Sharing our stories: With translation software?

Frank Berberich
Tokiwa University

Reference data:

Berberich, F. (2006). Sharing our stories: With translation software?.
In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.) *JALT2005 Conference Proceedings*. Tokyo: JALT.

Machine translation is now readily available in software packages and as a free option on many Web service sites and portals. Used with some skill, it can be helpful for reading texts in another language and, with considerably more effort, for producing text in a second language. This paper discusses what machine translation is, some theoretical issues in machine translation, how it is used, and how it can be used for language learning..

機械翻訳のためのソフトウェアは現在市販されている。それは多くの　Web　サイトとポータルと同じく無料な選択です。もしそれが少し技能で使われるなら、それは他の言語で本文を読むことに対して、助けになります。多くの努力をするなら、他の言語で本文を書くために使われることができます。この論文は機械翻訳が何であるか論じます。それは機械翻訳についていろいろな論理の問題を論じます。機械翻訳が使われる方法そしてどのように機械翻訳が言語学習者によって使われるかもしれないか論じます。

Automatic language translation with machines, known as 'machine translation' (MT) offers the possibility to communicate in another language without the need to learn that language. From the early history of computers, MT has been a goal in computer science, and the technology for written language is now widely available. There is still a long way to go, however, and the limits of current systems are obvious even to the casual user.

A story in computer science concerning MT, most likely apocryphal, tells about the unveiling of a (usually Chinese or Russian) system for translating from English. As a test, the well-known sentence, "*The spirit is willing but the flesh is weak.*" is entered, the system goes to work and finally produces—for the Russians, let's say—the Russian equivalent of, **The vodka is strong but the meat is putrid.** This story exemplifies some of the problems of MT, and suggests why it has taken so long to develop practical systems.

Only in the last 10 years or so has MT has become available in reasonably priced desktop software packages, and, more recently, now as a free option on various Web services and portals. These MT resources require a fair amount of skill in the second language to use effectively, both for reading and especially for production.

This paper discusses what MT is and how to use it. Section 2 introduces MT in more detail, and some theory of MT is presented in Section 3. Section 4 discusses how to use MT software, and Section 5 offers suggestions on MT's place in language learning. The basic conclusions are that MT is useful but needs a fair amount of skill in both languages, and usually should not be used by our students for class work.

What is MT?

MT is the use of a computer to convert natural language productions—speech or text—in one language to equivalent productions in another language. It is also the field of study involving MT systems.

MT systems work almost entirely with pairs of languages, so there are systems for translating between, say, English and German, French and Japanese, Arabic and Thai, and so forth. While certain computing processes within MT systems—"translation engines", for example—might be common to translation systems for any language pair, the majority of the data and analysis involved is specific to each language pair.

Some pairs of languages closely resemble each other, while other pairs are quite distant. The subtlety of a translation is usually inversely related to the distance of the pair. Portuguese and Spanish, for example, are mutually intelligible, and MT issues with these languages can address relatively fine details of expression and nuance. In contrast, Japanese and English—the language pair in focus in this paper—have little in common, and so the MT challenge is simply to get a rough equivalent to the original.

Some terms

MT is one branch of the large field of natural language processing (nlp). It is different from Neural Linguistic Programming. Nlp includes a range of functions from simple spelling check in a word processor to interpretation of speech spoken at normal speed. Nlp can be divided into the two major areas of speech processing and text processing.

In more detail, speech-processing involves extracting acoustical information or meaning from spoken language, using voice recognition (VR) or speech recognition (SR), while text processing is concerned with written language. It is worth noting, by the way, a common confusion of the terms VR and SR. VR is concerned with examining certain acoustical features of a voice to, for example, confirm a person's identity or estimate an emotional state such as the person's level of stress. SR is concerned with separating the speech sound stream into its constituent words. SR is often used to produce a text for subsequent text processing. Commercial applications of SR include the dictation packages that are intended to largely replace input with the keyboard.

Text processing includes a wide range of functions, such as character-recognition, corpus analysis, spelling and grammar checking, and extracting semantic content from text to interpret its meaning, or to translate it into another language.

