ISSN 2278-5612

Comment Evaluation and Revision in a Bilingual Electronic Meeting

Milam Aiken, Jamison Posey, Brian Reithel

School of Business Administration, University of Mississippi, University, MS 38677

maiken@bus.olemiss.edu

jposey@bus.olemiss.edu

breithel@bus.olemiss.edu

ABSTRACT

Translation accuracy continues to be a major problem in multilingual electronic meetings and a significant reason for such systems' lack of acceptance and use. One way of increasing accuracy is detecting potential errors before a comment is added to the discussion. By comparing the original message text with a round-trip translation (RTT) and correcting any wording mistakes, misunderstandings might be prevented in the overall conversation. In this study, one group used an electronic meeting system with automatic translation that detected differences between a participant's typed comment and a RTT. If there was a significant change, the group member was then given an opportunity to modify the text before submitting it to the transcript. Another group, serving as the control group, used an equivalent electronic meeting system without error detection. Results show that there was only a slight, non-significant increase in comprehension when comments were revised and translated to German with an 80% similarity threshold, but there would have been a significant increase in comprehension if a 50% threshold had been chosen.

Indexing terms/Keywords

Electronic meetings, Group Support Systems, Machine translation, Multilingual groups

Academic Discipline And Sub-Disciplines

Computer Science (machine translation)

TYPE (METHOD/APPROACH)

Lab Experiment

INTRODUCTION

Multilingual meetings with automatic machine translation have been conducted for over 30 years, and support has steadily improved. The need for such systems continues to increase, with the globalization of software and product development teams [9]. New languages are continually added, with one modern system able to support 80 languages [14]. However, translation accuracies often are not acceptable. One reason for the inaccuracies is the translation capability of the software, but another is that group members sometimes make mistakes in their comments. Typing errors, acronyms, slang, poor grammar, and idioms can affect a translation. Even if no mistake is made, a translation can use different words than intended or have an altered word order.

However, the message author is largely unaware of the content of a comment's translation; therefore, he or she can make incorrect assumptions about how well the idea was transmitted [12]. That is, a participant might believe others in the group have understood what he or she said or wrote, but they might have perceived only part of the information, if any. In one study [18], the researchers found that 1 out of 20 translated comments in an electronic meeting included misconceptions, while the comment originators did not know they were misunderstood. During an oral discussion, group members can indicate misunderstanding by facial expressions with looks of confusion. However, in an electronic meeting with typed comments, these emotions are difficult to convey. Group members could add a comment asking for clarification, but this takes extra time and adds to the length of the transcript. Further, many participants might not be willing to make the extra effort due to laziness or shyness.

Preventing misunderstandings in the first place might significantly reduce group confusion over the meanings of potentiallyconfusing comments. One way to do this is with a round-trip translation (RTT). That is, the comment text is translated to another language in a forward translation (FT), and then the results of FT are translated back into the original language in a backward translation (BT). If there are differences between the backward translation and the original comment, there might also be differences in meaning in the forward translation.

The purpose of this paper is to investigate how well this RTT error-detection technique performs in a simulated bilingual meeting with automatic translation. First, we discuss prior studies of RTT and then we describe the software. The paper then describes an experiment testing the system and presents a discussion of the experiment's results.

LITERATURE REVIEW

To our knowledge, Amikai's *AmiChat* was the first electronic meeting system that allowed users to indicate whether or not a comment was understood [8]. With this software, group members clicked a button when they were unclear about a message, and this sent a warning to the originator with a request to rephrase the text.

Perhaps the first system to use RTT for comment error detection in an electronic meeting was *AnnoChat* [19]. Using this software, the comment author compared the back translation with the text originally written, and if the BT was poorly phrased or used incorrect words, the author could revise the text and submit it again, repeating until a satisfactory result was obtained.

However, studies using this system have had mixed results. In one study [11], Japanese users were able to improve English translations with RTT, but they were not able to improve translations to Chinese and Korean. That is, changing the original text could increase the accuracy of a translation into one language, but it might decrease the translation accuracy to others.

