What is Missing in User-Centric MT?

Jennifer DeCamp
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22042, USA
jdecamp@mitre.org

Abstract

This paper describes some of the kinds of predictable errors in Machine Translation (MT). It then discusses means of alerting end-users of MT to the possible presence of such errors, including by providing training and/or by providing automated MT ratings, MT color coding and/or symbols, and footnotes and annotation. It also discusses the need for some kind of reliability measure and/or information to the MT consumer, and the likelihood of the MT user being open to using this kind of input. Some of the suggestions made for user-centric MT are also applicable to translator-centric MT.

1 Introduction

What is missing in MT? Some text may not be translated. Some relationships may be reversed. Some names may be wrongly translated. Some negatives may get lost. However, the text may read reasonably well, and the consumer may not realize substantive errors that may affect his/her understanding and decisions. This paper addresses some of the types of consistent errors and proposes means for communicating this variation in reliability to the consumer of Machine Translation (MT) output. Some of the suggestions made for user-centric MT may also be applicable to translator-centric MT.

2 What is User-Centric MT?

User-centric computing is a phenomenon that has emerged primarily in the last decade users are searching for, deciding on, and often translating the information they need. As Van der Meer observes in an eloquent description of user-centered computing (1994), "The one source of information provided by the product manufacturer, the government, the doctor, or the hospital is now being replaced by dozens—if not hundreds—of alternative and competing information sources. Tips and tricks from other users, prescriptions from multiple healthcare organizations, and analyses of government data from private sources may be much more valuable than the 'authoritative' information from the original publisher." These users then employ online MT to access the information in their language of choice: hence, "User-Centric MT".

Microsoft was one of the notable pioneers of this approach for product literature, providing the MT developed for internal use at their company to premium users as a perk. The MT enabled the users to translate and thus access larger sections of the online Microsoft website.

Government organizations are also making greater use of User-Centric MT. Bemish (2008) observed that "Using advanced tools like MT has allowed analysts and investigators to see data that would have taken years to translate and compile."

3. What do MT Providers Do for Users?

In the past few years, the Association for Machine Translation of the Americas (AMTA) and the MT Summit have provided an increasing focus at their conferences on providing tools for translators—primarily for post-editing—thus creating translator-centric MT. However, little has been done to support the users who are not translators and who are utilizing MT from free sites such as Bablefish, Altavista, or Systran's own websites, or from other free sites such as Google Translate, Free Translation (SDL International), ProMT, Gist-inTime, PARS, Microsoft Windows Live Translator, and others, or from intranet and/or licenses resources.

There are a few exceptions. LanguageWeaver added a confidence rating to some if its systems in 2007. However, Gerber comments that "not a lot of attention was drawn to it, and I believe they have never gotten any feedback on its usefulness." Systran for many years has enabled users to add

their own terms to their online MT at www.systranet.com.

However, tools for the end-user of MT seemed to have received little attention and/or to have fallen off the community's radar. There are many reasons for this lack of focus. One reason provided by Gerber (2009) is that "users (and more importantly buyers) don't demand such tools." Of course, if the users are unaware of such tools, they are unlikely to ask for them.

In addition, much of the user-centric MT has been with free MT systems on the Internet, so there has been little incentive for MT companies to commit additional development resources to provide tools. Some of the MT—such as Systran's free resources—was put online not for production purposes but for education. As Gachot (2005) pointed out, users became more knowledgeable about MT by playing with it.

Gerber also comments that "MT developers are aiming at so many different user environments, it can be hard to figure out which environment/users to target." Tapling (2008) pointed out that the MT field has been segmented by technology rather than by user needs. Perhaps as this focus shifts and as the volume of MT increases the feasibility of market segmentation, tools can be better targeted.

Even so, there is a significant market segmentation of people other than translators and post-editors using online MT systems. Each of these users has a stake in knowing the reliability of the MT output. Moreover, the fact that these users are employing MT indicates that many may not know the foreign language or have the time or resources to otherwise assess the reliability of the translation.

