Pre- and Post-editing MT text for better Communication and Documentation

S. Donomae
System Translation Group
Nagase & Co., Ltd

Introduction

Different from artistic writings, writings in business and industrial circles are to send or receive a message using symbols for conveying, confirming, arguing, discussing facts, data, idea, thoughts and opinion to let others do something or not let them something for some practical reason (mostly making money or avoid losing money). When these writings done in non-native tongue, translation occurs unless the person involved is a fully bilingual or bi-cultural. Many of such translated message most often have to be documented to share information by the people concerned.

For most people other than professional translators or interpreters, it is a time-consuming and pains-taking job for one to convert one language to another even he or she have had long foreign language education at school (6 to 8 years in Japan). Human intelligent capacity and time to spend to acquire foreign language reading and writing skill is very limited; there are more important things have to do in business and private life. Nonetheless, most people wish to be, or requested to be good command at English language as an international means of communication and documentation. Although this is particularly important for such people as Japanese who are completely outsider of European culture (European culture represents all western civilization including the United States of America, North and South American Countries, some part of middle East and Asia). and yet have to use English as a means of her survival. There is, however, one solution for this controversial and difficult problem: that is MT!

With above background, we have started using MT since 1991 when the first commercial MT WS was available for the purpose of translation and documentation of English literature of technical context. The first 6 month to one year was spent for compiling technical dictionaries and developing
pre-editing and post -editing skill for English text (customization of commercial system)

to increase the efficiency with the improved accuracy of translation.

Two years later, we started Japanese to English MT and found out that pre-editing of Japanese text was absolutely necessary to get the result make sense. For the past one year, we tied up with Kyushu Matsusita Electric Co., Ltd. devoted ourselves to develop integrated translation and documentation system because visual aids such as pictures, graphs, charts, diagrams, drawings, numerical and symbolic expressions sometimes play more important role in the documents for physical worlds rather than mere character strings. This new highly integrated graphic interface MT system is now commercially available under the Trade Name of "Honyaku Koboo"- literary means "Translation Laboratory". At the same time, thanks to the progresses in Personal computer, we place more emphasis on PC-based MT for data-communication purpose and plant to on market our own brand name PC-MT software under the Trade Name of "Uchino Honyakuya-san"- literary means "My Dear Translator".

For the purpose to briefly introduce you some of our basic idea how to pre-edit English text and post-edit the resulted Japanese text in English Japanese MT and vice versa., the following experiment was carried out using Toshiba's Transac as MT environment and the first few pages of "Sire- a year in the life of the Belgian royal family" published in 1985 by lannoo tielt weep as Text both in English and Japanese (translated by Mz. Mitue Hisatune). The choice of the text is just for "commonly interested topic" not for any other reason than to show how differ MT from a professional human translation and how to narrow the gap through pre-and post editing the text.

Experimental

The text has been produced by Henry Van Daele and translated by Mz Miutsue Hisatune at the initiative of INBEL, Brussels, and of the Lanno publishing company, Tielt, in collaboration with the Ministry of External Relations, the General Savings and Pensions Bank, the Belgian General Commissariat for the Tukuba Exhibition of 1985 and Agfa-Gevaert.

The first few page of the text titled "A Parliamentary Monarchy" in English and Japanese equivalent "議会君主制" were given to two my assistants of about same educational background - 8 years English learning at school with the last two years in major in English,

one for English text to Japanese and one for Japanese to English without telling there is an counter version and asking translation without consulting dictionaries and checking how many words or phrases each doesn't know.

The both texts in English and Japanese was degitalised for MS-DOS text file for MT. First, the texts were machine translated with only basic dictionary without making any modification and compared with original ones. Then they were machine translated using appropriate technical dictionary we have already compiled and user dictionary for adding some words and phrases. The key in this operation is in how effectively registered the phrases, not words, sometime clauses, in the equivalent in other language.

Results and discussions

Before touching on the results and discussions on the subject, I wish to simply review the most basic and common to all commercially available editing function of MT as follows:

1. Segmentation or division of a long sentence to shorter ones.

As there is some limitation in a length of sentence being able to be effectively machine translated, it is necessary to make segmentation of long sentences usually consisted of more than 30 words into two or three phrases or clauses according to sense group marked with comma, colon, semi-colon, etc. as is shown in the following example:

Example:
(E) In the meantime, please accept our sincere apologies for the inconvenience we may have caused.
(J) 合間に｜どうぞ、我々の誠実な謝罪を受け入れてください｜の順｜我々が引き起こしたかもしれない不便

This sentence can be segmented into three parts of |in the mean time please accept our sincere apologies for the inconvenience we may have caused and machine translated as 「ところで」、「どうぞ、我々の誠実な謝罪を受け入れてください」、「我々が引き起こしたかもしれない不便のために」。It is very easy for an average Japanese to rewrite this result into more natural expression as:

