Development of a Filipino-to-English Bidirectional Statistical Machine Translation System that dynamically updates via user feedback

Jasmine Ang, Marc Randell Chan, John Paolo Genato, Joyce Uy, Joel Ilao

College of Computer Studies
De La Salle University – Manila
jasmine_ang@dlsu.edu.ph, marc_chan@dlsu.edu.ph, paolo_genato@yahoo.com,
joyce_uy@dlsu.edu.ph, joel.ilao@delasalle.ph

Abstract

In this paper, we describe a Moses-based statistical machine translation (SMT) system, called FEBSMT, that incorporates periodic user feedback as a mechanism that allows the SMT system to adapt to prevailing translation preferences for commonly queried phrases, and assimilate new vocabulary elements in recognition of the dynamically changing nature of languages. A parallel corpus containing a total of ~22K sentences in the tourism domain was used in developing the system. Updating the SMT's language model and phrase tables via user feedback was modeled after the Post-Edit Propagation (PEPr) system [6]. Incremental training iterations were performed on the developed system via user feedback, which were collected in a duration of three months. The developed system was evaluated using the BLEU, NIST, METEOR, and TER metrics. We noted that the Filipino-to-English translations consistently scored higher than the English-to-Filipino translations. Over the course of 100 training iterations using randomly selected sentences taken from a closed set of sentences provided with user feedback, it was observed that the translation accuracy sharply improves within the first few iterations, which then gradually tapers after a peak translation performance has been reached.

1. Introduction

In the context of facilitating communications among citizens of ASEAN member countries especially as it prepares for economic integration in 2015, the ASEAN Machine Translation (ASEAN-MT) Project was launched [5]. The initial design of the ASEAN-MT system uses English as a pivot language to perform translation between pairs of major languages of the ASEAN member countries.

Furthermore, these machine translation systems can also contribute to the United Nations Millennium Goal of Developing a Global Partnership for Development [8]; since, one of its target condition is to "make available benefits of new technologies, especially information and communications" through the use of the Internet. However, not all pieces of information are available in English and not all are translated correctly. Hence, multiple improved machine translation systems are effective in developing bridges for information dissemination.

2. Related Work

2.1 Tools

2.1.1 Moses

Moses is an open-source toolkit for statistical machine translation that allows one to automatically train translation models for any language pairs [3]. It does training for any language pair with the use of a parallel corpus. The parallel corpus is separated into training, development and testing sets. The training set is where the bilingual phrases are extracted and their weights are learned. The development set is used to adjust the values of the parameters of the decoder, while the testing set is used for assessing the translation quality. For this project, Moses setting chosen for training the Filipino-English bidirectional SMT system are as follows: language model (LM) order of 3, cleaning range of 1-80, and the decoder's distortion limit of 6 [4]. The setting for FEBSMT was based from the previous Philippine Component of the ASEAN project in order to track the improvement in the machine translation technology.

2.1.2 PEPr

Post-edit Propagation (PEPr) is a phrase-based statistical machine translation system and uses an automatic post-editing (APE) setting with learning capabilities [6]. The APE system automatically post-edits the machine translation output into a proper text with human quality. Moreover, this approach aims to handle various errors, ranging from determiner selection to grammatical agreement. The APE system is built using the data comprised of the baseline translations and their post-edited counterparts.

In performing the Post-edit Propagation, the system has to undergo a cycle of two processes as shown in Figure 2.1.

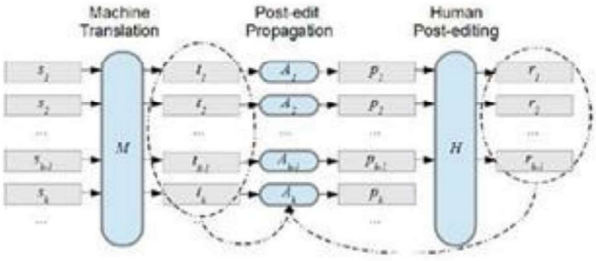

Figure 2.1 The Feedback Process of PEPr. Figure extracted from [6]

The first process involves the training of the baseline system, labeled M in Figure 2.1. The output from the baseline system is passed into the APE system. The baseline system is treated as a black box since no modification will be performed. The second process involves the APE system and a human post-editor. The baseline translations are subjected to human post-edits and these pairs of texts are used for the training of the APE system. Further version of APE systems were trained using the translations of the previous APE with their corresponding post-edits.