It may be that MT providers do not believe users are ready to accept such tools and may even be turned off the use of MT by being presented with too many caveats. The last decade has been characterized by considerable growth in the sophistication of users concerning computer tools and concerning realistic views of MT. A couple of weeks ago, a translator commented that her customers used to think that footnotes decreased the readability and thus the usability of translations, but that they now like footnotes.

Another reason may be a perceived lack of appropriate tools and underlying research. The automated MT evaluation tools at the forefront of MT assessment (e.g., BLEU, METEOR, etc.) require gold-standard reference translations of the same material. Such tools are thus probably not feasible for assessing the reliability of new translations where a reference translation is not available. These tools also oriented towards evaluating software development rather than the communicative value of a text, although new work on task-based metrics (e.g., Friedman and Strassell 2008) in the future may provide automated ratings more useful to end users.

There may also just be too many problems in MT to correct. It is significant that the Pan American Health Organization (PAHO) only color codes items that they are certain are correct (e.g., that are perfect matches in a Translation Memory or that come from an organizational terminology; Gerber 2009). To provide tools to correct all problems is not feasible. The only means of reasonably ensuring that all problems are addressed is to employ an excellent post-editor (i.e., a human) and preferably also an excellent second editor. Even so, the diminution of significant errors that may cause misunderstandings and bad decisions may still be a benefit to the users.

There are also those of us who are very concerned about unreviewed MT being used for any decision-making, due to the many problems with quality and reliability. However, despite our astute advice, people are increasingly using raw MT output.

One further reason for the lack of focus on user tools may just be the research focus that has permeated the MT community, particularly in the United States. For instance, in a presentation at the 2008 Conference of the Association for Machine Translation of the Americas (AMTA), Chang-Meadows described consistent errors with the Chinese particle "de" (的), resulting in confusion about who is doing what to whom or who reports to whom. When I raised the question of whether users could be alerted to such problems, the response from the DARPA program manager and his team was that the problem had been fixed. However, while the problem had been fixed from a research standpoint, it is still not fixed in the MT systems that are available to commercial and most Government users.

Part of this research focus and drive has been to provide the improved MT as opposed to providing the user with explanations of what is wrong or missing or with tools for the user to fix the prob-

lem himself. In addition, from a research and development standpoint, these problems are well known. They are old news and not cutting edge research.

In any case, it may be a good time to review ways to help the users of MT.

4 What is Missing in MT?

There is a wealth of information in the MT research, development, and post-editing communities concerning common and predictable problems of MT—including of specific MT systems. The following examples are a few from a study conducted by Chang-Meadows of comparative output of Google, Microsoft Translate, and SYSTRAN Chinese-to-English MT (2008).

4.1 Change in Subordination

Chang-Meadows found predictable errors in the use of the Chinese particle "de" (的), resulting in confusion about who is doing what to whom or who reports to whom.

For instance,

Original:

华建集团中国科学院直接投资成立的高科技企业

Human Translation:

The Huajian Group is (a high-tech enterprise invested and established directly by the China Academy of Sciences).

Google:

Hua Jian Group is a direct investment in the establishment of the Chinese Academy of Sciences of the high-tech enterprises.

The Google MT version could be read as the Hua Jian Group investing in the Chinese Academy of Sciences instead of the reverse, as in the human translation.

4.2 Blank Space

One high-risk practice in several MT systems is to omit text with no indication that something has been omitted. In LanguageWeaver MT, for instance, the default setting for handling unknown words is to simply omit them from the text.

The Microsoft translation for the example above was: Hua - group was direct investment set up high - tech enterprises", which omitted any reference to the Chinese Academy of Sciences.

A second example is as follows, where the Google example omits the name of the enterprise:

Original:

大三通是目前中国最大的GPS连锁企业和营运成绩最好的企业

Human Translation:

Dasantong is China's (largest GPS chain enterprise in China) and (the enterprise that has the best operational results.)

Google:

At present, China is the largest chain of businesses and operating GPS the best of the enterprise

4.3 Names, Acronyms, and Abbreviations

There are fairly consistent problems with names, acronyms, and abbeviations. In the example above, "Dasantong" was problematic and omitted. In some instances, the proper noun can be translated. For instance, in the example below, the Systran MT system translated the "Lanya" in the name as "blue", changing "the Wuhai City Lanya Chemical limited liability company" to the "The Wuhai blue Asia chemical industry Limited liability company".

