User choice as an evaluation metric for web translation services in cross language instant messaging applications

William Ogden,
Sieun An, and Yuki Ishikawa
New Mexico State University
PO Box 3001
Las Cruces, New Mexico 88003
ogden@nmsu.edu

Ron Zacharski

University of Mary Washington
1301 College Ave.
Fredericksburg, VA, USA 22401
raz@umw.edu

Abstract

A method for evaluating MT performance embedded in Cross-Language Instant Messaging (CLIM) systems is presented. A web interface that provided concurrent real-time translation for instant messaging from multiple MT services was developed and used by paid participants to collaborate on a photo identification task. The method showed a task performance benefit due to the availability of multiple translation alternatives. The method also provides a new evaluation metric for MT systems based on user's task motivated choices. This method was used to compare two English-Japanese online translation systems, one from Google, and one from Excite/Japan.

1 Introduction*

Cross-language instant messaging (CLIM) systems are intended to be used by groups of people who do not speak the same language but need to work together to accomplish common tasks. These systems can help mixed-language groups communicate by automatically translating text messages into the preferred language of each team member. Members of the group would be able to read and enter messages in their own language and still communicate with all members. The success of these systems will depend on the usefulness of the embedded machine translation (MT) software within the context of instant messaging.

We have been developing a method that would be sensitive to the evaluation of embedded machine translation in an instant messaging application. The goal is to study the performance of machine translation technology in the context of real tasks. In contrast, automatic evaluation methods, such as the popular BLEU method (Papineni et al., 2002), are used to compare systems against each other and are specifically designed to remove the potential bias a task context may impose on the evaluation. We believe one of the most important criteria of translation quality is its usefulness which can only be judged in the context of the task in which it is used. Therefore it is important to test the performance of a system by measuring the performance of the people who actually need to understand the translations to accomplish their goals.

In this study, we demonstrate a method of evaluating machine translation software for use in the context of online instant messaging. We have constructed an instant messaging application using a configurable set of available online web translation services. In particular, using various internet sites that provide instant translation of short texts we can provide CLIM users with concurrent translations from multiple sites. Most importantly, by asking test participants to use the

* Prepared through participation in the Advanced Decision Architectures Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under Cooperative Agreement DAAD19-01-2-0009. The views and conclusions contained in this document are those of the author, and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, or the U.S. Government.

system to work together with a cross-language partner to solve a collaborative task, we can evaluate the usefulness of any particular MT system by measuring task performance and by measuring which translation systems the participants are motivated to choose to help complete their cross-language communications.

2 Previous work

In our previous work (Ogden, 2009) we describe a series of studies that utilized task-based methods for evaluating embedded MT technology in CLIM systems. These studies involved user testing of a translingual instant messaging system interface (TrIM), developed by The MITRE Corporation to help multinational coalition partners to communicate using their own languages (Miller et. al., 2001). A series of studies were conducted in which pairs of participants using different languages are asked to work together using a CLIM system to accomplish a task. The tasks given to the participants evolved from study to study but all required that participants share information with a cross-language partner in order to complete tasks in a military logistics domain. Tasks were designed to invoke realistic conversations between participants in that domain.

Machine translation software typically suffers from a set of known problems. For example, syntactic and lexical ambiguity in the source and target languages can lead to poor translations due to word selection errors. So far, one of the most pervasive of findings in our investigations is the observation that these poorly translated messages slow participants down, but do not prevent them from communicating and sharing task relevant information. Participants engage in repair dialogs and other strategies that largely depend on knowledge of the task domain.

We see many examples of users trying to clarify misunderstood words or parts of the message by highlighting or echoing back just those parts. Unfortunately this strategy usually leads to more misunderstandings because the echoed parts are translated back to the originator differently. For example, suppose an English speaker types in the word *gun* as part of a sentence. This might get translated into Korean as 술잔, 글라스 (glass). The Korean speaker, confused by how this word is relevant to the conversation, echoes the word 술잔, 글라스 followed by a question mark back to the English speaker in an attempt to get clarification. The MT system translates this as glass which is displayed to the English speaker. From the English speaker's perspective, he types in something like *Do you have a crate of guns?* and he gets back as a response *glass*?