The APE system relies on the phrase table and language model of the previous APE version and combines them to the current. For the language model combination, linear mixture model is applied; while for the phrase table combination, linear interpolation is also applied. This process is used to broaden the vocabulary in the language model and balance the probability of the phrase table based on the post-edits.

When translating, the input text will first pass through the baseline system, and then pass through the latest APE system to be automatically post-edited. For this research, the concept of PEPr's APE system was applied due to its flexibility in taking user feedback or user post-edits as input to build the next APE system to improve machine translation quality.

3. System Design of FEBSMT

The aim of FEBSMT is to use post-editing approach to improve the translation accuracy of the machine translation. It is a web service application wherein users can translate Filipino and English bi-directionally. The development of FEBSMT is composed of three phases, namely, the training phase, the development phase, and the testing phase. The discussions of these phases are found in the succeeding sections.

3.1 Training Phase

The training phase consists of two parts. The first part of the training phase is data gathering and the second part is the data cleaning.

3.1.1 Data

The data was gathered from the Center for Language Technologies (CeLT) of De La Salle University (DLSU). It is a Filipino-to-English parallel corpus containing 22,031 sentences of a parallel corpus in the tourism domain. The parallel corpus is randomly split into 70% for the training of the baseline system (Block *M* in Figure 2.1). For evaluation, 10% of the data were selected for testing the machine translation accuracy and 20% were used for the development set.

3.1.2 Data Cleaning

Cleaning of data ensures that the data does not contain spelling errors, special characters, and tags. The data was also tokenized and re-cased into their lowercase.

3.2 Development Phase

For this phase, the development set from the data was used for the simulation of the feedback mechanism. This set of data was used to build multiple APE versions for 100 iterations. This is to observe the changes in the translation quality per APE version. For instance, the sentence "This is from room 208." is translated into "Ito mula sa room 208.". Although the translation is semantically correct, however it is grammatically incorrect. The proper translation should be "Mula ito sa room 208.". This is to be used as the feedback for the APE.

3.3 Testing Phase

A dataset containing 10% of the parallel corpus was used as the testing data for verifying the accuracy of the machine translation using the four evaluation metrics: BLEU [5], NIST [2], METEOR [1], and TER [7]. The testing data was made constant in order to provide a consistent evaluation of FEBSMT. For this project, APE was implemented and initially ran for five times on the same corpus. Every instance of the five iterations and the baseline system was subjected to the testing. The results showed a convergence of all the metric scores.

4. Results and Discussion

This section enumerates and explains the procedures of the succeeding experiments using both Filipino-to-English and English-to-Filipino sets of different human feedback and testing data. It also discusses the purpose of each experiment, along with its results. The results were analyzed and evaluated with the different evaluation metrics. Furthermore, the results will be the basis for the evaluation of the entire FEBSMT system.

For the experiments, the baseline development data and human feedback data were used in conducting the experiments with each set differing in size, content, and context. Two sets of testing data were used for testing the incremental training approach: the 10% baseline testing data and the 100 sentences randomly selected from the entire tourism corpus. This was done to maintain a consistent comparative reference to each other and to the baseline system.

For the automated evaluation metrics, four metrics were used, namely: BLEU, NIST, METEOR, and TER. BLEU and NIST are both precision-based metrics, which score the number of the target translation matches to the reference. METEOR, an F-score metric, measures precision and recall, the number of matches between the target, the reference and their explicit word ordering. TER counts the number of post-edits required to change the target translation to the reference. For the evaluation metric score of BLEU, NIST, and METEOR, a higher value means more matched words between the translation output and reference translation. If the score for TER is lower, the similarity between the translations is greater.