This example also shows the predictable distortion in translation of proper nouns:

Original:

乌海市兰亚化工有限责任公司

Human Translation:

Wuhai City Lanya Chemical limited liability company

Google:

Wuhai City LAN Ya Chemical Co., Ltd.

Systran:

The Wuhai blue Asia chemical industry Limited liability company

Microsoft:

Wuhai LAN Asia chemical co., Ltd.

4.4 Convoluted Complex Text

As Chang-Meadows points out, MT predictably does less well on convoluted and complex text:

Original:

该实验室多年来一直致力于环境工程和试验技术、可靠性工程和试验技术、环境测量分析和预计技术、电磁环境效应等方面的探索和研究工作，同时为各行业提供了大量的环境与可靠性试验服务。

Google:

The lab has for many years been committed to environmental engineering and test-

ing technology, reliability engineering and testing technology, environmental analysis and measurement is expected to technology, electromagnetic environmental effects, such as the exploration and research work, while for the industry to provide a large number of environment and reliable Test service.

Systran:

This laboratory has for many years devoted to the environment project and the experimental technology, the reliability project and the experimental technology, the environment survey analyzes and estimated that technical, aspect and so on electromagnetic environment effect explorations and the research work, simultaneously have provided the massive environment and the reliability test for various professions serve.

Microsoft:

The Laboratory efforts in environmental engineering and pilot technology, reliability engineering and pilot technical, environmental measurement analysis and estimated technology, electromagnetic environment effect aspects in the exploration and research work, at the same time for various industries provides a number of environmental and reliability testing services.

5 What Works Well?

As researchers and many editors point out, what works well with MT is simple structure and factual information.

5.1 Simple Structure

Bernth and McCord (2000) conducted studies showing the impact of simplified text on translation quality. Shubert and Spyridakis (1995) and Spyridakis, Homback, and Shubert (1997) showed that in many cases, the use of simplified English (as can be measured automatically) can improve HT results.

Consistent with this analysis was Chang-Meadows (2008) analysis of the best performance of Chinese-to-English MT output of Google, Systran, and Microsoft. She found that the best output occurred with simple parallel structures:

Original:

生产场地宽敞整洁, 生产设备一流, 生产技术先进

Google:

Production sites spacious and clean, first-class production equipment, advanced production technology.

Systran:

Produces the location spaciously neat, production equipment first-class, production technological advance.

Microsoft:

production venues spacious clean production equipment first-class production technology, advanced.

5.2 Factual Information

Good output also occurred with simple factual information about personnel, assets, and services:

Original:

集团公司拥有研发、流通和生产企业140余家，并在全球数十个国家和地区建立了近百家海外分支机构。至2007年底，资产总额近1500亿元，主营业务收入突破1300亿元，员工30万人。

Google:

Group owned research and development, production and circulation of more than 140 enterprises, and dozens of countries in the world and the establishment of nearly 100 overseas branches. To the end of 2007, with total assets of nearly 150 billion yuan, the main business income of 130 billion yuan breakthrough, employees 300,000 people.

Systran:

The Group has the research and development, the circulation and Production enterprise 140, and has established nearly hundred overseas Branch office in the entire nodule number ten countries and the area. By the end of 2007, the gross asset nearly 150,000,000,000 Yuan, the main business income tops 130,000,000,000 Yuan, the staff 300,000 people.

Microsoft:

owns r&d, circulation and production enterprise 140, and in the global dozens of countries and regions have established nearly 100 overseas branch offices. to the end of 2007, the total assets of nearly 1 500 billion, the primary business income

breakthrough 1,300 billion, an employee 30 000 people.

6 What Can We Do?

There are numerous strategies that could be tried to help users of MT manage their risk, including providing training, providing ratings, marking errors or high-risk output, providing tools to the user to evaluate the likelihood of errors given input, providing ratings, and/or providing footnotes and annotations.

