Using Machine Translation for the Localization of Electronic Support Content: Evaluating End-User Satisfaction

Osamuyimen Stewart, David Lubensky, Scott Macdonald, Julie Marcotte
International Business Machines (IBM)
1101 Kitchawan Road, Route 134
Yorktown Heights, NY 10598
{ostewart;davidlu;scott_macdonald;jhmarco}@us.ibm.com

Abstract

This paper discusses how to measure the impact of online content localized by machine translation in meeting the business need of commercial users, i.e., reducing the volume of telephone calls to the Call Center (call deflection). We address various design, conceptual and practical issues encountered in proving the value of machine translation and conclude that the approach that will give the best result is one that reconciles end-user (human evaluation) feedback with web and Call Center data.

1 Introduction

In a globally integrated (multinational) enterprise, like IBM, with linguistically diverse teams and customers, there is usually a deluge of online content that grows faster than with which the finite set of professional human translators can ever cope. These companies are faced with the ever increasing cost of translating online content, the majority of which is authored in English, as they continue to seek ways to mitigate the cost of translation. Furthermore, the translation problem is compounded because, oftentimes, on-going translations become obsolete even before the translation cycle is complete and the content is published. In general, a distinction is made between two categories of online content needing translation: formal (e.g., legal, contracts, etc.) and general (e.g., technical support, How To, etc.). The common trend inside the enterprise is too retain the services of professional human translators for translating the formal content (because it requires a higher degree of accuracy and rigor), and take advantage of some of the recent advances in statistical machine translation (SMT) for automatically translating the general support-related or How To content (because these required lesser accuracy and rigor). In this paper, we focus on the latter, specifically on the Electronic Support (e-Support) domain. e-Support provides online content to its customers that will enable them resolve IT-related issues on their own without having to place a call to the Support or Help Desk.

2 Description of the Problem

Consider the following typical scenario or use case for machine translation in e-Support: A customer (in China, Japan, Italy, etc.) is working on a time-sensitive task, and she encounters an issue that requires re-installing software. She reckons it will take too long to wait on the phone to get help from technical support (who most likely does not speak her language) and so elects to attempt to resolve this on her own by visiting the e-Support site to get the installation procedure and instructions. After a successful web search for the right material (authored in English), she translates the relevant content into her local language (using automatic real time machine translation), and uses the localized content to proceed to re-install the software. In this regard, the end user is able to consume content on the web in her local language based on machine translation, without having to make a telephone call (call deflection). In this scenario, end-user satisfaction is based on the ability to use the localized content to resolve the problem. However, it should be noted that machine translation is a derivative process that relies on the accuracy and/or effectiveness of the underlying (source) content (usually authored in English). Therefore, one challenge that is often encountered while trying to measure the impact of machine translation is how to assign end-user satisfaction value to such an embedded or derivative event. More specifically, the challenge is how to show the stakeholders that the quality of service (QOS) for the localized (translated) content can be measured in terms of the business need or service level agreement (SLA). In the case of e-Support, these include the percentage of client self assist through machine translated content and consequent reduction in the volume of telephone calls (call deflection).

In the Call Center, there are standard metrics for measuring end-user satisfaction including counts of first call resolution, call completion, call abandonment, etc. Similarly, in web-based client self assist solutions, end-user satisfaction may be determined from an analysis of web clicks (documents accessed). This is typically derived from the percentage of users (through post-transaction online surveys or context-sensitive "rate this content" feature) who state that the web content was sufficient in resolving their problem and avoiding making a telephone call. However, when you attempt to replicate a similar process for determining end-user satisfaction with machine translated (localized) web content, there are several problems encountered. First, machine translation is not perfect (even for the best tuned domain) and so what you see is not always on a par with human translation. Second, the effectiveness of the machine translation output is limited by the quality of the source material. Thus, errors in the underlying material (source) will obviously depreciate the value of the translated content. Third, end-user judgments are not granular (i.e., they are not determined on a sentence-by-sentence basis) but rather are holistic. Consequently, end-user satisfaction with localized web content (similar to telephone-based self service) spans all aspects of the interaction or translation. This includes perceptions of translation effectiveness across the document(s) viewed, from the first click (first page viewed) to exiting the site (irrespective of whether they were successful, or not, in resolving their issue). Additional complexity is introduced into the assessment of online end user satisfaction because the class of service that commercial users are assigned and which govern their web experience is usually opaque to the user. Thus, the back-end decisions pertaining to user experience and type of service which are not readily obvious to the commercial end-user, often factor into their overall evaluation of the effectiveness of localized content. All of these factors point to some of the fundamental issues surrounding how end-users evaluate machine translation. When deploying practical real-world systems like e-Support, we need to provide a way to communicate end-user satisfaction with results or outcomes that can be used to readily measure the QOS such that the stakeholders can quantify the value of the deployment of machine translation technology and the return on their investment.