4.1 APE Training with Human Feedback

The purpose of this experiment is to determine if incrementally training the system using human feedback will improve the machine translation quality.

For this experiment, 20% of the tourism corpus was used as the development set. The development set was translated in the baseline system and was subjected to manual post-editing to be used as the feedback for the 100 incremental training iterations. In each incremental training phase of the APE, a total of 1000 sentences were randomly selected from the baseline translation of the development set paired with their corresponding post-edited

counterpart. For evaluating the system, 10% of the same corpus was used as testing data.

The differences between the translation quality of Filipino-to-English and English-to-Filipino in terms of their evaluation scores can be observed in Figures 4.1 to 4.4. The range of scores throughout the 100 incremental training iterations for Filipino-to-English translation is between 0.36 to 0.38, while 0.32 and 0.33 for English-to-Filipino translation. Fluctuations and abrupt increase of scores occurred for both experiments. The peak in the translation scores occurred in the 6th iteration for the Filipino-to-English translation, and in the 15th iteration for the English-to-Filipino translation, obtaining a BLEU score of 0.3795 and 0.3346, respectively.

The scores of the 6th and 15th APE iterations, which obtained the highest scores, however, still have values lower than the baseline score. This means that while there is an inconsistency in the scores of the two experiments, the baseline system still displayed a translation that is closer to a human quality base from the BLEU and NIST scores but the TER evaluation metric was higher by 0.4071 and the METEOR score was very low. These observations suggest that the baseline system's translations have many extraneous words and incorrect word reordering.

A word can have many different translations coming from different contexts, and this tendency was observed to be the possible cause of the decrease in scores. In comparison, there is a more apparent decline in the scores of Filipino-to-English, unlike in the English-to-Filipino where the scores were significantly fluctuating.

4.2 Error Analysis

For thorough comparison, the results from baseline, 6th, 15th, and 100th incremental training were selected. Baseline is necessary to serve as the benchmark of comparison. The 6th and 15th incremental training was chosen for having the highest resulting BLEU score for English-to-Filipino and Filipino-to-English, respectively. This is to observe whether the value of the BLEU score has any effect on the actual translations. Lastly, the 100th incremental training was chosen as a representative of future incremental training of the APE system.

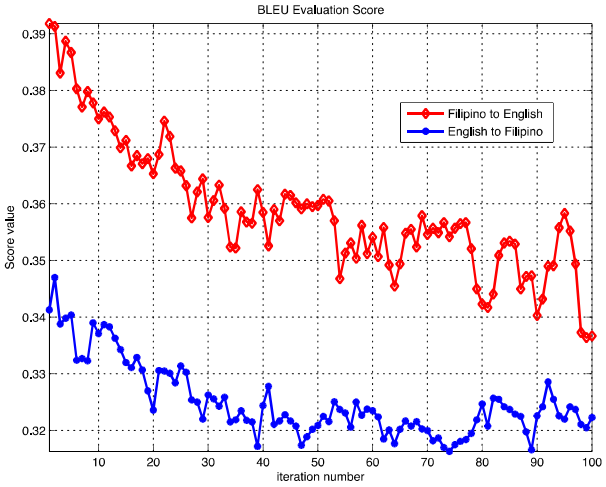

Figure 4.1: Bi-directional Filipino-to-English BLEU Evaluation Score using Human Feedback

