How Good Is Crowd Post-Editing?
Its Potential and Limitations

Midori Tatsumi
Toyohashi University of Technology
midori.tatsumi2@mail.dcu.ie

Takako Aikawa
Microsoft Research, Machine Translation team
takakoa@microsoft.com

Kentaro Yamamoto
Toyohashi University of Technology
yamamoto@lang.cs.tut.ac.jp

Hitoshi Isahara
Toyohashi University of Technology
isahara@tut.jp

Abstract

This paper is a partial report of a research effort on evaluating the effect of crowd-sourced post-editing. We first discuss the emerging trend of crowd-sourced post-editing of machine translation output, along with its benefits and drawbacks. Second, we describe the pilot study we have conducted on a platform that facilitates crowd-sourced post-editing. Finally, we provide our plans for further studies to have more insight on how effective crowd-sourced post-editing is.

1 Introduction

As the use of machine translation (MT) together with post-editing (PE) has become one of the common practices to achieve cost-effective and high quality translation (Fiederer & O'Brien 2009, Koehn 2009), and crowdsourcing is gaining popularity in many areas including translation (Désilets 2010, Zaidan & Callison-Burch 2011), one can easily imagine that 'crowd PE' is going to be a strong trend in the MT community in the near future.

This paper presents a preliminary investigation on the effectiveness of crowd PE. We conducted a pilot study using Collaborative Translation Framework (CTF) developed by the Machine Translation team at Microsoft Research. Having CTF as a platform of crowd PE, we translated the English websites of Toyohashi University of Technology (TUT)[1] into nine languages with very little cost (Aikawa et al. 2012). We analysed the results from this pilot study quantitatively in an attempt to evaluate the validity and the effectiveness of crowd PE.

The organization of this paper is as follows: In section 2, we discuss the past and the current situation of crowdsourcing in text and contents production, and state the goal of our research. Section 3 presents a brief explanation of our pilot study at TUT and its results. In section 4, we provide some results from the human evaluation on the quality of the crowd PE, and the results from the evaluation by means of an automatic metrics. Section 5 discusses the results from Section 4, while raising our renewed research questions. Section 6 summarises the paper.

We are aware that building and maintaining appropriate platforms and communities is an important aspect of crowd PE, and a number of research efforts are being made on those topics. Our paper, however, is focused on the quality we can expect from crowd members, and thus building and maintaining platforms and communities is out of the scope of this paper.

[1] http://www.tut.ac.jp/english/introduction/

2 Crowd Post-Editing or 'CPE'

The power of crowd resource in producing translation has been proven in a number of areas from fansubs (Cintas & Sánchez 2006, O'Hagan 2009) to social media such as Facebook (Losse 2008) to popular conference video site, TED[2], to community participation in product development at Adobe[3] and Symantec (Rickard 2009). This makes one think: if crowd translation has been successful, why not crowd post-editing? It may not be too extravagant to even speculate that crowd PE has more potential than crowd translation; considering that crowd members are often not professional translators or linguists, PE may seem to them as a less demanding task than translating from scratch (though in reality PE of MT sometimes can be more demanding than translation depending on the MT quality).

In fact, researchers and businesses have already started to study and test the potential of this method (Muntes & Paladini 2012). However, the current focus is mainly on developing platforms to facilitate the participation of crowd members and frameworks for quality control. The actual quality of the crowd PE outcome has not yet gained much attention.

Crowd PE can have different types of resources. Some cases may hire random crowd resources with a small monetary reward (e.g., Amazon's Mechanical Turk), others may be done by enthusiastic fans of the subject matter, or some others may even employ only the internal members of an organisation or a community, involving no payment. The latter cases can be more appropriately called 'Community PE' or 'Collaborative PE' than 'Crowd PE'. In this paper, we do not differentiate these different types of resources, and will use the acronym 'CPE'.

2.1 Advantages of CPE

MT + CPE, similar to crowd translation, can be advantageous in a number of aspects compared to MT + professional PE (i.e., post-editing done by professional translators and post-editors). The following lists such advantages.

