Machine translation: benefits and advantages of statistical machine translation and NRC's Portage

National Research Council of Canada. Information and Communication Technologies

https://doi.org/10.4224/21274926

NRC·CNRC Information and Communications Technologies

Machine Translation: Benefits and Advantages of Statistical Machine Translation and NRC's *Portage*

April 2015

National Research Council Canada
Conseil national de recherches Canada

Canada

Introduction

Machine translation (MT) – or automatic translation, as it is sometimes called – automatically transforms a text from one natural language into what is ideally a semantically equivalent and grammatically correct text in another natural language using a computer. From the point of view of an average user, an MT system is just a black box: you type in some text in a source language, or direct the system to a file, and seconds later you obtain a translated version of the text in the target language.

But, to dig a little deeper, different types of machine translation – statistical and rule-based – can have an impact on the results. Since 2003, the world-class research team at the National Research Council of Canada has been developing a proprietary statistical MT (SMT) system, called *Portage*.[1] This paper provides some background about SMT, and contrasts it with the other major current in machine translation: rule-based MT (RBMT). For most of this young field's history, RBMT was the dominant paradigm in MT. It can generally be traced back to 1948 when a memo sent out by Warren Weaver to a small group of scientific colleagues suggested that computers might be used for translation in the same way they had been used for code breaking during the Second World War. Later, as Chomskyan linguistics grew more dominant in the mid-1960's, MT came to adopt similar types of declarative rule formalisms.

System developers in that period tended to be linguists and computer scientists whose objective was to program the machine to emulate what they understood a human translator does: analyse the source text into some sort of meaningful representation and map that abstract representation into an equivalent target-language text. This required handcrafting large-scale computational grammars and dictionaries that contained many hundreds, if not thousands of linguistic rules – a challenging, time-consuming and costly endeavour. The resulting systems were not particularly robust and often failed to produce an output, largely because it is

[1] Actually, the Canadian government has had an active interest in MT since 1963, financing research in various universities and public labs over the years. *Portage* is the latest product of that long-standing commitment.

extremely difficult to write an exhaustive grammar and a complete dictionary for an object as complex and evolving as a natural language, not to mention the contrastive grammars and dictionaries required to map one language into another.

In 1988, a group at the IBM T.J. Watson Research Centre proposed statistical machine translation. Some of this group had previously been involved in the successful development of statistical speech recognition systems – another difficult problem that had long stymied the rule-based approach.

In contrast to RBMT, SMT systems have no declarative grammar rules or dictionary entries. Rather than relying on linguists' intuitions about language, these systems instead draw their linguistic knowledge directly from large bodies of previously translated text. Machine learning algorithms are applied to this text in order to automatically estimate the probability of some word or phrase *X* in the source language should be translated as *Y* (or W, or Z…) in the target language. An SMT system is composed of *two* major linguistic components which are both probabilistic: a translation model, which roughly corresponds to the bilingual dictionary in an RBMT system; and a language model, which ensures the fluency of output text by maximizing the probability of each word W of the translation given the choices already made for the *n* preceding words.

There are numerous other statistical modules, as well as a core algorithm called the decoder, which performs the actual search for an optimal target text in applying these probabilistic repositories of linguistic knowledge to the source text. Referring back to our preliminary definition above, one could say the translation model is responsible for the semantic fidelity of the translation, while the language model is responsible for its grammaticality.

SMT advantages

SMT offers a number of important advantages over RBMT, the first being greater **versatility**, or polyvalence. Under the rule-based approach, whenever a new language *T* has to be added, computational linguists specializing in that language have to be hired, along with bilingual lexicologists qualified in the pair *S* > *T* – if any such can be found.

In SMT, on the other hand, the same machine learning algorithms can be applied to virtually any language pair. Someone with an English > French SMT system who decides to tackle English > Spanish (or Portuguese > Bulgarian, and so on) can do so without delay – provided, of course, that the requisite training materials are available. And these, as mentioned above, are simply large bilingual volumes of well-translated texts.[2] Consequently, SMT reduces the time, effort and cost of developing new MT systems by orders of magnitude. A system that would take months and many hundreds of man-hours to develop under the rule-based approach can now (literally) be generated overnight.