6.1 Provide Training

One risk mitigation strategy would be to provide training to users of Fully Automated Machine Translation. The poor readability of FAMT used to be at least some warning to readers to be careful of using the results. However, the improvements in readability—particularly with SMT—have now increased the risk of users over-trusting the results.

Some U.S. Government MT systems provide a statement on the coversheet of the translation that the contents are machine translated and should be used with caution. However, there is currently no guidance on how to use those materials. What may be helpful for MT sites in general is a description of what to expect from the MT output and tips on how to improve the output by changing the input, in situations where changing the input is feasible.

There is still very little public training in understanding MT output. Free online MT services have enabled people to play with MT and to recognize both the potential and a few of the problems. However, limited play with a few usually short phrases is not sufficient preparation for using MT for real decision-making.

There are many efforts to provide language technology training, such as the Multilingual E-Learning in Language Engineering (MELLANGE) project (part of the European Leonardo da Vinci program) and the Localization Industry Standards Association (LISA) Education Initiative Taskforce (LEIT). Such efforts, however, focus on the translators and language technology specialists and not on the average user of fully automated machine translation.

Teaching the general public how to better understand and use MT may be good goal for professional organizations such as AMTA, its international counterparts, and the MT Summit to undertake during the next few years.

6.2 Provide Ratings

There have been numerous efforts to develop rating systems for machine translatability, as was discussed previously regarding LanguageWeaver and IBM. Uchimot, Hayashida, Ishida, and Isahara (2005) developed a system for rating MT quality without reference translations, specifically by using bidirectional translations. Many users of online MT have invented their own informal means of checking translation accuracy by using backwards MT. Of course, the use of bidirectional translations often creates new problems, since one translation pair is rarely the exact inverse of the reverse pair.

Clifford, Granoien, Jones, Shen, and Weinstein (2004) analyzed machine translation quality was affected by the level of text difficulty (as measured by the Interagency Language Roundtable Proficiency Scale). Various pre-editing and authoring systems also provide information on whether a document will translate well, as is discussed in the next section.

In the meantime, it may be possible to construct an automated rating system to help users based on the absence of problems in the source text that would be likely to create problems. Thus a source text with simple direct phrases and no known problems (such as “de”) in Chinese might get more stars or smiley faces than a convoluted sentence with some of the problems discussed earlier in this paper.

Providing overall confidence ratings presents significant problems, since as Egan (2008) points out, “A single error/omission/deletion can seriously compromise the utility of a particular translation even when judged 70% or 80% accurate” by some of the popular scoring methods such as BLEU.
In addition, some kind disclaimer would may need to be provided concerning the ratings, since the MT providers and raters would want to avoid legal liability for the MT (e.g., if the MT provided wrong information about product capabilities or prices).

6.3 Mark Input

Xerox in the early 1980s developed software to check source text and make recommendations to

writers about improvements to source text (e.g., shortening sentences) that would provide a more reliable MT output (Ruffino 1982; Ryan 1993). This type of checker—or even some of the analysis behind it—could be provided to that subset of consumers who are in a position to change the source text.

Bernth and Gdaniec (2001) identified characteristics of English text that resulted in higher quality. There are also a number of authoring systems such as Smart's MaxIT, Acrolynx, and AuthorIT which are designed to help authors write better input for MT. Some of this work could be tailored for this community.

6.4 Mark Output

There are many forms of MT markup that could be provided to users. Xerox Corporation in the 1980s color coded the output of MT to indicate areas needing post-editing by human translators. The marking was primarily on the basis of non-matches with a rule based system (SYSTRAN). SYSTRAN used to include include markup of their Russian-to-English system used by the National Air and Space Intelligence Center (NASIC). However, the marking could be expanded to reflect a broader array of potential errors.

6.5 Provide Footnotes and Annotation

Another method of improving the reliability of MT is to follow a common practice in human translation: to provide footnotes and/or inline or linked annotation. For instance, where a term does not have a direct equivalent in the target language, human translators frequently provide a footnote explaining the term. It would be possible to not only automate this process for FAMT but also to expand the footnotes and annotations to include warnings of common problems.