3 Analysis and Evaluation of Approaches

One current approach to evaluating machine translation is through BLEU (Bilingual Evaluation Understudy) (Papineni et al, 2002). BLEU is an algorithmic and quantitative tool that is designed to approximate human judgment from an aggregated corpus. Scores are calculated for individual translated segments—generally sentences—by comparing them with a set of high quality (human) reference translations. These scores are then averaged over the whole corpus to derive an estimate of the translation's overall quality. BLEU is a highly useful measurement index that has been shown to correlate with human judgments. However, in terms of correlating the QOS with stated SLA, it is very difficult to explain what a given BLEU score means for the consumability of the translated content in real-world deployments, and doing so in a consistent way that can be understood by the non-initiated or non-technical stakeholder. For example, a high BLEU score may not necessarily guarantee correlation with a positive end-user satisfaction score. From this standpoint, the question that needs to be answered is whether the localized content satisfied the end-user (i.e., the QOS), in such a way that it prevented a telephone call (i.e., the contractual SLA).

In speech-to-speech machine translation applications like IBM's MASTOR (Zhou et al, 2004), the measurement of translation quality is also based on assigned algorithmic scores which (unlike BLEU) are associated with conversational tasks or units. For example, a yes/no task measures whether an utterance is a logical/sequential response to the preceding question. If the preceding question was "When does John like to go to the ball game?" and the response was "yes", it will be assigned a failing score, whereas if the question was "Does John like to go to the ballgame?" and the response is "yes" it is assigned a passing score. In this approach, scores are assigned based on relevance and accuracy of conversational units derived from semantic constituents in an utterance. Thus, a text may have a low BLEU score and still be deemed effective since a sub-part (the relevant sub-part) of the text contributes to the reader's understanding of parts of the overall semantic import. Although this approach offers matrices that can be more easily understood in terms of end-user satisfaction, it still suffers from the weakness of fragmenting the content being reviewed (rather than a holistic view), which does not imply that the localized content was satisfactorily effective in resolving the users issue.

The foregoing analysis underscores a fundamental and conceptual issue surrounding how end-users evaluate machine translation from a consumability perspective: whether they focus on a sentence-by-sentence, on discourse-by-discourse, or on the overall readability of the entire content? For the unit-based approaches like BLEU and other algorithmic tools that assign scores to segments or discourse fragments, the dilemmas that are faced are (a) how to translate the scores into the holistic and comprehensible terms that match the end-user's perception of success, and (b) how to show an aggregation or a matrix that can be easily understood by the stakeholders in terms of correlating the QOS with the SLA (to quantify the return on investment). For all intents and purposes, the bottom line question seems to be a functional one, i.e., does the translated content enable the end-user to resolve their problem and prevent a telephone call? Algorithmic tools do not yet readily answer this question satisfactorily nor do they offer outcomes or results that can be easily converted to determine the level of end-user satisfaction.

In light of the foregoing, it appears that an acceptable (desirable) approach for estimating the real-world end-user satisfaction with machine translation evaluation is one that is based on a matrix derived from the input of human evaluators in real-time. In this regard, studies by Microsoft (Wendt 2008) and Intel (Burgett and Chang 2008) provide some useful parameters for designing such an approach. In a pre-deployment phase, they describe the process involving human evaluation of quality as follows: 3-5 independent human (QA) evaluators are asked to rank machine translation quality for about 250-300 sentences drawn from a variety of content in the same domain. This involves using human (subjective) judgment to determine translation quality or accuracy based on a scale of 1-4:

4 (Ideal; Excellent) grammatically correct, all information included
3 (Acceptable; Good) not perfect, but definitely comprehensible, and with accurate transfer of all important information
2 (Probably Acceptable; Fair) may be interpretable given context/time, some information transferred accurately
1 (Unacceptable; Poor) absolutely not comprehensible, and/or little or no information transferred accurately

Table 1: Parameters for human evaluation of effectiveness

While this kind of human evaluation is currently done offline, there is no doubt that these parameters will readily correlate to real-time end-user evaluation of machine translation. For example an average score of 4 or 3 implies that the information was useful in deflecting a call, an average score of 2 indicates borderline satisfaction (which may or may not have resulted in call deflection and so inconclusive for measuring the impact of localized content), and finally an average score of 1 is a clear indication that the localized content does not support call deflection. Consequently, we propose that real-time (end user) evaluation of content localized by machine translation consists of three evaluation parameters (a) holistic (is the information complete), (b) comprehensible (is the information understandable), and (c) functional (is the information useful). Furthermore, it generally assumed that offline or lab experiments (e.g., offline testing by human evaluators) may not mirror actual performance in real-world deployments due to uncontrollable external (real-world) conditions. Thus, offline human evaluation is a best guess of performance but not an actual indicator that can be used to validate the business case for machine translation. Ultimately, what is required is that the evaluation be done by the end-users of the localized content using it to resolve IT-related issues.