Another important advantage that SMT systems have over their RBMT counterparts is their **robustness**. In SMT, every source language sentence has a multitude of possible target language renderings, although some are usually much more probable than others. As a result, SMT systems never fail to produce an output and, in this sense, they are naturally more robust than the older RBMT systems, which often depended on a complete grammatical analysis of the source sentence. Of course, this robustness would be of little interest if the quality of all the generated translations were poor.

[2] How large a bilingual corpus is required to train an SMT system? People who work in the field tend to say the larger the corpus, the better. However, there is a minimal size for a training corpus, below which the acquired probabilities will not be reliable. A figure that is often cited is a hundred thousand sentence pairs, or roughly a million words.

However, it has been conclusively demonstrated in open international competitions like those organized by the American military and the National Institute of Standards and Technology (NIST) that SMT systems tend to produce **better-quality** translations than their RBMT counterparts. The best MT systems available today are all statistical; SMT is state of the art.

NRC's *Portage* system

Statistical systems completely dominate the MT landscape today. Some of the better known of these systems, like *Google Translate* and *Microsoft Translator*, can be accessed for free over the Internet, or are available for very modest user fees. Other open-source SMT systems (most of which are based on *Moses*[3]) are being marketed by vendors who promise to tailor them to the particular needs of their clients. With all of these options, someone who is considering MT may initially wonder why they should bother looking at *Portage*, which is not offered as freeware, but there are a number of good reasons to do so.

Among statistical MT systems, *Portage* is one of the very best available in the world today. This was clearly demonstrated by the system's showing at recent NIST competitions, where *Portage* finished first among all systems translating from Chinese to English, and second among all systems translating from Arabic to English – no easy feat! NRC's *Portage* team has twice been invited to participate in large and prestigious projects funded by DARPA, the US Defense Department's Advanced Research Projects Agency, alongside such highly regarded labs as Raytheon BBN and RWTH (Germany). As well, the group has actively collaborated with major commercial partners like *Systran*, a company that has dominated the MT business for decades.

[3] *Moses* is a full-blown, open source SMT system originally developed under the leadership of Philipp Koehn as an academic alternative to the proprietary systems developed by large corporations like Google and Microsoft. Supported by funding from the European Commission, *Moses* has recently been extended to numerous European language pairs.

Many potential MT users have legitimate concerns over the security of their texts and are reluctant to send them out of the country for translation, or over the Internet into the cloud. Fewer still are prepared to see their texts used to improve another company's MT engine, over which they have no rights.[4] *Portage* provides a straight-forward solution to all of these concerns. The system is normally installed on a local server, within the user's secure premises. No one else has access to it, or to the texts it processes. And as we will discuss, *Portage* can be easily integrated within a user's existing document processing environment.

Opportunities to benefit from *Portage*

From the early origins of machine translation right through to the end of the Cold War, military and intelligence communities were the principal users of MT, and also provided the primary source of funding for MT research and development. The vast volumes of material that these services needed to process far exceeded the productive capacities of all available human translators -- MT offered the only possible hope.

Furthermore, geopolitical considerations determined the major foreign languages that MT systems were developed for. It was no accident, in other words, that for the first public demonstration of machine translation that was given in New York in 1954, the language direction was from Russian into English. Nor was it fortuitous that the American government heavily subsidized the development of a Vietnamese-to-English MT system in the 1960's, during the Vietnam War, and a Farsi-to-English system in the years before and after the fall of the Shah of Iran.[5] In all such cases, MT was seen as responding to the urgent needs of national security.

[4] This is the case with *Microsoft Translator*, although the company insists that no other client can recover and reconstitute the texts a user has uploaded onto its translation servers.
[5] In the case of both Vietnamese and Farsi, the principal beneficiary of these US government investments was the firm of Logos Corporation.