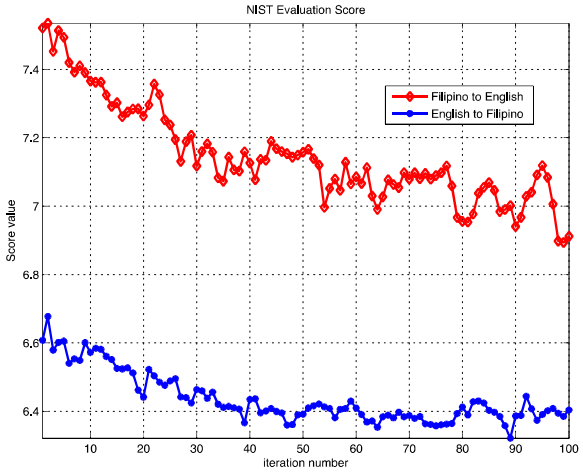

Figure 4.2: Bi-directional Filipino-to-English NIST Evaluation Score using Human Feedback

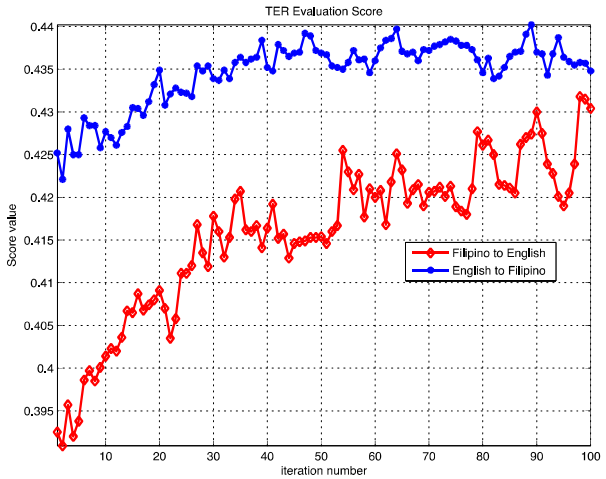

Figure 4.3: Bi-directional Filipino-to-English TER Evaluation Score using Human Feedback

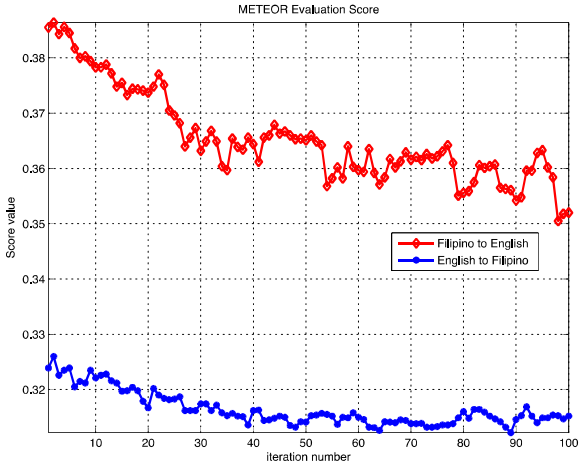

Figure 4.4: Bi-directional Filipino-to-English METEOR Evaluation Score using Human Feedback

Table 4.1: Frequency Count of Errors Based on Phrase Length using the professional translator's Target Translation

5.	Complete	Under Translation	Over Translation	Equal Translation
Eng-to-Fil Baseline	11	26	40	23
Eng-to-Fil 101st APE	20	26	30	24
Fil-to-Eng Baseline	41	26	20	13
Fil-to-Eng 101st APE	28	24	27	21

Table 4.2: Frequency Count of the Types of Error for the Target Translation

6.	Mistranslation	Parts of Speech	Word Order	Untranslated
Eng-to-Fil Baseline	27	19	13	11
Eng-to-Fil 101st APE	30	24	14	11
Fil-to-Eng Baseline	17	14	13	15
Fil-to-Eng 101st APE	29	15	22	9

In performing error analysis and evaluation of FEBSMT, a professional translator provided 100 sentences as feedback for comparison purposes. The professional translator also provided a set of categories for classifying the errors, namely *complete*, *under translation*, *over translation*, and *equal*. For an error to fall under this set of categories, the word count of the translation output is compared against the word count of the professional translator's feedback.

