Evaluation of Machine-Translated User Generated Content: A pilot study based on User Ratings

Linda Mitchell
SALIS
Dublin City University
Ballymun, Dublin 9, Ireland
linda.mitchell7@mail.dcu.ie

Johann Roturier
Symantec SES EMEA Research
Ballycoolin Business Park
Blanchardstown, Dublin 15, Ireland
johann_roturier@symantec.com

Abstract

This paper presents the results of an experimental pilot user study, focusing on the evaluation of machine-translated user-generated content by users of an online community forum and how those users interact with the MT content that is presented to them. Preliminary results show that ratings are very difficult to obtain, that a low percentage of posts (21%) was rated, that users need to be well informed about their task and that there is a weak correlation between the length of the post (number of words) and its comprehensibility.

1 Introduction

This study follows up on the work described in Roturier and Bensadoun (2011), in which four machine translation systems were compared in order to evaluate their suitability in translating user-generated content. In the present study, the objective is different since feedback on machine-translated content is solicited from actual users of an existing community forum (rather than using linguists or bilingual technical support agents). Thus, an additional objective is to analyse how users interact with the MT content presented to them. This paper is divided into four parts: in Section 2, related work is briefly discussed. In Section 3, the experimental design of this study is presented, while in Section 4 preliminary results are reported. In Section 5, we make some conclusions and outline possible future work.

2 Related Work

The machine-translation of user-generated content has been identified as being potentially useful to allow communication between various user groups that do not share a common language (Flournoy and Rueppel, 2010). Indeed, it was announced in 2010 that TripAdvisor would be using Language Weaver-powered translations to make hotel reviews available in multiple languages [1]. More recently, Facebook announced they would be using Microsoft's Bing Translate for Page content[2]. Recent research work has also been performed in this area, including Roturier and Bensadoun (2011) and Banerjee et al. (2011). However, no study has focused on how machine-translated content would be received in-context by existing users of a forum community.

3 Experimental Design

The current German Norton Community forum[3] is composed of multiple sections, known as "boards". We decided to create a specific board, where machine-translated content would be published[4]. In the introduction to this board, it was explained to users that this board is used to show machine translated posts (from English into German). A user (Max_MÜ) was created; a fictitious "MT robot" whose name is used to post machine-translated content. Additionally, a de-

[1] http://blogs.forrester.com/tim_walters/10-07-15-sdl_casts_vote_machine_translation_language_weaver_acquisition

[2] https://www.facebook.com/photo.php?fbid=10150491112449572&set=a.121044129571.125587.10381469571&type=1

[3] http://de.community.norton.com

[4] http://de.community.norton.com/t5/L%C3%B6sung-nicht-gefunden/bd-p/Max

scription was added to Max_MÜ's profile[5], introducing himself and explaining the study. Max_MÜ's signature says explicitly that each of its post has been machine translated. One communication thread was opened and floated to the top to explain the study and present the users' feedback option including the voting mechanism. Feedback options consist of the newly developed voting mechanism, and the already existing options of commenting on the machine-translated posts and giving kudos, "a way for you to give approval to content that you think is helpful, well-formed, insightful, or otherwise generally valuable in the community"[6].

3.1 Voting Mechanism

To collect genuine user feedback, a voting mechanism was developed. This mechanism consists of the question whether the machine translated post was comprehensible and the option of selecting either "yes" or "no", which is then send to a database via the "vote" button. This was written in Javascript and included on every page of the MT board. It was inserted to the left of each post in the MT board, as shown below:

Figure 1. Feedback mechanism

3.2 Evaluation Criteria

In this experiment, comprehensibility, which refers to "the extent to which the text as a whole is easy to understand" (Hovy et al. 2002), is measured for machine translated user-generated content. It is evaluated in this study using a binary evaluation system: The user answers the question of whether a post was comprehensible or not, with either "yes" or "no" (see Figure 1).

3.3 Evaluation Data

The evaluation data was obtained from the English Norton forum[7]. In a first step, ninety threads were identified from different boards (Norton Internet Security, Norton 360, Online Family and Norton Mac). The threads had to fulfill the condition that in addition to a question, they had to have one post marked as a solution. The process of retrieving two messages per thread (question and answer) from the English forum was automated using API requests and a script in Python. For the translation of the posts the API of the Microsoft Translator system was used[8] since it is the system that had obtained the highest comprehensibility and fidelity scores in Roturier and Bensadoun (2011).