This is a problem with the asymmetrical design of most MT systems. The system going from one language is different than the one going back to the original. Furthermore, MT systems usually do not save the mapping from the original to the translation. If a mechanism could be implemented that would allow feedback to be given to the originator of the message about parts that were incorrectly translated we predict a great benefit.

In one study, we experimented with an interface feature designed to aid dialog repair by providing a set of "meta-buttons" which mapped the seven most common types of repair messages onto a set of function buttons to deliver fixed, human translated versions of those messages. The addition of the meta-buttons had an effect on a participant's method of solving the communication problems and helped participants communicate important information faster and perhaps better.

The more important findings of this work suggest that MT software does not have to be perfect to be very useful. This is because there are intelligent language users on both sides of the MT software in CLIM applications. However, while we have observed many successful repair dialogs, there is nothing in the underlying technology of the system or of the user interface which supports this repair activity. Although our initial evaluations of interface enhancements, such as 'meta buttons,' show promise, more effective collaboration tools are needed to minimize potential communication breakdowns among distributed multi-lingual team members.

3 Multiple translations and user choice

Our previous work suggest CLIM systems can be used to accomplish meaningful work, but that the translation quality of automated systems still require communication partners to repair and clarify misunderstandings due to translation problems. Another, more available feature of some MT systems is the ability to provide alternate

translations. Currently TrIM shows only one possible translation. We predict a benefit if users could request alternate translations directly from the MT software, rather than solely from their partner. Alternatively, one could use multiple translation systems to provide users with alternate translations. We predict that the need for clarification will decrease and task performance measures will improve if the participants are provided multiple translations of each message.

In the present study we also wanted to make the task a bit more accessible to a more general, non-military population. In addition, we wanted it to be less influenced by specific domain vocabulary problems in the target and source languages. So we abandoned the military logistic task primarily used in our previous work, and developed a photo identification task which could be presented without the built-in language bias inherent in language presentations. Participants were presented with a set of real-world photographs and they needed to identify which photo was being viewed by their chat partner by texting natural language questions.

4 Method

4.1 Participants

Sixteen native Japanese speakers were each paired with a native English speaker to form 16 two-person teams. In addition, 12 additional English speakers were paired to form 6 teams to serve as a Control group. Nearly all the participants were university students and were proficient typists in their native languages. Each participant was paid $15 per hour, usually receiving between $20 and $25 for their participation.

4.2 Procedure

Each participant sat in a small room with a computer and was asked to communicate using an instant messaging system with a partner who sat in an adjacent, but physically separate room. The two-person team could not see or hear each other and could only communicate via text messages. The participants were asked to take turns identifying which of nine photos presented on their screen was also being presented on their partner's screen. Their partner would have one of the nine photos (the target) showing on their screen. Figure 1 shows a sample of the questioner's screen showing nine photos.

The participant used the text messaging system (shown in Figure 2) to ask questions about the target photo. When the questioner had enough information, they would use the mouse to select one of the photos and then select the "Submit" button. The participants would then both see a screen showing the target photo and the selected photo (but only if different from the target). The time required to solve the task was also displayed. The next trial would start with a new task when both participants selected a "Continue" button. The role of questioner would shift to the other participant on each trial.