In the United States, military and intelligence communities continue to make extensive use of MT. If anything, their reliance on machine translation has dramatically increased in recent years, since they now monitor not just written texts but spoken communications, videos and social media as well. If nothing else, the attacks of September 11 served to sharpen the interest of the American Defence Department in machine translation, prompting it to reinvest heavily in MT R&D and organize the international competitions that were alluded to above.
And of course, Arabic suddenly became a principal focus of attention, with Chinese not far behind. Funding agencies have emphasized the importance of rapid system development, on the basis of smaller and more varied training materials. It is safe to say that much of the remarkable progress that has recently been achieved in machine translation is attributable to the large investments in the technology made by the military and intelligence communities.

In this context, it is important to note that MT is primarily being used for **information-gathering** purposes. Even if the quality of the MT output was absolutely perfect (which it is not), no intelligence service could possibly afford the time to read through such volumes of translated materials. Instead, MT serves as a crucial component of a larger chain designed for **triage**, helping to sift through massive quantities of foreign-language texts with a view to locating those of interest, which may then be translated by humans. And in this role, today's MT systems function very well.

Interestingly, the military and intelligence agencies are no longer the only ones interested in gathering information that appears in a foreign language. As a result of the increasing globalization of trade and commerce, businesses today are also eager to know what people are saying about their products and services in foreign-language markets. In recent years, a whole new industry has emerged that focuses on **business intelligence**, i.e. on the gathering and analysis of large amounts of data, particularly on the increasingly important social media, which is intended to help businesses adjust and hone their commercial strategies.

A few years ago, it was primarily large multinationals that were actively involved in so-called 'foreign' markets. Today, more and more SMEs are venturing abroad, and they too have a pressing need to comprehend what is being said about them in a range of foreign languages. Firms specializing in business intelligence are therefore looking to machine translation to help them respond to this demand for large-scale multilingual text processing and analysis. In fact, NRC's *Portage* team has already participated in some exploratory work in cross-linguistic sentiment analysis with the Canadian firm MediaMiser.[6]

Machine translation for dissemination purposes

As we mentioned above, in their information gathering activities, the intelligence and business communities are using MT essentially for **assimilation** purposes, to help people read and comprehend part of the content of a foreign language text. But what about MT for **dissemination** purposes, where the system is used to generate and eventually publish a text in a foreign language? This, after all, was the original aim and ultimate goal of machine translation.

Since MT's inception, there have been many attempts to employ the technology in this way, both in the public and private sectors. In the United States, companies like Systran, Weidner and Logos (to name just a few) were founded in the second half of the last century with just this goal. In Europe, with all the official languages of the EU creating an enormous demand for translation, the European Commission invested substantial effort and large sums of money in adapting a commercial MT system to help it meet its translation needs, before launching its own ambitious MT development program.[7]

[6] Sentiment analysis refers to the process of computationally categorizing opinions expressed in a piece of text.
[7] The program in question was called Eurotra, and it ran from 1978 to 1992. To this day, the EC continues to finance numerous R&D projects in MT and MT-related technology.

In Canada, the government's Translation Bureau mounted extensive trials of several commercial MT systems in order to determine if the technology could help it better meet a spiralling demand for its services and hopefully reduce its costs. Viewed retrospectively, however, the great majority of these efforts came to naught. Until recently, the successful implementations of MT for dissemination purposes have been relatively few and far between.

Although reliable figures are difficult to obtain, there can be little doubt that MT's share of the world translation market has until now been very modest, not to say minimal. And the reason for this is quite simple: the quality of the raw MT output has not generally been good enough to allow for its direct publication or cost-effective exploitation. Even when the MT system is coupled with a human post-editor who systematically revises, corrects and improves the computer's output, this arrangement will not prove cost-effective unless the machine's first draft achieves a certain level of quality. Otherwise, the post-editor finds herself spending more time correcting the machine's proposals than it would take to translate the text from scratch.

Canada's own *Météo* system, a system specifically designed for the translation of the weather forecasts issued daily by Environment Canada, is among the few exceptional cases where the raw machine output does attain close-to-publishable quality. The reason this system succeeds, however, is that it is designed specifically for weather bulletins, a very particular and narrow sublanguage. Language service providers (LSP's), for their part, are normally called upon to handle a much wider variety of texts, most of which are far more challenging than weather bulletins, and until recently there has been no compelling business case for them to adopt machine translation in order to help them produce the high-quality texts their clients require. And the same is true for most in-house translation services that are part of large corporations.