(R) ご迷惑をおかけしたことを心から深くお詫び申しあげます。

2 Clarification of a part of speech

Most often MT fails to specify the correct part of speech in the sentence such as is in the case of as follows:

Example:
(E) He then got back into the truck.
(J) 彼は、そのときバックをトラックの状態にした。

When specify "back" as adverb, we will get the result make sense as follows:
(C) 彼はそのときトラックに戻った。

3. Specifying phrase

Though it is easy for a human to grasp the cluster of words as a unit of sense as phrase, sometimes it is very difficult for MT to a certain word combination as phrase unless there is no dictionary entry as is the following sentence:

(E) Because of moving into new offices and family illness, we were unable to advise earlier about this matter.
(J) 新しい事務所、そして、家族病気に働くことのために、我々はこの件について早朝あなたを忠告することができなかった。
By specifying "moving into new offices " as noun phrase, we can get the result make sense as follows:
(R) 新しい事務所への引っ越しと家族の病気のために、この件について事前に連絡できなかった。

4 Specifying the sentence pattern

In order to let computer understand the structure of a sentence for correct and quick translation as is the case of the following sentence, specifying the sentence pattern is very effective:
Example:
(E) This offers a patient sharper vision.

(J) これは、患者詐欺師視覚を提供する。

When specifying the above sentence as type IV, we can get the correct translation:

(c) これはさらに鋭い視覚を患者に提供する。

It is easier to rewrite the above into more natural Japanese as follows:

(R) この製品を使えば、目の悪い人でも、もっとものをはきっりと見ることができる。

5. Specifying a part of speech of the phrase

As same as specifying a part of speech and phrase, we can get acceptable translation as is the case of the following sentence:

Example:

(E) Here is a new book for you to read.

(J) ここにあなたが読むために、新しい本がある。

When specifying "for you to read" as adjective phrase to modify "book", the translation will be:

(c) ここにあなたが読む新しい本がある。　And then, rewrite to:

(R) 君が読んだら良い本だ。

6. Specifying parallelism in the sentence:

One of the weakest point of MT is in its inability to identify parallelism in the sentence:

Example:

(E) Both in Japan and abroad, we have many installations for industrial and municipal water and waste water.

(J) 日本の双方とも｜そして｜海外へ｜我々は、産業の、市の水、および、廃水のための多くの設備を持つ。

By specifying "in Japan" as adverbial phrase and as parallel with "abroad", MT will produce the following correct sentence:

(c) 日本において、そして海外へ、、以下同じ。Then we can rewrite it as easily as follows:

(R) 国内においても、海外においても、当社には、産業用水、飲料水、廃水処理用に多くの販売（据え付け）実績があります。

7. Specifying non-translation

Unless registering in the dictionary, MT cannot distinguish proper nouns from common nouns, thus it translates Mr. Smith as 「鍛冶屋氏」 and Bob John as 「ジョンを切れ」 as is in the following case:

Example:
(E) President Bush may have to speak out fast to avoid such a fate.
(J) 大統領潅木は、そのような運命を回避するために、速く明言しなければならないかもしれない。
By specifying "Bush" not to translate, and replace all "Bush" in the text as 「ブッシュ」
And then rewrite as follows:
(R) ブッシュ大統領は、そのような事態に陥らないように態度を速く表明せざるを得ないであろう。

8. Specifying the insertion of omitted word(s) or Phrase(s).

There is some difference in the way of the eclipses of word(s). Unless otherwise to insert omitted word(s) into right place, the translation will make no sense:
Example:
(E) I will do all I can but you must know you have left it very late.
(J) 私は、すべてをするでしょう｜私｜缶｜あなたは、あなたがそれを非常に遅い状態にしておいたということを知らなければなりません。
By inserting "do" in between "can" and "but", and rewrite it as follows:
(R) 私にできることは何でもいたしますが、御連絡を頂くのが遅すぎました。

All above examples is the case of translation of English text into Japanese for assimilation where it is more important what is written than how it is expressed. And many instances, there is no need to rewrite the machine translated Japanese to natural fluent Japanese unless such translation is done for documentation or publishing. Although eight basic pre-editing skills are introduced, the most commonly used one in daily operation is segmentation and entry of user dictionary for the exact meaning for the combination of words according to the text. If key words or phrases are correctly and accurately translated into Japanese, some other syntactical errors or grammatical mistake is not so serious because human brain automatically works to correct it for assimilation. In this respect, MT is very useful tools for cultural assimilation.

Most English text for MT is well written and edited by the well trained writer; however ,

we occasionally encounter the problematic heavy writings needed to be rewrite before MT processing. In this case, the following 5 points should be considered as the factors that makes writings difficult to understand in English.

(1) Big words or phony fancies

(2) Nounry

(3) Passive voice

(4) Wasted words

(5) Grammatical errors.

Though it is not main objective to get in touch too much details of the subject, it is worth remember that the rewriting or the deletion of a part of original text in accordance with above greatly improve the quality of machine translation.