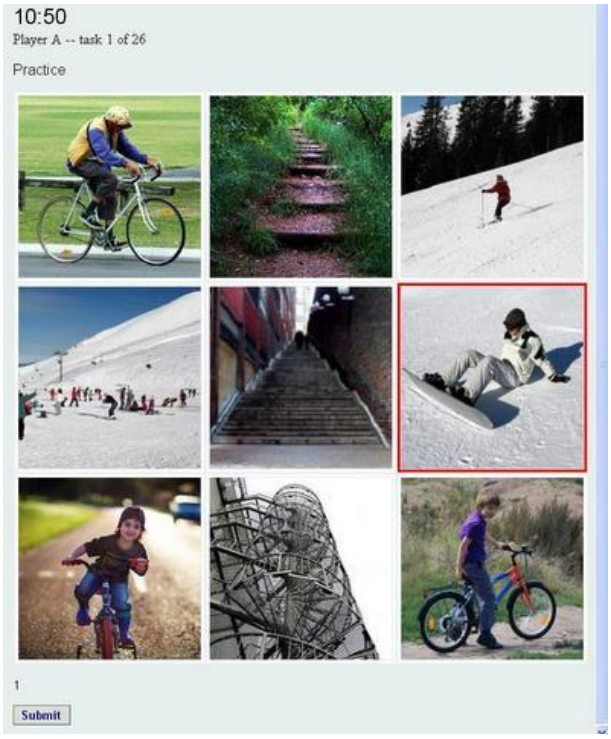

Figure 1. The photo identification task.

Three sets of photo triads were selected for each trial. Photo triads were three photos of a similar theme (e.g. person on bicycle, skiers, stairs, etc) and were selected from stock.xchng, a free stock photo web site (http://www.sxc.hu/). All photos were displayed in color.

Participants used the CLIM client window shown in Figure 2 to ask questions about the target

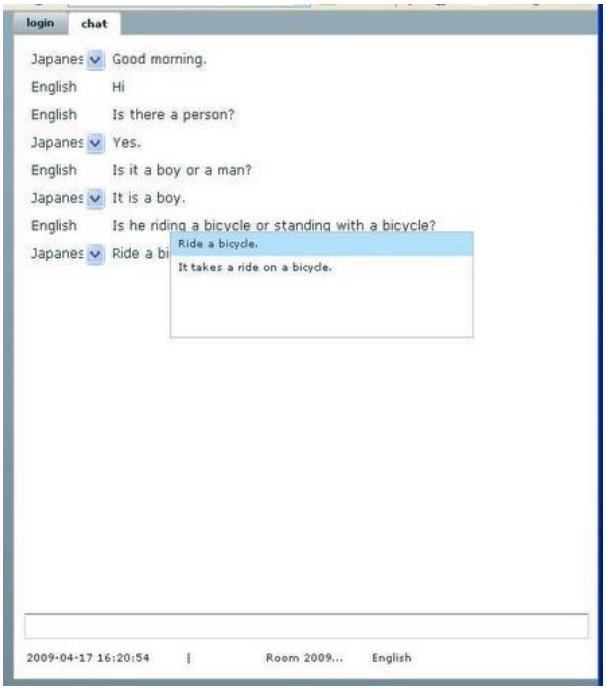

Figure 2. The CLIM client used in the experiment.

photo. They could ask any question they liked. To make the task a little more challenging, however, the partner was instructed to only answer questions and not offer new information about the target photo.

Japanese participants entered text in Japanese and saw automatic translated English to Japanese responses. English participants entered text in English and saw automatic translated Japanese to English responses (except when an English participant was paired with another English participant in the Control condition in which case no translation occurred. The instant messaging interface shown was embedding in a web browser implemented as a Flash client.

Translations were obtained from one of three online translation sites, Google, BizLingo through Excite (Japan), and Amikai. We accessed each translation engine using a separate PHP script (a translation wrapper) specialized for each engine. Each script ran as a separate process decoupling requests for translation from the main flow of execution. Once a translation was received, it was placed in a database. A separate script served these translations to the Flash client.

The Google translation wrapper used the Google AJAX Language API. The Google server returned a JSON encoded result, which the wrapper processed. The BizLingo translation wrapper used the standard Excite (Japan) web interface (http://www.excite.co.jp/world/english). The Amikai translation wrapper also uses a standard web interface designed to be used for demonstrations (http://www.amikai.com/demo.jsp). Both of these translation wrappers generated http encoded translation requests and extracted translations from the returned web page using regular expressions.