The advent of the new statistical paradigm in MT is in the process of changing all that. As mentioned above, SMT systems have been shown, objectively and

convincingly, to produce better-quality translations than their rule-based counterparts. Furthermore, the ease with which **specialized SMT systems** can be trained now makes it possible to develop systems that are tailored to the particular text type, style and terminology of any number of clients or client departments. And here too, it has been shown that these specialized systems can produce better-quality translations than such general-purpose systems as *Google Translate*.

Just as important as the advances in the technology, people's attitude and expectations with regard to MT have also undergone a significant change in recent years. Not very long ago, many who turned to machine translation expected that these computerized systems would allow them to dispense with costly, slow-moving human translators and obtain a high-quality translation, literally at the push of a button. After all, if computers helped us put a man on the moon and are now capable of defeating the world's best chess masters, why can't they automate the age-old process of language translation?[8]

Today, due in large part to the public's greater exposure to this technology, through easily accessible websites like *Google Translate* and *Bing Translator*, we no longer entertain such naïve views. Human translators are not about to go the way of hand weavers. That said, the improved quality of the translations produced by the new breed of statistical MT systems does seem to have reached a point where a **productive partnership** with human post-editors may finally be cost-effective. And this, in conjunction with the ever-increasing demand for translation and intense market pressures to reduce costs and turnaround time, is leading many LSP's and in-house translation services to take a fresh look at MT.

In the previous paragraph, we mentioned the public's increased exposure to machine translation and a consequent change in its attitude to the technology. Among translators too, there has been an important shift in attitudes. It used to be the case that many professional translators feared that MT would actually put them

[8] Why indeed? The question warrants a more fully developed answer than we have space for here.

out of a job. Far fewer translators believe that today, for they have seen firsthand what the technology can and cannot do, and they know that the demand for translation continues to spiral upward, far exceeding the collective capacity of the profession.

Instead, they are coming to view MT as a welcome assistant, which may well help them increase their productivity and, in certain cases, actually relieve them of the need to tackle the most repetitive and least interesting texts.[9] Translation service managers are also changing their attitude and expectations with regard to machine translation. In the past, many of those who were eager to test the technology did so in the hope that MT would save them money by replacing all the other tools and resources that their translators had previously been using. Not so today. Now, more and more translation managers understand that MT is best implemented as an additional component in the translator's varied arsenal, supplementing rather than replacing the other tools the translator has come to rely on and use productively.

Most popular among these are undoubtedly the translation memory (TM) systems that automatically detect and recover the translation of previously encountered sentences. At first, many translators resisted the introduction of these systems; now, the overwhelming majority use them on a day-to-day basis -- and would be very unhappy to relinquish them. But what about the many sentences that haven't been previously encountered, for which a TM system will have nothing to propose? An effective way of combining the two technologies in this situation is to have the MT system insert a machine translation in the TM editor, in what would otherwise be an empty target cell.[10]

Needless to say, the translator is free to accept, modify, or reject the machine's proposal, just as with the proposals that come from the translation memory. But if the MT system has been trained on an adequate corpus that corresponds to the

[9] Indeed, this was the original impetus that led to the development of the Météo system: translator ennui with boringly repetitive weather bulletins.
[10] Or a copy of the source segment in the target cell.

domain of the source text, chances are that the machine's proposal will contain at least some elements that the translator will be able to profitably recycle.

This way of combining MT and TM is all the more natural given that the core of the training material used to generate an SMT system is often drawn from the very same database of past translations as the translation memory. Another important point that TM's and SMT have in common is that both improve with use. As new, human-approved translations are added to the TM database, that system should be able to find matches for more and more sentences; and the SMT system will also have more data on which to calculate and improve its translation probabilities and coverage.

The impressive recent advances in SMT, together with this kind of productive partnership between different translation technologies are leading an increasing number of translation services and language service providers not just to consider machine translation, but to actually implement it. According to the January 2015 newsletter published by TAUS (the Translation Automation Users' Society), 40% of all translators are already making use of machine translation in one way or another.