3.4 Experiment Procedure

For three weeks, the MT board was solely opened to the gurus (eight users). During this period, six valid votes were received. This test period showed that the voting mechanism worked and that users would have to be motivated by posts constantly to vote. The board was opened to the public (users and non-users) on 11 January 2012. Every week, ten new threads were posted to the MT board.

4 Results

4.1 Sections

During the evaluation time frame, votes were recorded for non-machine-translated content; after repeatedly specifying that users should only vote for content posted by Max_MÜ. This suggests that users do not necessarily read the introduction to the board or any other related post. Figure 2 shows the number of ratings collected per week. While there was an increase in votes initially, the number of ratings decreased noticeably after week 12. This might be related to user motivation and is a topic that will need to be addressed in the future. The number of different users who voted per week never exceeded five. While the users mostly voted for one or two posts at a time, there were instances of users voting for more posts (e.g. 18 posts Wk6).

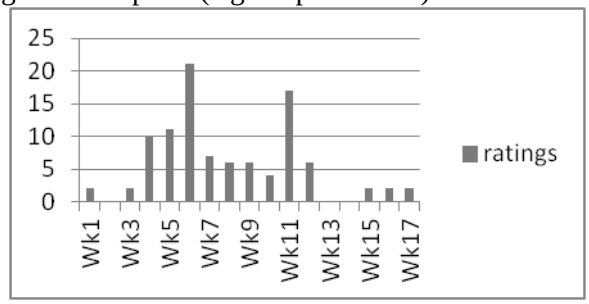

Figure 2. Number of valid ratings per week

[5] http://de.community.norton.com/t5/user/viewprofilepage/user-id/6115
[6] http://community.norton.com/t5/Announcements/New-Feature-KUDOS/m-p/9713
[7] http://community.norton.com/
[8] http://www.microsofttranslator.com/dev/

Between 20 December 2011 and 04 April 201294 valid ratings and 18 invalid ratings, e.g. ratings for non-machine-translated content were collected. Out of the valid ratings, 57 (61%) ratings were "yes", i.e. the machine translated content was rated as comprehensible, and 37 (39%) were "no", the machine-translated content was deemed incomprehensible. There were two more ratings for answers (48) than for questions (46). It is apparent from the results that, both for question and answers, "yes" was the preferred rating, as shown in Figure 3:

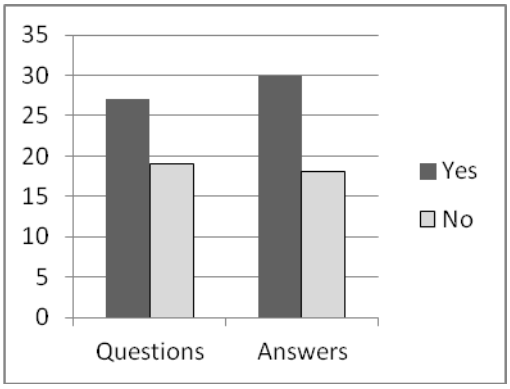

Figure 3. Ratings grouped into questions and answers

4.2 Interpretation of collected ratings

While these are only preliminary results, Figure 3 suggests that some machine-translated posts can be understood by users who do not have access to the source text. This confirms the results from Roturier and Bensadoun (2011), where on average, machine-translated posts were rated 2.6 on a scale of 5 (in terms of comprehensibility). We are interested in finding out whether some textual characteristics, such the length of a post, may have an impact of the comprehensibility ratings. For instance, the average number of words per post in those that were rated as comprehensible was 56, whereas it was 93 for those that were rated as incomprehensible. This suggests that the longer the post, the less likely it is to be comprehensible. This is only supported by a weak correlation (-0.35) between the two variables - when comprehensibility is expressed by 1 or "yes". Thus, the relationship between length of post and comprehensibility seems to be more complex. More research needs to be conducted, e.g. on whether more context increases a post's comprehensibility.

The users made sparse use of the other feedback options available to them (kudos, comments). Five posts received two ratings. Two times, the users voted for the same answer, three times they voted differently. There were 210 threads (420 posts) available in the MT board. Only 88 (21%) of those posts received a rating. No kudos was given to any posts in the MT board. All of the comments (four) received in the MT board indicated that the users had not grasped the concept of the MT board, e.g. they mistook posts by other users as machine-translated content or they did not realise that the content was machine-translated. None of the comments were related to the quality of the actual MT output.