Cost: CPE is less expensive than professional PE, which is especially important for non-profit organisations and/or the types of contents that need to be updated frequently. This, however, only applies to the per-word cost, and the initial investment on developing the platform, framework, interface, etc. needs to be taken into account when evaluating the total cost.

Speed: Crowdsourcing often proves to be equally quick, or sometimes even quicker, than the traditional style commercial works[4].

Domain Knowledge: Although crowd members are not expected to have linguistic expertise, they are often highly knowledgeable in specific domains.

Community Development: Crowd members can get the sense of community by participating in CPE. In addition, CPE might give the contributors an opportunity to become more familiarised with the community topics and issues as they try to read and understand the contents more deeply than they would as a mere reader.

Confidentiality: CPE also has a potential to be an ideal solution for translating sensitive contents in an organisation. Translating the text by an MT system and have internal members to perform CPE can eliminate the fear for information leakage (provided enough resources can be secured within the organisation).

2.2 Drawbacks of CPE

One big challenge CPE would face is how to assure the quality of CPE. To address this issue, most, if not all, of the crowdsourcing platforms provide one or more ways to control the quality of the crowd-sourced products. One of the common methods is to have one or more moderators to check and ensure the quality of the product. Another common method is rewarding and/or ranking mechanism that gives various rewards and/or quality statuses to the crowd members based on the past performance. Such mechanisms are designed to encourage the participants to make more contribution with higher quality jobs.

[2] http://www.ted.com/OpenTranslationProject

[3] The blog article written by Dirk Meyer is available at: http://blogs.adobe.com/globalization/collaborative-translation-helps-adobe-business-catalyst-add-new-languages/

[4] One example is the translation of movie subtitles in China (Chipchase, J. & Wang, F. "subtitle team, crowd sourced translation in China". Available at: http://janchipchase.com/2011/09/chinese-bandit-translation-teams/).

These solutions can help to overcome the quality assurance issue, but it can also incur a great amount of effort and investment to develop and maintain complicated frameworks and platforms. If we know what level of quality we can expect from CPE, it would help to make a necessary and sufficient investment on quality assurance. This paper is a step stone to this goal.

3 Pilot Study

This section provides a brief description of our pilot project conducted at TUT, which we mentioned at the beginning of the paper.

3.1 Motivation and setting

TUT has more than 200 foreign students from various countries, and the demand to localise the information on their websites into various languages has always been strong. Yet, localising the websites using professional translators is just too expensive. To make the university information more accessible to current foreign students and to prospective students, the university created an English version of their websites. However, still many foreign students had problems in understanding the information because of the language barrier. To overcome this issue, the university decided to translate the English websites into nine languages by means of Microsoft Translator's Widget, and have their foreign students to post-edit the MT output.[5]

Foreign students at TUT were ideal crowd resource for this project as they are familiar with the contents of the TUT's websites, and they are willing to make a contribution to this project with a small monetary reward. We hired a total of 22 foreign students [6] with nine different language backgrounds shown in Table 1.

3.2 Conducting the CPE Session

Prior to starting the project, we gave the students a brief introduction on how to use CTF user interface and explained the background of the project. We also provided the following CPE guidelines:

Avoid over-editing: don't try to over-edit if the existing translation(s) (whether they are MT output or other human edits) are grammatical and readable.

Ignore stylistic differences: don't try to modify stylistic differences unless they are critical or matter for readability.

Start from scratch: if the quality of MT output is too low, provide your translation from scratch (as opposed to modifying MT output).

It is important to note here that we did not prevent the students from modifying existing CPE results provided by other students. The students are allowed to modify not only the MT output but also any one of the previous CPE results as they think is necessary.

We assigned each student 30 hours for performing CPE. The CPE sessions were conducted in November-December, 2011. The details on the workflow of the CPE and the design of CTF are provided in (Aikawa et al. 2012).

3.3 Results

Table 1 shows the descriptive statistics of the results of the pilot study.