Jost Zetzsche, in the 243rd issue of his highly respected Tool Box Journal, cites data from Memsource, a cloud-based TM system with more than thirty thousand registered users, suggesting that over 50% of their clients regularly consult the free online MT systems offered by *Google Translate* and *Microsoft Translator*. And in Canada, two of this country's largest LSP's are now using *Portage* on a day-to-day basis, because they have found that the system allows them to increase productivity, thereby saving time and money, without compromising quality or the private nature of their clients' texts.

A host of additional applications

So far, we have focused on two ways of using machine translation which could be considered to be at opposite poles of the application spectrum: on one hand, raw MT, used principally by the intelligence and business communities for information-gathering purposes; and, on the other, revised MT, used principally by translation services and LSP's for publication purposes. But between these two poles, there is a wide variety of other possible applications for MT, some of which we will briefly mention here.

The phenomenal growth of the social media, along with the steadily increasing popularity of e-commerce, has indirectly created an enormous new demand for translation. Millions of people every day post comments and user reviews on social media and business websites, which others, who don't necessarily understand the language of the post, nevertheless want to read. They can, of course, cut-and-paste that user-generated content into a free online MT system like *Google Translate.* But some companies have found that they can provide their users with better-quality translations by actively partnering with an MT developer in order to create a semi-customized system for their particular type of texts; this is what TripAdvisor has done with SDL. Other popular online sites, such as eBay, have invested in the development of their own proprietary MT technology.

Now in most such cases, the lifespan of the posted texts is very short, and so there is little point in having the machine translations carefully revised. Hence, what the users generally obtain is raw MT output; and most often, it is enough to allow them to make a decision as to whether they want to patronize a given hotel or restaurant, or purchase a given product. In the eBay scenario, moreover, the users may actually interact with the vendor in a conversation that is partially mediated by machine translation.

A slightly different application of unedited machine translation has been implemented at Microsoft, where the company is constantly updating enormous

online knowledge bases that support its wide array of products. Human translation in this context is clearly out of the question, and so Microsoft has applied its own MT system to the task. The content of the knowledge bases is drafted in English, but Microsoft clients can now query them in seven different languages (at last count), and hopefully find answers to their questions in their mother tongue. The company has also surveyed users on their satisfaction with the answers they obtained and, interestingly, there appears to be little difference in the satisfaction rate of those who query the knowledge bases in English and those who obtain their answers via machine translation.

Finally, between raw, unedited machine translation and fully edited MT output that aims for the same level of quality as human translation, various intermediate levels of post-editing are of course possible. One well-known compromise is commonly referred to as 'light post-editing'. What this means is that the reviser limits themselves only to those corrections that will interfere with understanding on the part of the reader. Hence, lightly post-edited MT output may contain minor grammatical errors, or stylistic infelicities – as long as they won't prevent the reader from grasping the meaning and all the essential content of the translated text. Lightly post-edited MT is often used for in-house communications or, in one well-known case, for internal memos and preliminary drafts of documents at the European Commission.

Needless to say, NRC's *Portage* could easily be adapted to any and all of the above scenarios.

Summary

Machine translation has progressed remarkably in recent years, particularly since the advent of the new statistical paradigm. Among SMT systems, NRC's *Portage* is among the very best in the world, as repeatedly demonstrated in various international competitions. What is more, the system is currently being used to cost-effective advantage by major language service providers in Canada, while being tested by businesses in other domains, as well as by the military and intelligence services in the United States.

The NRC is eager to extend the use of *Portage* to a broader range of partners as there are significant opportunities for the system to provide benefits in multiple areas. *Portage* is not only state of the art as far as the MT industry is concerned, but also less costly than most of its major competitors.

Organizations interested in finding out more about *Portage* –for practical demonstration or to inquire about licensing fees and collaboration opportunities – are invited to contact NRC to learn more:

Pierre Charron, Client Relationship Leader
Email: pierre.charron@nrc-cnrc.gc.ca
Tel.: +1 613 990-0336