The CLIM client could be configured to show single or multiple translations. Translations were shown as soon as they were available in the database. For purposes of our experiment, we chose not to reveal the source of the translation to the user but it was recorded in the database. When multiple translations were available, an arrow button appeared next to the translation. When a participant clicked the arrow button, a pop-up window would appear showing all available translations for that message. The participant could click on one of the translations in the pop-up and it would replace the currently shown translation. Figure 2 shows that the participant clicked the arrow button next to the last translation on the screen. The pop-up shows two alternative translations. The CLIM client recorded both the "view" events, (i.e. when the participant clicked the arrow button to view the alternative translations) and the "replace" events (i.e. when the participant clicked an alternative translation in the pop-up).

Eight of the Japanese-English teams saw multiple translations and the other eight teams saw only one translation. Unfortunately, half way through the study, the Amikai translation service became unavailable. Therefore, four of the teams seeing multiple translations only saw two, not three alternatives and four of the teams seeing single translation alternated between two instead of three translation services. For all teams, the initially presented translation came from one translation service for each trial. The source of the initially presented translation then rotated among the two or three available services for each subsequent trial. Thus equal numbers of trials occurred with each translation source being presented first.

Participants were given two practice trials followed by 24 trials. Tasks were presented in a different random order to each team.

5 Results and Discussion

5.1 Task measures

Table 1 shows the average task performance measures for groups of teams in the study. Time to solve the task was only measure that showed a significant difference between groups. The Control group teams, whose members communicated with each other in English and required no translation, were faster than the groups that required translation, $t(17) = 3.7$, $p < .001$. This matches the observations we have made in our previous work that translation slows but does not prevent task completion. While not significant, however, it does appear that there is a speed advantage when participants have more alternate translations to view. Teams that had three alternatives available were on average faster than those teams that had two which were faster than those teams who had only a single translation to view.

Available Translations	N	Time (sec)	Per cent correct	Message count
Single	8	177	83	8.80
Multiple 2	4	148	83	7.78
Multiple 3	4	130	86	7.74
English Control	6	109	94	7.74

Table 1: Average task completion performance.

What is even more interesting in the present data is the lack of differences. For example, the teams that could view multiple translations required on average the same number of messages to complete tasks as did the teams using English only. Thus it would appear that participants could make up for a bad translation by viewing an alternative one. The teams without an alternative translation to view took on average one additional message to solve each task. Again, it would appear that without the alternative translations, participants had to use more messages to clarify and repair miscommunications. Indeed, a count of the number of meta-messages (messages that were used to repair the conversation) showed the Single condition teams used on average 11 meta-messages to complete all 24 tasks whereas the Multiple condition teams only used an average of two meta-messages which is a significant difference $t(14) = 3.25$, $p < .005$.

While it may seem having more than one translation provides a benefit, it could be that one of the translation engines is much better than the other and the Multiple condition just makes it more likely that the good system's translations are seen. We discuss measurable differences between the translation systems in the next section.

5.2 Translation quality measures

This study provides an interesting task related measure of translation quality. When the quality of a translation is unsatisfactory for the purposes of completing the task, participants should be much more likely to view an alternative translation than when the quality of the translation is good. Therefore, counting the number of times a participant views an alternative translation can be considered a measure of translation quality which is directly related to translation usefulness in the context of the task and instant messaging.

In the following analysis we will compute the number of times a participant viewed an alternative translation for each task as a function of which translation service provided the initial translation. Remember, the translation service providing the initial translation alternated from trial to trial. We will only consider two translation services, Google and BizLingo because the Amikai service stopped working during the experiment. We can directly compare Google and BizLingo because they each had equal numbers of tasks in which they were used as the provider of the initial translation.

Initial Translation	View count	Time (sec)	Per cent correct	Message count
BizLingo	1.83	132	79	7.41
Google	3.87	150	89	7.98

Table 2: Average task performance as a function of initial translation provider.

As can be clearly seen in Table 2, there was a big difference in the number of times participants decided to view alternative translations depending on which translation was showing. Google translations prompted twice as many view events (3.87) than were prompted by BizLingo translations (1.83) which is a highly significant difference, $t(7) = 5.71$, $p < .001$. Participants were

less likely to view alternative translations when they were viewing translations provided by BizLingo and were more likely when the translations were provided by Google. From this evidence it would be safe to say that BizLingo provides more useful translations for this task.