Language	Participants	Sentences	Edits
Arabic	2	397	723
Chinese[7]	6	1637	2269
French	2	512	647
German	1	147	192
Indonesian	2	1285	1559
Korean	2	598	707
Portuguese	1	204	308
Spanish	4	1841	3643
Vietnamese	2	1341	1929

Table 1. Summary of the results

The Sentences column shows the number of the sentences that were edited, [8] and Edits column shows the total number of sentences resulted from

[5] This is a collaboration project between TUT and Microsoft Reseach. See Yamamoto et al. (2012) for our initial report.

[6] Strictly speaking, the total number of student participants was 21 as one of the students edited both Arabic and French MT output.

[7] This study involved only simplified Chinese.

[8] Note that there were cases where no CPE was provided as MT output were acceptable enough. We did not study such cases as the focus of this study is the effect of CPE, and not the quality of MT

CPE, for each language. The gap between the two indicates that some sentences have received multiple CPE. Following is an example where multiple CPE were performed for Spanish:

[English source text]
You must show this table to the banker before sending your money.

[MT output]
Se debe mostrar esta tabla para el banquero antes de enviar su dinero.

[First CPE result]
Se debe mostrar esta tabla al banquero antes de enviar su dinero.

[Second CPE result]
Se debe mostrar esta tabla al empleado del banco antes de enviar su dinero.

Overall, the figures in the table show that the combination of Microsoft Translator's Widget and CTF has been well adapted as a community translation environment such as university websites. We have received a fair number of CPE outputs from the participant students, which demonstrates their enthusiasm. Using the crowdsourcing power of the foreign students at TUT, the majority of the university's English websites was localised into nine languages within two months with inexpensive cost.

We asked the participant students to give feedback about their experience as a CPE contributor. The students, though not having professional translation experience or linguistic expertise, seemed to have worked quite comfortably and confidently in the provided CPE environment, and their overall feedback was very positive. They also mentioned that participating in this project as a CPE contributor gave them the strong sense of community.

Now the important question we need to ask is: how good was the quality of CPE? We address this question in the next section.

4 Quantitative Analysis

4.1 Human evaluation

Among the nine languages post-edited for this pilot study, we chose four languages that had higher number of sentences post-edited than other languages, namely, Chinese, Indonesian, Spanish, and Vietnamese, to evaluate the CPE results. To this end, we hired professional translators and asked them to choose the best translation among all the translations (which consist of MT output and CPE results) in the sense that it reflects the meaning of the source text. We advised them not to worry about stylistic or registry differences. We also asked them to provide their own translation in case none of the existing translations conveyed the correct meaning of the source text. To make this evaluation a blind test, we randomised the order of the MT output and all the CPE results. This way, the evaluators (professional translators) could not tell which translation was from MT or CPE based on the order of the sentences.

For the purpose of a cross-language comparison, we focused only on the test sentences that had been post-edited for all four languages; there were 567 such sentences.

The following table shows the frequency of the occurrences of single and multiple CPE for each of the 567 test sentences.

Number of CPE	Chinese	Indonesian	Spanish	Vietnamese
1	372	441	196	350
2	137	95	175	154
3	41	24	88	43
4	10	5	46	17
5	6	2	28	2
6	1		22	1
7			8	
8			1	
9			3	
Average Number of CPE	1.49	1.29	2.39	1.54

Table 2. Frequency of multiple CPE

According to Table 2, except for Spanish, more than 60% of the test sentences had only one CPE output, and more than 95% less than three CPE outputs.

Here we make a simple assumption: among all CPEs, the last one should be the best one, assuming that the last one is the result of the collective intelligence of all the CPE contributors worked on a given sentence. When a sentence is post-edited by more than one person, the second person onward can see not only the MT output but also the previous contributors' editing results, thus can gain better idea of what an acceptable translation should be like, by learning from other people's editing.

In order to find out if this is true, we distinguish the last CPE output from other CPE outputs. In the analyses and descriptions below, we will use the following terms:

MT: Machine Translation output

LCPE: The Last CPE output for each test sentence. When there is only one CPE output, it becomes the LCPE.

XthCPE: All CPE outputs other than LCPE.

Revision: Revised text provided by the professional translators.