4 Discussion of our Proposed Approach

In this section, we discuss our systematic attempt to collect real-time human (commercial end-user) evaluation in a real-world s-Support deployment. For the measurement, we adopted our three proposed evaluation parameters for the real-time human evaluation of machine translation localization, and will now describe our implementation. First, we provide some background about the application. Next, we discuss the design of end user evaluation which is implemented as "document-level feedback" (DLF). Finally, we present the results from the DLF and outline some of the drawbacks of our approach as well as some of the workarounds being considered.

4.1 Electronic Support Web Application

The e-Support portal provides online content (solutions) for end-users to resolve IT-related problems on their own without placing a call to the Help Desk. This application handles an average monthly traffic of 2 million visits (total end users) and 5 million page views (page transactions). 50% of the traffic (visits and page views) come from non-English speaking countries which means that machine translation may be used to enhance the user experience for about 2.5 million monthly page views. Nine languages are served by machine translation including Chinese (Simplified), Chinese (Traditional), French, German, Italian, Japanese, Korean, Portuguese (Brazilian), and Spanish. For an application of this size, the significance of measuring the impact of the localized content in meeting the business need (client self assist and consequent call deflection) is a no brainer.

4.2 Document-Level Feedback

In general, a distinction is made between fully automated high quality translation (FAHQT) and fully automated useful translation (FAUT) (Burgett and Chang 2008; TAUS Report 2007, etc.) where the former focuses on quality (grammaticality, rigor, linguistic and stylistic accuracy) and the latter focuses on usefulness or effectiveness in communicating the overall message (semantic import, "gist", comprehensibility, etc.). In our implementation, we assumed the FAUT for the measurement of translation effectiveness based on the concept of goal achievement, i.e., did the localized content help the user resolve the problem (which is similar to the Call Center's first call resolution or call completion). Thus, goal achievement serves as a funnel for call deflection from the web based on the number of end-users who are satisfied with the localized content from machine translation and did not have to call the Help Desk. In order to collect end-user judgments about goal achievement, we created the document-level feedback and asked the end-user to rate the effectiveness of the localized document(s) viewed:

Rate this translation (please take a moment to complete this form to help us better serve you)
The translated information is useful ➢ Strongly agree ➢ Agree ➢ Neutral ➢ Disagree ➢ Strongly disagree
The underlying information is clear and easy to understand ➢ Strongly agree ➢ Agree ➢ Neutral ➢ Disagree ➢ Strongly disagree

Table 2: Verbiage of rate this translation for DLF

It is anticipated that the end-user responses to question #1 would map nicely into the human evaluation parameters such that Strongly agree and Agree imply that the localized content was useful for goal achievement and so deflected a call, while neutral is just that, and both Disagree and Strongly disagree may be taken as evidence that the localized content was not useful and so did not deflect a call. It was also important to ask a follow on question that will help understand if the responses to question #1 are judgments about the source of the translation rather than the translated content. Based on responses to question #2 (the underlying information (source) language is clear and easy to understand), one can envision a scenario where a user indicates "Strongly disagree". This would help shed light on the corresponding response to question #1 since the translation is only as good as the source itself.

4.3 Results

The purpose of the document-level feedback is to enable end-users to answer the question about goal achievement, i.e., whether they gleaned enough of the “gist” from the localized content to be able to resolve their problem without having to contact the human agent over the telephone. At the top of each page view, a user interface function for “Rate this Translation” was provided in the context of the problem resolution (context-sensitive survey). When end-users click on this function button, the two DLF questions are presented. We will illustrate the results from the DLF questions based on data for a single month (we only recently started collecting this data):

Language	# of Feedbacks	Question #1 (average rating)	Question #2 (average rating)
Chinese (S)	300	4.3	4.3
Chinese (T)	100	4.0	4.0
French	500	4.0	4.0
German	700	2.4	2.7
Italian	200	3.0	3.5
Japanese	400	3.0	3.5
Korean	100	5.0	5.0
Portuguese Br	100	4.0	4.0
Spanish	1100	4.0	4.2

Table 3: Document-level feedback for 1 month

The average rating shown in Table 3 is derived from the following taxonomy: all responses tagged as Strongly agree are assigned 5 points, Agree is 4 points, Neutral is 3 points, Disagree is 2 points, and Strongly Disagree is 1 point. When we apply the parameters from our adaptation of the human evaluation matrix for measuring end-user satisfaction, we can extrapolate that 63% of end-users found the localized content useful (2,100 ratings of 4 and above), while 17% were not sure it was useful, and 20% found it not useful.