In a study using a different multilingual electronic meeting system [4], there was no significant difference in FT accuracies between a group using RTT error detection and another group that was not. However, those using RTT error detection were more likely to believe that speakers of other languages would be able to understand their translated comments.

In another study using this system [2], group members significantly underestimated how well the forward translations were understood when using RTT error detection. Unfortunately, only 4.1% of the comments generated by the group were revised for possible better translation.

Several researchers have argued that there is no value to RTT for error detection, because, for example, a poor RTT could be due to a bad FT, a bad BT, or both [17]. Alternatively, a good RTT might have a bad FT, but the BT resembles the original text. However, others have argued that if the final and original texts are identical or very similar after RTT, the technique provides strong evidence that the FT is accurate (e.g. [6]). Two studies [15] [16] found positive correlations between forward and back translation accuracies when using Japanese and English. Furthermore, the latter study found that back translation accuracies underestimated the associated forward translations.

Two studies using English and Korean [3] [4] found significant correlations between FT and BT accuracies. That is, if the RTT was bad, the FT was also bad. Another study using English, German, and Spanish. Finally, Aiken & Hazarika [2] also found moderately positive, yet significant correlations between forward and backward translation accuracies.

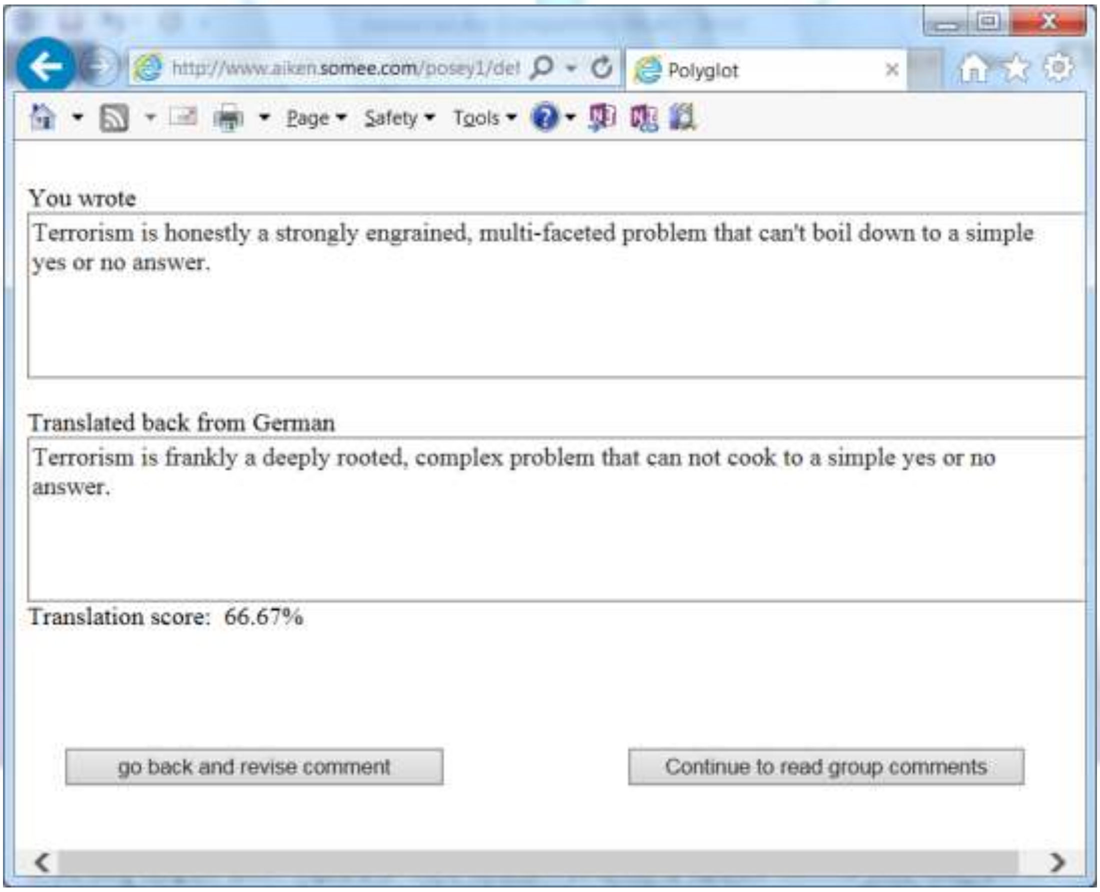
MULTILINGUAL MEETING SOFTWARE

Prior studies of multilingual electronic meetings with RTT always presented the comment originators with back translations for evaluation. However, this process could slow down a meeting and present an annoyance to participants. Alternatively, a system could present the BT text only if the software detected a possible problem, e.g. differences in word orders or the number of words used.

We designed a multilingual electronic meeting system with RTT error detection that presents the back translation to a comment author only in situations in which the percentage similarity between the original text and the back translation falls to 80% or below, a threshold that was arbitrarily selected as a starting point for this study because this is the first time this particular automated notification technique has been attempted in a multilingual electronic meeting system context. The software compares how similar the words are in two samples or text, with no regard for word order, synonyms, or spelling errors. The Bilingual Evaluation Understudy (BLEU) technique [10] and others are much more sophisticated and possibly a better choice, but an analysis using data from a prior study [2] showed that the correlation between percentage word similarity in the back translation and actual human understanding in the forward translation was strong and significant (English-German: R=0.611, p<.001; English-Spanish: R=0.572, p=0.001). Therefore, we believe a simple comparison of word similarity between the original text and back translation could be useful in predicting forward translation accuracy.