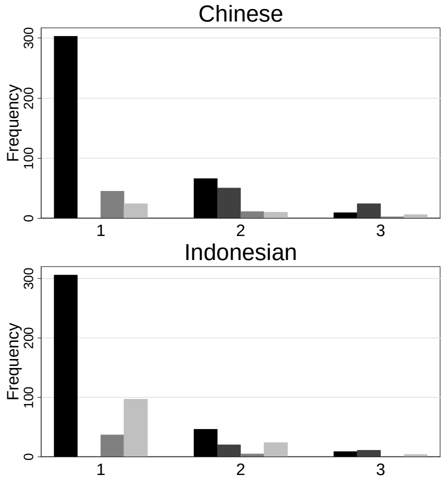
(When we just say ‘CPE’, it includes both XthCPE and LCPE.)

The following table shows the human evaluation results and the numbers of the cases where LCPE, XthCPE, or MT was selected or a Revision was provided for each language. The greyed area indicates the percentages. Note that when MT or XthCPE was selected and when it was exactly the same as LCPE, we counted that into LCPE. Likewise, when MT was selected and it was exactly the same as an XthCPE, we counted that into XthCPE.

Selected as Best/Revised	Chinese	Indonesian	Spanish	Vietnamese
LCPE	383	364	261	334
	68%	64%	46%	59%
XthCPE	85	34	154	67
	15%	6%	27%	12%
MT	58	42	50	22
	10%	7%	9%	4%
Revision	41	127	102	144
	7%	22%	18%	25%
Total	567	567	567	567

Table 3. Human evaluation results

Overall, LCPE is the most frequent choice for all languages, though the percentage varies from the highest of 68% for Chinese to the lowest of 46% for Spanish. This is generally good news, but it also means that our assumption that LCPE should be the best was not right for around 30 to 50% of the cases. XthCPE was selected as the best translation in 6 to 27% of the time, and MT 4 to 10% of the time. This means that one or more CPE contributors transformed the MT or existing CPE results that had acceptable quality into the one that did not. In order to further investigate this, we looked at the evaluation ratio by the number of CPE outputs. The following figures show the results (we only looked at the cases where one, two, or three CPE was performed, as there were not many cases for which more than three CPE outputs were available). Note that there is no bar for XthCPE for the category 1, as this is the case where there is only one CPE, that is, LCPE.



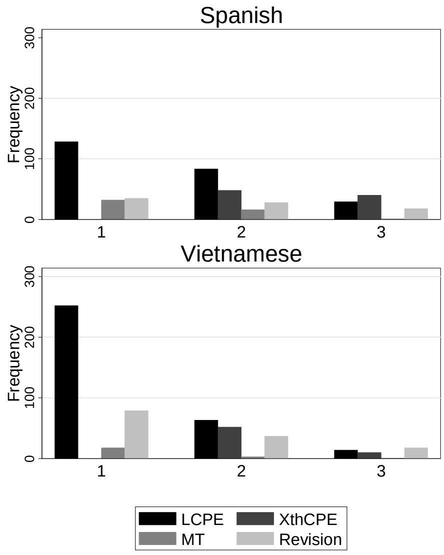


Figure 1. Relationship between the number of CPE output and the evaluation

As the figures show, for most of the cases where LCPE was selected, LCPE was the only CPE output (category 1). Interestingly, LCPE is still the best choice when one more CPE was done (category 2), but for the test sentences where CPE was performed three times (category 3), XthCPE was slightly more frequently chosen as the best translation, except for Vietnamese. This may mean that after the third CPE, the quality of the CPE output tended to deteriorate. We would like to investigate this issue further in the future.

There are 7 to 25% of the cases where professional translators did not find any satisfactory translation and provided a Revision. We were interested in finding out if there are any prominent source text characteristics that may have caused low quality CPE. As a starting point, we compared the average source sentence length in words between the sentences for which LCPE was chosen as the best translation and those for which Revision was provided. The following figure shows the result.