A full-scale real-time language-interpreting system—one that would "listen" to normal speech in language A and simultaneously produce the equivalent speech in language B—would require SR, MT, and artificial speech production technology. Such a system is not currently readily available. By far the most practical MT systems available operate on text files, such as produced by a word processor. This is the kind of MT system considered in this paper. Thus, we are discussing MT that converts a text in language A—the 'source' language—to its equivalent text in language B—the 'target' language.

What does "equivalent" mean? Intuitively, of course, it means that the two texts have the same meaning in their respective languages, but this definition veers into philosophy and is difficult to quantify. A more practical operational definition of "equivalent" is simply that if text B, in the target language, is accurately translated back ('back-translated') into A, the source language, the resulting text will be identical to the original. For example,

Example 1.
Source: *This is a book.*
Target: これは本です。
Back-translation: **This is a book.**

The notion of back-translation tends to assume that the translation and back-translation will perform with the same accuracy. Most systems are built to work both ways, using the same engine, but it is certainly possible that a system could work well one way but not the other.

It turns out that back-translation is the main practical technique to improve the accuracy of an MT production. Essentially, one uses the system to confirm its own translation, revises unsuitable words or phrases, and tries again. Along these lines, a revealing experiment is to re-translate a text several times, without revising the intermediate results. One finds that several iterations of translation and back-translation without intervention by the user results in a text that gradually gets farther from the original meaning until it settles onto a text that is quite different from the original. So, the user and the MT system must work together for best results.

Some theory

MT is usually a chapter in artificial intelligence textbooks. A clear and readily accessible introduction to MT can be found in Arnold (2002).Techniques in MT can be as crude as simply substituting, word by word, the target for the source. For example,

Example 2.
*Source:*これは　本　です。
Target: **this book is.**

Even in this simple example, the shortcomings of the approach are evident. We can add refinements by using transformations based on the syntax and grammar of the target language, giving,

Example 3.
Target a: **this is book**
Target b: **this is a book**
Target c: **This is a book.**

For this simple example, such mechanical techniques work reasonably well. Unfortunately, the richness of natural language means that things very quickly get far more

complicated. Some of the problems include:

Ambiguity—a term has multiple meanings

Idioms—the literal meaning and the understood meaning are quite different

Nuance—irony, sarcasm, disappointment, etc.

Poor writing in the source

Semantic discontinuity—the semantic categories of source and target are different (traffic lights are red and green in English but red and blue in Japanese)

Among these, ambiguity is perhaps most frequent and challenging. For an example at the simplest level, consider the single word 'fly' and some of its meanings.

Example 4.
The word *'fly'* can mean:
a kind of insect
~like a bird
~an airplane
an opening in men's pants
a fishing lure
a type of fishing
A type of curtain on a stage
flee
(etc.)

Ambiguity, of course, extends to the sentence, paragraph, and even the full text level, as anyone who has read, say, Finnegan's "Wake" will know. A well-known ambiguous text, widely attributed to Groucho Marx, is:

Example 5.
"Time flies like an arrow; fruit flies like a banana."

This sentence can mean:

Using a stopwatch, time the flight of flies as quickly as an arrow.
Time goes very quickly.
A kind of fly called a 'time fly' likes an arrow.
Certain small insects like a banana.
Fruit takes wing like a banana.

Another widely quoted example of ambiguity at a higher level, from Chomsky, is:

Example 6.
Colorless green ideas sleep furiously.

This sentence is structurally correct, but semantic nonsense.

Here are some examples of ambiguous sentences and their back-translations using a typical commercial MT package:

Example 7.
Time flies like an arrow.
時は矢の様に飛びます。
Fly like an arrow in the time.

Example 8.
The spirit is willing but the flesh is weak.
精神は喜んであるが、肉は弱い。
Mind is rejoicing, but the meat is weak.

These examples show the difficulty even a single word or sentence can present to an MT system. In all of them, the meaning of the source cannot be reliably extracted from the words themselves, so a simple dictionary lookup does not work. More information on the text is needed.