(1) Big words or phony fancies are those fuzzy words or beurocratic, pedantic expressions make the writings difficult to understand. Just rewriting such words as "provides for the requirement that you---" or "provides a means whereby you may---" to simply "enables you to--" is significantly improve the clarity of the resulted Japanese.

(2) Nounry or overuse of nouns sometime destroy communication as is the sentence " All aspects of the situation should be taken into careful consideration prior to the implementation of any corrective action. When this is rewrite to " Now, don't change anything here until you checked it thoroughly", the better translation will be resulted. Also rewriting beurocratic, pedantic euphemism as those " interior intrusion detection systems," "cost impact consideration systems," " cost of living adjustment modifications" to "burglar alarms," "price," "rising prices," respectively will be resulted in better communication from practical view point.

Try to avoid turning verbs into nouns as in the cases of " have tendency to," "give encouragement to," express appreciation," "take us consideration," "be in agreement," make and examination of," make mention of," "take recognizance of," "in receipt of," and use a verb as a verb instead turning into nounry expression.

(3) Passive voice sometime causes ambiguity in the meaning as is the case of "The picture was identified as the one taken during the operation by the radiologist". This sentence may have two meanings; (a) The radiologist identified the picture as the one taken during the picture. (b) The picture was identified as the one the radiologist had taken during the operation. As a rule, it is better to use active voice unless there is a clear need to use the passive voice for better communication.

(4) Wasted words or redundancy makes deteriorate the quality of the machine translation. It is better to eliminate these wasted words before processing MT. These words "angle," "area," "aspect," "case" "character" "circumstance" "factor" "function" "field" "happen" "lines" "nature" "point" "respect" "situation" "thing" "situation" "type" "variety" etc. tends to be added meaningless to other word combination. Better to take out for clarity.
Also it is better to take out redundancies such as "if and only it" "each and every man and woman" "and so as a result" "and moreover" "but nevertheless" "never at any time" "result so far achieved" "make an effort to try" "at this point in time" draw final conclusions" "a hypothetical situation that does not now exist" etc.,. Just write respectively as "if" "every one" "and so" "moreover" "nevertheless" "results" "try" "now" "conclude" "a hypothetical" etc.,.

(5) Grammatical mistake often doesn't matter for human beings except in classroom but is serious for MT processing. Most commonly found mistakes to be corrected before MT processing are (a) misplaced modifiers, (b) dangling modifiers, (c) unclear agreement, (d) faulty pronoun reference (e) unparalleled construction (f) faulty punctuation. Thanks to long unpractical English education at school in Japan, most Japanese have ability to correct these grammatical errors.

Besides above, it seems very difficult for MT understands parallelism in syntactical point view as well as idiomatic expressions, proverbs, quotations written in classic Greek and/or Latin from semantic view point, and some basic knowledge of science, technical, financial and business field, for which human brain must assist MT making use of intelligence and/or knowledge for interpretation and understanding the true meaning behind expressions. That is the task for us instead of memorizing English vocabularies or detailed grammatical rules mechanically . Let these mechanical jobs a computer do. It dose incredibly good job with unbelievable accuracy and the high speed. This is the reason why we are using MT.

Different from the translation of Foreign Language into Mother Tongue of which purpose is assimilation, the purpose of the translation from Mother Tongue into Foreign Language is dissemination where how to express is the major concern, So the different approach must be taken for pre-editing the Japanese text before MT processing. More than 80% of Japanese text for MT processing must , smaller or greater, be rewrite to more simple and direct form in order to get the resulted English make sense or at least readable. This

seems to be caused by the failure of morphological and syntactical analysis of Japanese text due to too much flexibility and complexity of Japanese language and varieties in signifying it with 4 deferent characters or mode of writing, i.e., 漢字、ひらかな、カタカナ、and Roman Alphabet.

In addition, Japanese grammar now adopted at school since 1800s as the result of Europeanization of the country seems unpractical due to facts that ours is completely out of the category of Latin Grammar and the different way of thinking and feeling from the logical and analytical way of yours. Generally, Japanese sentence tends to be longer and there is no fixed rule for punctuation and paragraphing. If we read a series of periodic short sentences, we feel them childish and lack of intelligence. Yet we have the shortest form for poet (Haiku-17 characters segmented 5, 7,5 and Waka- 31 characters). And so many people involves creating Haiku or Waka. Until Meiji Restoration, we were strongly under the influence of Chinese Culture as you owed much t o Greek and Latin. Yet we developed very unique culture though the history due to geographical and mental isolation from the Continent (China). These cultural background makes an average Japanese with long English education difficult to express his or her idea in English or any other foreign language even though he or she could read and understand the text written in foreign language. Thus MT user's point of view, English/ Japanese MT is quite useful and help very much for assimilate while Japanese/ English MT seems not so much successful though there is strong demand.