A translation is *complete* if the word count for both sentences is equivalent, with the context of the sentences being the same. If the contexts of the compared sentences are different, the translation will be classified under the *equal* category. *Under translation* means that the word count of the target translation is less than the feedback, while *over translation* means it has exceeded. When a translation is classified as *under*, *over* or *equal*, the type of errors that caused the failure in translation is checked.
The four main types of translation errors that were considered are *mistranslation*, *parts of speech*, *word order*, and *untranslated*,.explained as follows:

- *Mistranslation* – This is the failure to translate a part of the source sentence to its correct translation, or when the target sentence is unintelligible.
- *Parts of Speech* – This error occurs when there is a wrong usage of pronouns, tenses or verb agreements.
- *Word Order* – This error occurs when a word is misplaced in a sentence.
- *Untranslated* – This error occurs when words from the source sentence are retained in the target translation.

As shown in Table 4.1, the results for both the baseline and the 101st incremental training of English-to-Filipino are close to each other. Although, the 101st incremental training managed to obtain more complete sentences, it lessened the over translated sentences, and increased the number of equally translated sentences by 1. However, the number of mistranslation, parts of speech, and word order slightly increased based on the results shown in Table 4.2. This can denote that while the errors increased for some sentences, there were also some sentences, which were completely fixed by the 101st incremental training. On the contrary, the result for the Filipino-to-English translation showed an obvious deterioration as the number of complete sentences decreased greatly while the number of mistranslation and word order errors increased. Also evident in Table 4.4, the most frequent occurring errors are mistranslation and parts of speech errors.

6.1 Quantity-Based Human Feedback

Another experiment was conducted in order to observe the effect of merging several feedback data of APE and treating them as a single feedback data. There were a total of 15 APE incremental training systems combined together consisting of the 1st to the 15th APE in a single APE incremental training. This is for better comparison against the 15th APE, which is the highest scoring APE for the English-to-Filipino. A single APE contains 1,000 sentences as their training data; hence, there are a total of 15,000 feedback sentences trained for English-to-Filipino in this experiment. On the contrary, there were 6 incremental trainings of APE combined together for Filipino to English, which consists of 6,000 sentences in total.

Table 4.3: English-to-Filipino Comparison of 15th and Merged APE

7.	BLEU	NIST	TER	METEOR
15th APE	0.3333	6.5343	0.4293	0.3218
Merged (1st to 15th APE)	0.2726	6.0478	0.4750	0.3036

Table 4.4: Filipino-to-English Comparison of 6th and Merged APE

8.	BLEU	NIST	TER	METEOR
6th APE	0.3795	7.4114	0.3995	0.3795
Merged (1st to 6th APE)	0.3610	7.2671	0.4100	0.3765

As a result, training the feedback data in smaller sets is still better than training them in larger sized data. The scores in Table 4.3 and Table 4.4 show that the 6th and 15th APE got higher precision scores for the four metrics compared to their merged counterpart. This denotes that doing incremental training allowed more word matches, clustered words, and lesser corrections needed to be done to match the reference sentences. With lesser amount of data, the system is able to learn better as the weighted sum of the probability values in the language model is taken. However, given that the feedback data is trained as a whole, the system will only take in the current probability value causing a poor translation quality. Applying interpolation and combination methods limits the probability increase or decrease of n-grams in the language model and preserves previous phrase pairs, which limits the amount of changes done to the translation.

8.1.1 Unique Quantity-Based Human Feedback

Table 4.5: English-to-Filipino Comparison of 8th and Merged APE

	BLEU	NIST	TER	METEOR
8th APE	0.3300	6.4999	0.4305	0.3195
Merged (1st to 8th APE)	0.3352	6.5512	0.4271	0.3227

Table 4.6: Filipino-to-English Comparison of 8th and Merged APE

	BLEU	NIST	TER	METEOR
8th APE	0.3743	7.3323	0.4049	0.3753
Merged (1st to 8th APE)	0.3892	7.5288	0.3903	0.3887

The previous 100 APE incremental training phases were trained between baseline development data that was first translated in the baseline and corresponding human post-edited feedback. However, the baseline development data contains duplicates that could result to repetitions in the training data of the APE phases. Since having more repetitions increases the probability values for both the language model and the phrase table, it is necessary to observe how unique sets of feedback will improve the translation when trained in the APE setting.