More sophisticated techniques involve adding semantic knowledge concerning related words and structures, and inferring meanings from the larger body of the surrounding text. For example, a system might have a comprehensive database of "semantic nets" of related words so that if some these words are found in the text, the probability of a correct interpretation is improved. Thus, if the source includes words like 'insect', 'wings', 'buzz', etc., the insect interpretation of 'fly' is likely to be correct.

Very sophisticated approaches begin to resemble human thought processes and indeed involve artificial intelligence. A system might have a great store of common knowledge about many life situations and use this to help interpret the text. For example, a restaurant situation involves language about food, servers, checks, tipping, etc, and this knowledge would help a system make inferences from the text.

Clearly, the more ambiguous, idiomatic, or nuanced a text is, the more difficult it is likely to be for MT. For this reason, MT currently works best with concrete material such as instruction manuals, newspaper reports, technical papers and specifications, and so forth. MT does not work well with poetry (though systems for writing poetry often come up with surprisingly original and entertaining works).

A very simple example of how subtle the process of translation can be is the well-known place name in Japan, 青山, usually translated as "Blue Mountain", which has little flavor in English. It would be very difficult to build an MT knowledge-base so extensive it could infer that the word 'Vermont' from French probably conveys the nuance of 青山 far better.

How to use MT

MT can be used for both reading and writing by exchanging the roles of source and target. For reading, the source is the L2 text and the target is in L1, while for writing, the L1 text is the source used to produce the L2 target text.

Reading

As would be expected, reading with MT is usually much less demanding on the user than writing. For reading, the translations of current systems are useful even for persons not familiar with the source language. The topic and gist of the source are usually evident, but more detailed relationships among the main points in the text might be obscured by strange grammar usages and plain mistranslation. It is often necessary to read the source to understand, for example, who or what are the subject, direct, and indirect object of some action. Thus, a basic reading knowledge of the source is very helpful to assure understanding. In this case, MT's greatest merit for reading Japanese can be simply the bulk translation of the kanji in a text.

It is sometimes helpful to experiment with the target. One can rewrite an unclear passage, back-translate it into the source, and then translate this revision. If the revised source is similar to the original and the translation is clearer, the rewritten target is likely the more accurate translation. With sufficient knowledge of the source language, one can try paraphrasing an unclear passage and seeing if the translation gives the expected meaning.

Writing

While writing with MT is a demanding task, this writer can attest that it is far easier than producing original text with only the help of dictionaries. The basic routine is a loop: translate, back-translate, rewrite, translate, and so on. First, the source text is produced and translated. The target is then back translated and the text that produces errors in the target is fixed in the source. Typical problems include selection of unintended meanings, obscure vocabulary, mistranslation of complex sentences, and so forth. This process is repeated until the back-translation and source are satisfactorily similar, so the target appears to be a good approximation of the source. Several iterations—up to 10 or so—may be needed. LogoVista's コリャ英語！——一発翻訳 (2005) facilitates this with an automatic back-translation feature.

Especially with a distant pair like Japanese and English, translations are usually rough approximations and lack expression and nuance. These shortcomings can be exacerbated by the style of the source. If the source prose is turgid and convoluted with complex subordinate clauses, or heavily idiomatic, the target will likely be nearly incomprehensible. It is best to keep sentences short, avoid ambiguous vocabulary and complex grammar, subordinate and relative clauses, idioms, and social jargon.

Here is an example of a public notice (edited slightly to protect the guilty) in Japanese and its target. Native Japanese speakers report that the source is obscure:

> **Example 9.**
> (*X-place*)開業で、*X*駅北周辺は、市自転車等放置防条例に基づき、「自転車等放置禁止区域」として指定されました。
>
> The city bicycle was based on the leaving prevention regulations, and the circumference of the north of X station was specified by the (X) opening "in such cases as the bicycle, leaving prohibition zone".

An amusing and effective example of the practical use of MT for writing occurred when, during a search for an MT package, this writer made an email inquiry to LogoVista, the company producing the MT package mentioned previously. The inquiry was in English and the reply was in English. While not perfect, the reply was fully comprehensible and more than adequate in detail. The concluding paragraph went on to explain that the note had been translated from Japanese by the very software under inquiry. This decided the sale. Similarly, the Japanese abstract of this paper was written using the same software and the Google Language Tools Webpage.