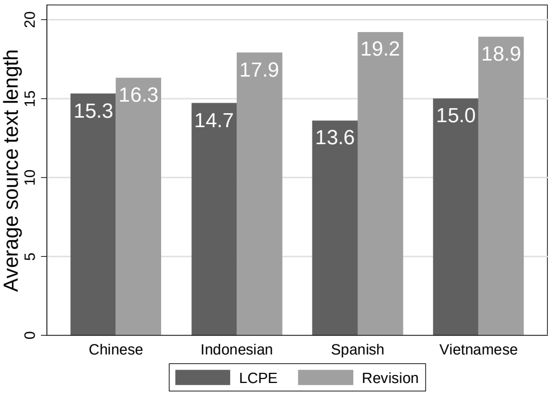


Figure 2. Comparison of source sentence length in two different cases

As shown in Figure 2, the average length of the source sentences that ended up having professional translators to provide the Revisions was longer than that of the sentences where LCPE achieved a good enough quality.[9] The average length for all 567 source sentences is 15.9 words.

4.2 Evaluation with TER

Next, we focused on two cases: Case I, where LCPE was selected as the best translation, and Case II, where the professional translator revised LCPE to produce acceptable translation.[10] This was to see 1) how much editing was done by CPE contributors in order to transform the MT output of unacceptable quality to the translation of acceptable quality, and 2) when LCPE was better than MT or XthCPE yet not quite good enough to be regarded as an acceptable translation, how much editing was necessary by the professional translators to produce Revisions.

To measure these, we used TER (Translation Edit/Error Rate)[11]. TER (Snover et al. 2006) is one of the automatic metrics developed for MT quality evaluation. It compares two sentences and calculates a score based on the number of minimum editing operations necessary to

[9] P<0.01 for Indonesian, Spanish, and Vietnamese. Statistical significance was not observed for Chinese.

[10] This, however, involves some subjectiveness. When the translator provided Revisions, the revised text is inserted next to the text that the translator thought was the closest to the acceptable translation. However, the revised text sometimes ends up in becoming closer to other text than the one they have chosen.

[11] http://www.cs.umd.edu/~snover/tercom/

transform one sentence to another. The perfect match gets a score of 0 (0 edits needed), and the score gets higher as the difference between the two sentences becomes larger. As it uses word as an editing unit, we used Stanford Chinese Word Segmenter[12] to tokenise the Chinese text.

We took TER scores between MT and LCPE for Case I, and between LCPE and Revision for Case II mentioned above. The average TER scores for the two cases are shown in Table 4.

Language	Case I: TER between MT and LCPE	Case II: TER between LCPE and Revision
Chinese	54	27
Indonesian	38	26
Spanish	40	34
Vietnamese	49	27

Table 4. Average TER for the two cases

The results show that, for Case I, Chinese got the highest score among four languages, which means that, on average, it took CPE contributors more editing to transform an MT output into an acceptable quality translation in Chinese than other languages.

On the other hand, for Case II, Spanish got the highest score, which means that it took professional translators more editing to fix LCPE to produce an acceptable level translation than other languages. We plan to investigate such language differences in more details in the future.

Overall, for all languages, the average TER scores between LCPE and Revision are significantly smaller than the scores between MT and LCPE.[13] This may suggest that even when LCPE could not achieve acceptable quality, the amount of Revision work necessary to improve such text to an acceptable level quality can be smaller than revising the MT output from scratch.

5 Discussions and Ongoing Studies

Overall, the above mentioned results suggest the following:

- Around 50 to 70% of the time LCPE produces good enough translation
- Longer source sentences may cause difficulty for CPE contributors to produce acceptable quality
- Even when LCPE result is not good enough, the amount of necessary additional revision work may be rather small

These results generally show that CPE can be a great help in raising the MT output quality to an acceptable level. However, there were still cases where professional translators found LCPE results unsatisfying or LCPE having lower quality than XthCPE or even MT. This gave us renewed research questions (RQ) listed below.

RQ1: In what kind of cases do CPE contributors fail to produce acceptable translations?

We found that the number of the cases professional found LCPE results unacceptable varies among the languages. However, the numbers alone do not tell us 'why'. In order to understand more deeply in what cases and in what way LCPE failed to produce an acceptable translation, we will need to examine the results qualitatively.

RQ2: Would having the larger number of CPE contributors be of help in achieving acceptable quality?

