is also active in this field, with at least four or five active groups (See *Box 3*). Given the multiplicity of languages in India, such efforts are very relevant, and need to be intensified further.

Conclusion

Two phenomena have given a new impetus to machine translation work – the globalisation of the world economy, and the explosion of the Internet and the World Wide Web. Both these developments mean that there is a need for making an immense collection of natural language documents available to a multi-

lingual global audience, and translation tools and systems can go a long way in meeting that need. The global translation market is estimated to be at least 12 billion dollars. Systems that automatically translate Kalidasa and Shakespeare may still be a distant dream, but systems that translate stock market reports, weather bulletins and technical manuals are a reality today, and will continue to play an increasingly important role in the society of the next millennium.

Suggested Reading

Address for Correspondence
Durgesh D Rao
National Centre for Software Technology
Gulmohar Road 9, Juhu,
Mumbai 400049
email: durgesh@ncst.ernet.in

[1] **The ACL NLP/CL Universe, *http://www.cs.columbia.edu/~radev/u/ db/ acl/html/* Contains a large amount of pointers to on-line NLP and MT information and resources.**

[2] **John W Hutchins. *Introduction to Machine Translation*. Academic Press, 1992.A good introductory book on machine translation.**

[3] **The Altavista Service, *http:altavista.digital.com/* This is a good site to search for more information on anything, including Machine Translation. It also runs a simple translation service based on the Systran system.**