Typical operation of MT software

Most MT packages for desktops use the same basic screen layout, which is two large spaces, or "panes" side-by-side—one, usually the left side, for the source and the other for the target. The source text is entered and then one clicks on a 'translate' button and the target text appears in the other pane. Source can be entered directly from the keyboard, or pasted or dragged from another program such as a word processor or emailer, making it very easy to translate email notes. In the writer's university, for example, it is gradually becoming policy to use email for most intradepartmental correspondence, both for general convenience and so that non-Japanese users can use MT to help keep up.

Most Web-based, free MT systems, such as found on Google (2005) or Yahoo (2005), offer the same basic layout and can handle a text of up to several thousand words. Package MT systems offer refinements such as file-saving, integrated dictionaries, dividing the text into sentences, and a highlighting function that allows one to see translations of specific words.

MT in the classroom

By now, most of us have seen an example or two of homework written in a style that no student we have ever had would produce. Odd, obscure, sometimes archaic vocabulary is embedded in convoluted sentences in which "it" is often the subject. Passive constructions abound. All of these are evidence of MT. The student has likely written or pasted the text into a Web MT and then printed out or hand-copied the result. Clearly, this process does not fulfill the purpose of a writing assignment in a language class.

Here is an example of a paragraph written by a student directly in English, and an MT production from the Japanese version written by the same student.

> **Example 10.**
> *My hobby is fishing. I go fishing nearby ponds. I like to use big lure and to fish big basses. I always aim at sixty centimeters bass. But I fish fifty-five centimeters bass.*
>
> My hobby is fishing. I go to a pond nearby. I use a big lure, and it is good to catch a big bus. I always aim at a bus of 60 centimeters. However, it is not to have caught only bass of 55 centimeters.

A simple, and fun, way to discourage this kind of MT usage is to write a sentence in English that is a bit complex but understandable to the students and then show the MT result. As the laughter dies down, the students will see for themselves, in their own language, how rough MT is. Comparing the original source with the back-translated source is even more convincing.

Is there any use of MT appropriate for students? Perhaps, as long as it is not a substitute for learning. One possibility is to conduct simple experiments using the MT system. The student can try out various basic structures and see if the intended target is produced. Another might be to read advanced material—on the Web, for example—that the student would otherwise not attempt. Another use would be to conduct correspondence for some specific purpose, like the product inquiry mentioned before. More broadly, MT might be appropriate in cases where communicating the content is more important than the producing the form. If a Japanese student is writing a paper for an English publication, rather than struggling to produce original English, it might be more efficient to start with a Japanese version and then go through the refinement iterations.

Conclusions

MT is a practical and an accessible fact, and it is improving. The present state of the industry is that even Web-based free MT can be useful, especially if the user has some knowledge of the second language. Even those unfamiliar with the second language can at least get some sense of the source text. Important words in a translated text should be checked and key phrases rewritten and re-translated to confirm

meaning. Producing written material with MT requires a great deal of back-translation and rewriting. Nevertheless, with sufficient time spent, and especially depending on the user's familiarity with the source, MT can help save time overall and improve understanding, and can facilitate the writing of text in a second language. Language learners can use MT to assist with reading, and even to help check written assignments, but should not make this use a substitute for learning.

Frank Berberich teaches courses in English and Computer courses, in both English and Japanese. His graduate degrees are in ethnomusicology and psychology of music.

References

Arnold, D. (2002). Machine Translation: an Introductory Guide. [Online]. Available: <http://www.essex.ac.uk/linguistics/clmt/MTbook/>

Berberich, F. (2005). *Sharing Our Stories: With Translation Software? (Conference Presentation).* [Online] Available: < http://www.tokiwa.ac.jp/~frank/trans/>

Google (2005). *Language Tools. [Online]* Available: <http://www.google.com/language_tools?hl=en>

Logo Vista. (2005). *Promotional Page*. [Online] Available:<http://www.logovista.co.jp/ product/product.mac.html>

Yahoo. (2005). *Yahoo! Japan* 翻訳. [Online] Available: <http://honyaku.yahoo.co.jp/transtext>