4.3 Visibility of the machine-translated posts and its impacts on rating behaviour

In the previous section, we have shown that machine-translated content could sometimes be understood by users, hence suggesting that it can be of value to these users. We are also interested in determining whether the content that is rated as comprehensible relates to important user issues. To achieve this, we analysed the top search terms on the German Norton Forum for the MT board, but found that no search queries were submitted during that time period. For the German Norton Forum in general, we found that error codes, such as "fehler 3040,20063" or "8920.201" were prevalent. While there was one MT post that had an error code in its subject "Fehler: 8.920.223", there were no searches performed for that particular error code; however, both question and answer received a rating for this thread. This may suggest that posts including an error code (in the subject) are possibly the posts that are most accessible to the users. As the number of available searches is small for the German Norton forum (e.g. 116 single term searches within two months), we analysed the searches in the English Norton community (e.g. 52923 single term searches within two months) in order to determine possible candidates for keywords and to consequently re-rank the posts or change the way of selecting new posts. It was found, for example, that the Norton products are often searched for, as well as different browsers in connection with the Norton toolbar. This and information gained from reports on searches performed in independent search engines will be included in the selection process of threads to be machine translated in future.

Figure 5 shows the number of ratings posts collected depending on their position in the board at the time the rating was submitted. This figure

suggests that most of the ratings are likely to have been generated by users who went to the MT board deliberately, voted for some of the posts on the first page and subsequently left the board. Only six of the ratings received went below thread 9 on a page. The median number of posts voted for by a user in one session is 2, i.e. the number of posts voted for by the same user within a very short time frame. (The average number of posts voted for in one session is 3.5.)

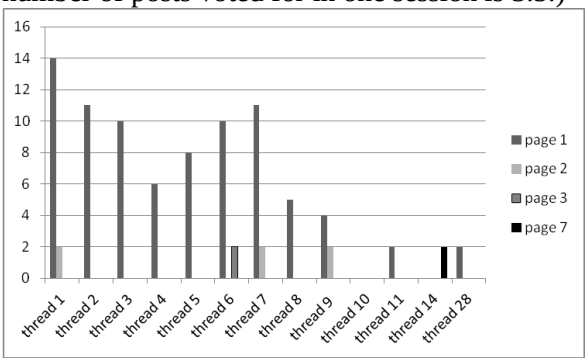

Figure 5. Position of threads voted for in MT board

5 Conclusion

This paper presented the setup and results of a pilot study focusing on the evaluation of machine translated user-generated content in an online community environment. These results point towards the content as being rated comprehensible slightly more often than not. The decision of rating a post as comprehensible may be influenced by the length of the post. The drawbacks of this study were that a limited number of ratings were collected. This is connected to the issue of motivation. It can be concluded from this study that the users need to be constantly reminded and, more importantly, motivated to vote. A possible reason for the low motivation to vote may be that a platform for German speakers is already in existence. Thus, it would be beneficial to the project, to see whether motivation to vote would increase for a language that does not have a community yet, e.g. in a Spanish board. By broadening the setup, we are hoping to receive a larger number of votes and a more general idea of whether MT content is acceptable for the users of an online community.

In addition to this, the machine-translated content could be made more relevant to the user by selecting the threads based on the findings of the analysis of search queries performed within and outside the Norton community. Some of these issues will be tackled within the framework of an FP7-funded project, ACCEPT[9].

References

Banerjee, P., Naskar, S. K., Roturier, J., Way, A. and van Genabith, J. 2011. Domain Adaptation in Statistical Machine Translation of User-Forum Data using Component Level Mixture Modelling. In Proceedings of the Thirteenth Machine Translation Summit, pages 285–292, Xiamen, China.

Flournoy, R., and Rueppel, J. 2010. One Technology: Many Solutions. Proceedings of AMTA 2010: the Ninth Conference of the Association for Machine Translation in the Americas. Denver, Colorado.

Hovy, E., King, M. and Popescu-Belis, A. 2002. Principles of Context-Based Machine Translation Evaluation. Machine Translation 17 (1), 43-75.

Roturier, J. and Bensadoun, A. 2011. Evaluation of MT Systems to Translate User Generated Content. In Proceedings of the Thirteenth Machine Translation Summit, pages 244–251, Xiamen,China.

[9] http://www.accept.unige.ch/Description.html