Using this software, a meeting participant types a comment in a textbox, and then clicks on a button to submit it to the group. The system translates the comment into a target foreign language (or multiple foreign languages), for example, English to German, resulting in a "forward translation" (FT). Next, the software translates the German text back to English again, giving a back translation (BT). The system compares word similarity between the original comment and the back translation giving a score of 0% - 100%. If the score is at 80% or below, the comment author sees a new screen (Figure 1) that shows the original text and the back translation, and the user is able to go back and revise the original comment before it is finally submitted into the discussion.

Figure 1. Round-trip translation comment error detection


http://www.aikensomee.com/posey1/det
Polyglot
Page ▾
Safety ▾
Tools ▾

You wrote

Terrorism is honestly a strongly engrained, multi-faceted problem that can't boil down to a simple yes or no answer.

Translated back from German

Terrorism is frankly a deeply rooted, complex problem that can not cook to a simple yes or no answer.

Translation score: 66.67%

go back and revise comment
Continue to read group comments

The assumption is that low similarity scores might be correlated with poor forward translations, and the comment author should probably correct any text that does not meet the selected threshold. However, the word similarity technique is not perfect, as illustrated in the examples below. In comment 1, there is a perfect match, and chances are good that the forward translation would probably be understood. Comment 2 was not perfect, but above the 80% limit and the meanings are identical. In comments 3 and 4, deliberate spelling errors were made, and comprehension is proportional to the similarity scores. Comment 5 shows intentional junk text, and there was a perfect match. Comment 6 is very poorly worded, and although there is an 80% match, both sets of text are difficult to understand. Comments 7 and 8 are more serious, and we believe their comprehension is proportional to their similarity scores.

1. Original: The rain in Spain falls mainly on the plain.

   Back translation: The rain in Spain falls mainly on the plain. [100% match]

2. Original: I think we need more parking spaces.

   Back translation: I think we need more parking. [85.71% match]

3. Original: Wht is you name?

   Back translation: Wht is her name? [75% match]

4. Original: se shells sea shells by th see shore

   Back translation: se of th lake shore mussels mussels [30% match]

5. Original: xxx aaaa zzzzz

   Back translation: xxx aaaa zzzzz [100% match]

6. Original: brown kitten show what now?

   Back translation: brown catkins show what now? [80% match]

7. Original: Muslims are feeling the heat of a fierce backlash following last a terror attack against French satirical magazine Charlie Hebdo.