7 Conclusion

User-centric computing has changed the paradigms for at least one major segment of our MT user community. Users with little or no background in the source language or in MT are conducting a significant amount of machine translation, often to use for decision-making. As a community of MT professionals, we need to better educate these users on what they are receiving and on what they are missing. We also need to examine how we can better provide them with the kinds of tools now being used by researchers, authors, and post-editors—or better yet, more tailored tools—in order for them to at least better understand the quality of the translated information.

Acknowledgements

I would like to express my appreciation to Shin Chang-Meadows for her many outstanding examples of problems in Chinese-English MT. I would like to thank Shin and Laurie Gerber for their review of this paper. In addition, I would like to thank the United States Defense Intelligence Agency Foreign Language Program Office for sponsoring my participation in the MT Summit.

References

JuliaAymerich and Hermes Camelo. 2006. Post-Editing of MT Output in a Production Setting. *Proceedings from the Association for Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications.* Cambridge, MA.

Nicholas Bemish. 2008. Can MT Really Help the Department of Defense? *Proceedings from the Association for Machine Translation in the Americas (AMTA 2008).* Cambridge, MA.

Arendse Bernth and Claudia Gdaniec. 2001. MTranslatability. *Machine Translation* Vol 16, 3, 175-218.

Will Burgett and Julie Chang. 2008. The Triple Advantage Factor of Machine Translation: Cost, Time-to-Market and FAUT. *Proceedings from the Association for Machine Translation in the Americas (AMTA 2008).* Cambridge, MA.

Shin Chang-Meadows. 2008. MT Errors in CH-to-EN MT Systems: User Feedback. *Proceedings from the Association for Machine Translation in the Americas (AMTA 2008).* Cambridge, MA.

Ray Clifford, Neil Granoien, Douglas Jones, Wade Shen, and Clifford Weinstein. 2004. The Effect of Text Difficulty on Machine Translation Performance – A Pilot Study with ILR-Rated texts in Spanish, Farsi, Arabic, Russian and Korean". *Proceedings of Language Resources*

and Evaluation Conference (LREC 2004), Lisbon.

Kathleen Egan. 2008. User-Centered Development and Implementation. *Proceedings from the Association for Machine Translation in the Americas (AMTA 2008).* Cambridge, MA.

Lauren Friedman and Stephanie Strassell. 2008. Identifying Common Challenges for Human and Machine Translation: A Case Study from the GALE Program. Proceedings *from the Association for Machine Translation in the Americas (AMTA 2008).* Cambridge, MA.

Denis Gachot, 2005. Personal conversation.

John Hutchins, 2001. Machine translation and human translation: in competition or in complementation? *International Journal of Translation*, 13(1-2), 5–20.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02).* Philadelphia, PA, 311-318.

Richard Ruffino. 1982. Coping with Machine Translation. *Practical Experience of Machine Translation.* Lawson (ed.) 57.

Serena Shubert and Jan Spyridakis, The Translatability of Simplified English Documents. Matching Information to Audience. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00554897

JoAnn Ryan. 1993. Machine Translation: Matching Reality to Expectations. *Progress in Machine Translation*, ed. Sergei Nirenburg . Amsterdam: IOS Press, 225-235.

Greg Sanders. 2006. Post-Editing in the GALE-Program I. *Proceedings from the Association for Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications.* Cambridge, MA.

Jan Spyridakis, Heather Holmback, and Serena Shubert. 1997. Measuring the Translatability of Simplified English in Procedural Documents. IEEE Transactions on Professional Communication, Vol 40, No 1, 4-12.

Stephanie Strassel. 2006. Post-Editing in the GALE Program II. *Proceedings from the Association for Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications.* Cambridge, MA.

Kiyotaka Uchimoto, Naoko Hayashida, Toru Ishida, and Hitoshi Isahara. 2005. Automatic Rating of Machine Translatability. No publication information provided. MT Archive PDF http://www.mt-archive.info/MTS-2005-Uchimoto.pdf.

J. van deer Meer, J. 2006. The Emergence of FAUT: Fully Automatic Useful Translation. In Keynote at the 11th Conference of the European Association for Machine Translation. Oslo, Norway.