We found that 46 to 68% of the time LCPE was selected, but would the percentage increase if we ensure each MT output is post-edited by certain number of CPE contributors? Would the quality keep increasing to the point where the professionals' intervention becomes unnecessary?

In order to answer these questions, we are now in the process of the following two further studies.

5.1 Qualitative Analysis

In order to answer RQ1, we are having one native speaker of each target language, who has some translation experience, but not the same person who did the evaluation task explained in section 4.1, to explain the difference between CPE

[12] http://nlp.stanford.edu/software/segmenter.shtml

[13] P<0.05 for Spanish. P<0.001 for Chinese, Indonesian, and Vietnamese.

outcome and its Revision. The interview sessions will be held in August 2012.

Based on the results of the interviews, we are hoping to have insights into what kinds of necessary editing CPE contributors tend to achieve or fail to achieve, for each language, and also for all languages.

5.2 Controlled Experiment

In order to answer RQ2, we are conducting a controlled experiment in which all the sentences are ensured to be post-edited by certain number of CPE contributors.

We predict that, after certain number of editors, there will be nothing left to improve, and hence editing would become 'saturated'.

We are interested in finding out the following:

- *Would the percentage of LCPE selected by the professional translator increase when we have more CPE contributors?*
- *If that is the case, how many is enough?*

We are currently running an experiment to answer these questions.

6 Concluding Remarks

In this paper we first discussed the current situation and the potential of crowd PE. Then we explained our pilot study on the impact of crowd PE, presenting some quantitative results from the human evaluation and the evaluation by means of TER. Finally, we stated our further research questions and introduced our ongoing research effort.

Acknowledgments

This project was funded by International Affairs Division at Toyohashi University of Technology, and we would like to give special thanks to all the members of International Affairs Division for their support during the project. We are also thankful to Dr. Anthony Hartley for his support on conducting the experiment.

References

Aikawa, T., Yamamoto, K., & Isahara, H. (2012) The Impact of Crowdsourcing Post-editing with the Collaborative Translation Framework. In: Proceedings of the 8th International Conference on Natural Language Processing, Kanazawa, Japan

Cintas, J.D. & Sánchez P.M. (2006) Fansubs: Audiovisual Translation in an Amateur Environment. *The Journal of Specialised Translation,* 6, pp. 37-52

Désilets, A. (2010) Collaborative translation: technology, crowdsourcing, and the translator perspective. Introduction to workshop at *AMTA 2010: the Ninth conference of the Association for Machine Translation in the Americas*, Denver, Colorado, October 31, 2010; 2pp

Fiederer, R. & O'Brien, S. (2009) Quality and machine translation: a realistic objective? *Journal of Specialised Translation,* 11, pp. 52-74.

Koehn, P. (2009) A process study of computer-aided translation. *Machine Translation,* 23, 4, pp. 241-263.

Losse, K. (2008) "Achieving Quality in a Crowd-sourced Translation Environment". Keynote Presentation at the 13th Localisation Research Conference Localisation4All, Marino Institute of Education, Dublin, 2-3 October 2008

Muntes, V. & Paladini, P. (2012) "Crowd Localization: bringing the crowd in the post-editing process". Presentation at Translation in the 21st Century – Eight Things to Change, Paris, May 31 – June 1 2012

O'Hagan, M. (2009) Evolution of User-generated Translation: Fansubs, Translation Hacking and Crowdsourcing. *The Journal of Internationalisation and Localisation,* Volume I 2009, pp. 94-121

Rickard, J. (2009) Translation in the Community, Presentation at LRC XIV conference, Localisation in the Cloud, 24-25 September, Limerick, Ireland, available at http://www.localisation.ie/resources/conferences/2009/presentations/LRC_L10N_in_the_Cloud.pdf

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006) A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, pp. 223-231.

Yamamoto, K., Aikawa, T. & Isahara, H. (2012) 機械翻訳出力の後編集の集合知による省力化. In: Proceedings of the 18th Annual Meeting of the Association for Natural Language Processing, Hiroshima, Japan, pp. 499-500

Zaidan, O. & Callison-Burch, C. 2011. Crowdsourcing translation: professional quality from non-professionals. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 1220–1229.