   Back translation: Muslims feel the warmth of a violent backlash after a recent terrorist attack on French satirical magazine Charlie Hebdo. [63.64% match]

8. Original: Terrorists' hatred is responsible.

   Back translation: Terrorists hate responsible. [25% match]

As shown in Figure 1, the software shows the reverse translation only if the similarity score falls below a designated threshold. If the back translation is different, but still acceptable, the user merely presses the "continue to read group comments" button and the comment is added to the transcript. Thus, we believe the system will be less intrusive and will not unduly bother the comment author with suggested corrections, thereby improving the flow of the electronic meeting experience.

EXPERIMENTAL STUDY

Purpose

The purpose of this study is to evaluate how well an electronic meeting system can detect possible errors in a participant's typed comment. That is, will participants retype comments if the RTT measure falls to the stated threshold of 80%? Will group members find the system too difficult to use?

Subjects and Task Description

We recruited a sample of 40 undergraduate business students from a large public university in the southern United States. In the first treatment, 22 students used the software described above for about 10 minutes to exchange comments in English about how to combat terrorism. In the second treatment, 18 students used the electronic meeting software with automatic translation to discuss the same topic for the same amount of time, but the system did not evaluate their comments for accuracy, unlike the participants in the first treatment. Afterwards, students completed a survey asking questions about how usable and useful the system was. The system simulated a bilingual meeting in that the students were told a German speaker would be evaluating their translated comments afterward.

Results

Table 1 shows a summary of the experimental results. Students in both groups believed their respective electronic meeting systems were easy to use and useful, and there was no significant difference between the treatments ($F = 0.55$, $p = 0.46$, $F = 0.48$, $p = 0.494$, respectively). However, group 1 generated significantly more comments in the same amount of time ($F = 18.86$, $p < 0.01$). A correlation analysis showed that, as might be expected, those who found the system easy to use indicated a greater likelihood of using the system ($R = 0.609$, $p < 0.001$), but neither ease of use nor usefulness were predictors of the number of comments ($R= -0.28$, $p=0.073$) and ($R = -0.074$, $p=0.649$), respectively.

Table 1. Overall summary of variables

Variable	Min	Max	Mean	Std. Dev
Overall				
Ease of use	1	7	4.90**	1.59
Usefulness	1	7	4.40	1.71
Comments per person	0	10	1.98	2.33
Group 1				
Ease of use	2	7	4.73*	1.28
Usefulness	1	6	4.23	1.45
Comments	0	10	3.23	2.37
Group 2				
Ease of use	3	7	5.11*	1.97
Usefulness	3	7	4.61	2.06
Comments per person	0	5	0.53	1.31

* Significantly different from the neutral measure of 4 at α = 0.05

** Significantly different from the neutral measure of 4 at α = 0.01

Comment analysis

Group 1 generated 82 comments with 65 sentences and 1,015 words. The text had 12.6 words per sentence and a Flesch Reading Ease score of 68.1 (0 = difficult, 100=easy) and a Flesch-Kincaid Grade Level of 6.8 [7]. Group 2, with no error checking, generated 36 comments with 45 sentences and 568 words. The text was also relatively easy with 11.9 words per sentence and a Flesch Reading Ease score of 71.0 and a Flesch-Kincaid Grade Level of 6.1

Of the 82 comments typed by Group 1 with RTT error detection, participants opted to revise their comments 22 times (26.8%). The following discussion provides an illustration of how and when group member's revised their comments:

> One participant wrote, "We collect intelligence on potential terrorism attacks." The back translation was "We gather information about potential terrorist attacks." with a 42.86% similarity score; the user chose to revise the comment. About 49 seconds later, the group member submitted the revised comment "We gather intelligence on potential terrorism attacks." That gave a back translation of "We gather information about potential terrorist attacks." with a 57.14% similarity score. The group member decided this comment was acceptable and submitted it as final.
>
> On the other hand, another group member wrote, "You can fight terrorism by first learning why they're doing terrorist acts first." This gave a back translation of "You can terrorism by first learn why they are doing, acts of terrorism first fight." with a similarity score of 68.42%. About 45 seconds later, the participant submitted a slightly modified "You can learn how to fight terrorism by first learning why they want to commit the acts of terror." The back translation had a 100% similarity score and the comment was added to the transcript.
>
> As a third example, a participant wrote, "We should not involve ourselves with other countries unless they're allies." This rendered a back translation of "We do not have to deal with other countries if they are allies." with a 50.00% similarity score. This translation is different in meaning, and the group member went back and submitted an altered comment two minutes and two seconds later: "Americans need to be concerned with Americans, unless the country is an ally." The back translation was equivalent with a 100.00% similarity score and the group member's comment was added to the transcript.