The other measures in Table 2 show that Google trials took a bit longer, had fewer errors, and resulted in more messages, but these differences were not significant.

Finally, we wanted to compare the translation quality measure obtained by task performance with a more traditional measure of translation quality, human judgment. Therefore we asked two bilingual Japanese-English judges to rate the quality of a random selection of two translations from each trail in the experiment for Google and BizLingo generated translations (1280 messages). Each judge independently rated each message translation pair for Adequacy on a five point scale for Information Content: 1) None, 2) Little, 3) Much, 4) Most 5) All. The same judges, who were native Japanese, rated the same English to Japanese translations for Fluency on a five point scale: 1) Nonsense, 2) Disfluent, 3) Non-native, 4) Good, 5) Flawless. A second set of native English judges rated the Japanese to English translations using the same 5-point scale for Fluency. The linear weighted kappa score for inter-judge agreement was 0.552, indicating a "moderate" level of agreement. The average rating results are displayed in Table 3.

.

	Adequacy	
	English	Japanese
BizLingo	4.30	4.40
Google	2.89	3.11

	Fluency	
	English	Japanese
BizLingo	3.98	3.70
Google	2.64	3.47

Table 3: Average ratings for two judges as a function of source language and translation service

The human translation quality ratings match those obtained with our task measure. Both indicate that the BizLingo translations were higher quality than the Google translations.

We still don't know if providing multiple translations has a benefit over just ensuring that at least one good translation service is provided. While none of the differences in Table 4 are statistically significant, they are suggestive on this point. In the Single condition, participants were slightly faster and used less messages when using BizLingo translations compared to when using Google translations. But when comparing task measures between Single and Multiple conditions, Multiple is better for each translation service so it seems likely that there may be some advantage for providing translations from multiple sources even for the best translation service. BizLingo performs best when presented with multiple translations perhaps because while it might be wrong less often, on those occasions when it is wrong another translation service may be right.

.

Available Translations	Translation Service	Time (sec)	Message count
Multiple	BizLingo	135	7.43
Multiple	Google	150	8.00
Single	BizLingo	172	8.46
Single	Google	194	9.49

Table 4: Average task performance by translation service and condition

6 Summary and Conclusion

User choice appears to be valid metric for evaluating the quality of MT systems in cross language instant messaging applications. When viewing a translation generated by a system that is highly rated by human judges, participants choose to view alternative translations less often than when viewing a translation generated by a lower rated system. Thus, at least preliminarily, we can claim that the task performance measure reflects human judgments of translation quality. This is an important finding because some task performance measures are easier and less costly to obtain than expert human judgments.

However, does evaluating the quality of a translation system really matter? If even the best translation system can benefit by making alternative translations available, maybe it's enough to ensure users have a choice because even a bad translation systems can be right occasionally

and in the world of CLIM, that might make a difference.

References

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

John R. Joyce. (2003). Coalition Interoperability Tested at Dahlgren During JWID 2003. CHIPS - The Department of the Navy Information Technology Magazine. Fall 2003, Retrieved from http://www.chips.navy.mil/archives/03_fall/PDF/JWID_2003.pdf

Keith J. Miller, Florence Reeder, Lynette Hirschman, David D. Palmer. (2001). Multilingual Processing for Operational Users. MITRE Technical Report, Retrieved from http://www.mitre.org/work/tech_papers/tech_papers_01/miller_multilingual/index.html

Ogden, W. C. and Bernick P. (1997). "Using natural language interfaces." In Helender, M., Landauer T., & Prabhu P. (Eds.), Handbook of Human Computer Interaction, 2nd Edition, Amsterdam: North Holland.American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Ogden, W. C. (2009). "A Task-based Evaluation Method for Embedded Machine Translation in Instant Messaging Systems." In Advanced Decision Architectures for the Warfighter: Foundation and Technology. Alion Science and Technology Corp., McLean, Va.