4.4 Discussion and Implications

Based on a single month’s data, we have observed that 63% of end-users indicate that they are satisfied with the localized content provided through machine translation. However, there is a problem with using this data alone to measure the impact of machine translation because of the small number of responses (less than 1% of monthly visits.) Therefore, the DLF approach is greatly inhibited by the low response rate from the end-users. This is a problem for the DLF approach because end-users are not motivated to respond to the questions, and in fact, in some cases we have observed that it is mostly those who are dissatisfied who take their time to provide feedback. In spite of this inherent weakness of the DLF, there is an important question that Table 3 allows us to investigate. That is, what is the implication when a given percentage of end-users state that they are satisfied with the localized content provided through machine translation? For example, does this imply that there is a corresponding 63% call deflection?

The correct answer will depend on several factors. For instance, Burgett and Chang (2008) based on data spanning 7 months claim that 44% of end-users said that content localized by machine translation helped to answer their question (Spanish), while 54% said no. In our case, we have only collected data for one month and need to have statistically significant numbers in order to reach clear conclusions. Meanwhile, it would appear that the confirmation of the actual impact of machine translation may come from an examination of web data (problem management records) and Call Center statistics (first call completion by categories). The assumption would be that the number of satisfied end-users from the DLF responses should reflect a corresponding reduction in both Call Center data (for the relevant categories) as well as web data on problem management records. We are currently collecting these information and the initial results show that the combination of DLF, problem management records and Call Center data will prove to be the better approach for measuring the QOS to determine the impact of machine translation in meeting the business need (i.e., the SLA).

Finally, we turn to another data point from Table 1, which is the assertion that machine translation can only be as good as the source (underlying information). This is essentially correct. As shown in Table 1: in Chinese (Traditional and Simplified), French, Korean, and Portuguese (Brazilian), the localized content is scored the same value as the source (underlying information), whereas in German, Italian, Japanese and Spanish, the localized content has a lower average rating than the source (underlying information). Remarkably, there is no instance where the source has a lower average than the localized content translated by machine translation.

4.5 Future Work

We plan to collect more DLF data by improving the user interface to make it easier for end-users to respond.

In order to ensure balanced response from both satisfied and dissatisfied end-users, we will work on culturally sensitive incentives (motivation) to help increase the response rate.

We intend to pursue the three-prong approach of reconciling DLF data with web traffic (problem management records) and Call Center data.

Finally, just like currently done in speech deployments, we will explore separating the measurement of core technology (accuracy of engines) from the measurement of QOS (end-user satisfaction) by developing a rejection algorithm which "translates" into end-user satisfaction (automation) rates.

5 Conclusion

In other for machine translation to gain wider acceptance with stakeholders, it is important to have a way to quantitatively measure its impact on the SLA so that they can quantify the return on investment. Currently, there are good algorithmic tools available for measuring quality and some of them like BLEU have been shown to correlate with human judgments. In this paper, we have focused on our work in e-Support localization and the challenges of trying to figure out how to articulate the quantitative measurements to correlate QOS with SLA relevant for calculating the business benefits. We proposed an approach for the real-time human evaluation of localized content that contains three evaluation parameters (a) holistic (is the information complete), (b) comprehensible (is the information understandable), and (c) functional (is the information useful). Thus, we postulate that the subjective evaluation of translation effectiveness of the localized content by commercial end-users is a function of the sum of the holistic, comprehensible, and functional and not limited to just the unit-based measurement of accuracy like BLEU. In this regard, we examined the approach of collecting end-user response for determining QOS through document-level feedback (DLF). Although the DLF approach shows great promise, it is constrained by low response rate from the end-users. Moreover, it is not clear if the percentage of satisfied users matches (equals) corresponding decreases in the volume of calls (call deflection). As one workaround, we advocate that measurements of translation effectiveness should combine (and reconcile) DLF feedback with web data analytics (such as problem management records) along with Call Center data. This may prove to be a better approach that may offer significant results for proving the value of machine translation and its attendant return on investment.

Acknowledgments

Do not number the acknowledgment section.

References

Bowen Zhou, Daniel Dechelotte, and Yuqing Gao. 2004. Two-way speech-to-speech translation on handheld devices. *International Conference of Spoken Language Processing (ICSLP), Korea, October 2004*

Chris. Wendt. 2008. Large-scale deployment of statistical machine translation: Microsoft example. AMTA 2008, Waikiki, Hawaii October 21-25, 2008

Kishore Papineni, Roukos Salim, Ward Todd, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of ACL*, pp. 311-318.

TAUS Report. 2007. *TAUS (Translation Automation User Society) Starter's Guide to Machine Translation: Technologies, Case Studies and Good Practices. Release 1: April 2007*

Will. Burgett and Julie Chang. 2008. The Triple advantage factor of machine translation: Cost, time-to-market and FAUT. AMTA 2008, Waikiki, Hawaii, October 21-25, 2008