Comprehension analysis

A German speaker evaluated all of the forward translations provided by the electronic meeting systems and he was able to understand 87.26% of the comments in group 1 (std dev =19.14%) and 93.65% in group 2 (std dev = 13.06%). However, there was no significant difference in comprehension of the forward translations between the two groups' transcripts (T = 1.81, p = 0.08). This result is consistent with a prior study of RTT [4]. Also, this level of accuracy is consistent with prior studies of the electronic meeting system, e.g. 86.57% comprehension in meetings with English, German, and Spanish, and 94.97% in meetings with English, German, Italian, Spanish, and Swedish [13]. Further, several studies have suggested that this level of understanding is probably sufficient for informal discussions (e.g., [5]).

An analysis of the system with error checking revealed that the back translation scores significantly underestimated the actual forward translation comprehension (74.85% vs. 87.26%, respectively, T=3.71, p < 0.001). Thus, it seems the system erred on the safe side. This pattern is consistent with a prior study [2].

Also, some comments were not understood by the German reviewer because the original English comment was confusing, e.g. one group 2 member wrote “THEN WHO WAS PHONE?” Several acronyms and slang were not translated but simply repeated, and some idioms, e.g., “It is a slippery slope” were translated literally, potentially causing confusion. None of these errors were caught by RTT. In addition, longer comments often were more understandable because of the additional words to add context.

There was only a slight and insignificant improvement (T = -0.645, p = 0.259) in German comprehension of comments which the group member decided to revise (after receiving a low back translation score), 83.12% for the original (std dev = 18.34%) vs. 87.69% for the revised (std dev = 16.54%). However, there was a significant difference in German comprehension if the threshold was changed from 80% to 50% (T = -2.87, p = 0.021). Comprehension of forward translations to German was 85.0% (std dev = 8.66%) for those comments with BT similarity scores falling below 50%, but comprehension rose to 94.2% (std dev = 7.10%) for the subsequent, revised comments. Also, in some cases, the content of the original comment was completely changed, so the comparison might not be relevant. For example, one participant wrote, “Don't give power to them.” The group member subsequently revised the comment after receiving a back translation score of 80% to “Stop the killing of american citizens.”

CONCLUSION

Summary

Group members used an electronic meeting system with round-trip translation in an attempt to increase the comprehension of forward translations to German, forming a simulated bilingual meeting. Results showed that there was improvement (83.12% to 87.69% comprehension), but the improvement was not statistically significant. However, if a different similarity threshold of 50% had been selected, there would have been significant improvement. Members of both treatment groups – with and without error detection – thought the systems were usable and useful. Ultimately, comprehension was probably sufficient for a bilingual discussion.

Limitations

The first limitation is that the meetings only simulated bilingual discussions as participants entered comments in English and the German speaker evaluated the transcript only after the meeting had ended. Nevertheless, the objective was to detect possible comment errors in an electronic discussion.

Second, only two languages were used in the simulated meeting, English and German. Other language combinations, e.g. Japanese and Hindi might have worse or better translation accuracy. That is, translating from Hindi to Japanese and back to Hindi again will likely give a different RTT evaluation score.

Third, only one German speaker was used to evaluate the forward translations. Others might have produced different opinions about how well passages convey their intended meaning.

Fourth, a simple, similarity score algorithm with an arbitrary 80% error detection threshold was selected. Using a more sophisticated technique such as BLEU with a lower error detection critical threshold value might yield different results.

Future Research

Studies of multilingual groups ideally should involve participants typing in different languages. However, it is difficult to obtain fluent speakers of languages other than English in many American universities. By hosting electronic meetings on the Web, however, group members from many different nations could contribute to the conversation. Furthermore, additional language combinations and/or translation comparison techniques might produce improved results.

REFERENCES

[1] Aiken, M. (2009). Transterpreting multilingual electronic meetings. International Journal of Management and Information Systems. 13(1), 35-46.

[2] Aiken, M. and Hazarika, B. (2012). Evaluation and revision of translation-mediated communication in a multilingual electronic meeting. International Journal of Business and Systems Research. 6(4), 379-394.

[3] Aiken, M. and Park, M. (2010). The efficacy of round-trip translation for MT evaluation. Translation Journal. 14(1).

[4] Aiken, M. and Park, M. (2009). Enhancing bilingual electronic group meeting comprehension with round-trip translations. International Journal of Systems and Change Management. 4(2), 103-116.

[5] Aiken, M., Wang, J., Gu, L., and Paolillo, J. (2011). An exploratory study of how technology supports communication in multilingual groups. International Journal of e-Collaboration (IJeC), 7(1), 17-29.

[6] Chan, S. (2006). A Dictionary of Translation Technology. Hong Kong: Chinese University Press.

[7] Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221-233.

[8] Flournoy, R. and Callison-Burch, C. (2000). Reconciling user expectations and translation technology to create a useful real-world application. Proceedings of the 22nd International Conference on Translating and the Computer. 16–17 November, London, UK.

[9] Funakoshi, K., Yamamoto, A., Nomura, S., & Ishida, T. (2003). Lessons learned from multilingual collaboration in global virtual teams. In Tenth International Conference on Human Computer Interaction, 22-27 June, Crete, Greece.

[10] Noorbehbahani, F. and Kardan, A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. Computers & Education. 56(2), 337-345.

[11] Ogura, K., Hayashi, Y., Nomura, S. and Ishida, T. (2004). User adaptation in MT-mediated communication. The First International Joint Conference on Natural Language Processing (IJCNLP-04). 596–601.

[12] O'Hagan, M. and Ashworth, D. (2002). Translation-mediated communication in a digital world – Facing the challenges of globalization and localization. Multilingual Matters: Toronto, Ontario, Canada.

[13] Pepper, W., Aiken, M., and Garner, B. (2011). Usefulness and usability of a multilingual electronic meeting system. Global Journal of Computer Science and Technology (GJCST), 11(10), 35-40.

[14] Posey, J. and Aiken, M. (2014). Large-scale, distributed, multilingual, electronic meetings: A pilot study of usability and comprehension. International Journal of Computers & Technology, 14(3), 5578-5585.

[15] Shigenobu, T. (2007). Evaluation and usability of back translation for intercultural communication. Usability and Internationalization: Global and Local User Interfaces. Berlin: Springer, pp.259–265.

[16] Shigenobu, T., Yoshino, T., Nadamoto, A. and Ishida, T. (2007). Accuracy and usability of back translation. International Workshop on Intercultural Collaboration (IWIC2007), 477–482.

[17] Somers, H. (2007). The use of machine translation by law librarians - a reply to Yates. Law Library Journal, 99, 611-619.

[18] Yamashita, N. and Ishida, T. (2006a). Automatic prediction of misconceptions in multilingual computer-mediated communication. Proceedings of the 11th International Conference on Intelligent User Interfaces. Sydney, Australia, 62–69.

[19] Yamashita, N. and Ishida, T. (2006b). Effects of machine translation on collaborative work. Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work. Banff, Alberta, Canada, 515–524.

Authors' biographies

Dr. Milam Aiken is a Professor and Chair of Management Information Systems in the School of Business Administration at the University of Mississippi. His research interests include machine translation and multilingual meeting systems.

Dr. Jamison Posey is a Clinical Assistant Professor of Management Information Systems in the School of Business Administration at the University of Mississippi. His research interests include machine translation, multilingual meeting systems, impression management, and outsourced information systems development teams.

Dr. Brian Reithel is a Professor of Management Information Systems in the School of Business Administration at the University of Mississippi. His research interests include digital forensics, software development processes (including team communication enhancement techniques), and the strategic use of information technology in an organizational setting.