Of the total of 4,406 sentences in the gathered human feedback for both English and Filipino, there were a total of 4,111 unique English sentences and 4,174 unique Filipino sentences. In order to also investigate the effect of changing the size of human feedback for each incremental training iteration, the size of the training data for each iteration was changed from 1000 to 500 sentences. Eight sets of APE incremental training data were built. The goal of the experiment is to compare between the 8 APE incremental training phases and the merged APE, composed of the same 8 phases of the APE. In all other aspects, this experiment was similar to that of the Quantity-Based Human Feedback (Section 4.3). The merged APE in the previous experiment contained duplicates, which was a possible factor for the lower translation quality because duplicate entries increase the probabilities of wrong translation pairs. With these ~4K unique sentences, the merged APE can be analyzed without the factor of incorrect duplicate translation pairs.

In Table 4.5 and Table 4.6, for both English-to-Filipino and Filipino-to-English translations, the merged APE incremental training phase has better evaluation metric scores compared to the 8 separate APE incremental training phases. The main reason for the increase in score is its uniqueness. Since there were no duplicates, the APE phase was able to learn all sentences equally wherein it calculated a more accurate computation of the probabilities. The number of training data for an APE phase does not directly mean the decrease in translation quality.

Table 4.7: English-to-Filipino Unique Incremental APE Phases

9.	BLEU	NIST	TER	METEOR
1st APE	0.3316	6.5162	0.4317	0.3194
2nd APE	0.3282	6.4722	0.4347	0.3188
3rd APE	0.3239	6.4550	0.4350	0.3176
4th APE	0.3323	6.5203	0.4308	0.3215
5th APE	0.3313	6.152	0.4304	0.3212
6th APE	0.3323	6.5214	0.4301	0.3216
7th APE	0.3345	6.5402	0.4287	0.3220
8th APE	0.3352	6.5512	0.4271	0.3227

Table 4.8: Filipino-to-English Unique Incremental APE Phases

10.	BLEU	NIST	TER	METEOR
1st APE	0.3714	7.3506	0.4028	0.3798
2nd APE	0.3729	7.3610	0.4036	0.3785
3rd APE	0.3783	7.4151	0.3991	0.3821
4th APE	0.3816	7.4386	0.3971	0.3833
5th APE	0.3796	7.4249	0.3976	0.3827
6th APE	0.3804	7.4254	0.3968	0.3832
7th APE	0.3834	7.4539	0.3950	0.3850
8th APE	0.3892	7.5288	0.3903	0.3887

In addition, for a single unique APE incremental training phase, more unique sentences would entail better machine translation quality.

The automated evaluation scores are listed in Table 4.7 and Table 4.8. The general trends for BLEU and NIST for both English-to-Filipino and Filipino-to-English translations are shown in Figure 4.5 and Figure 4.6.

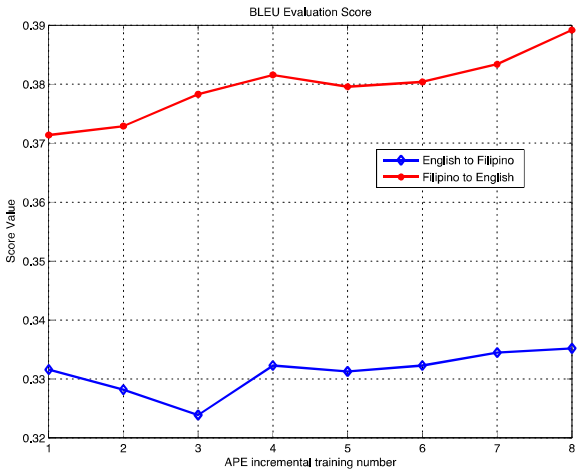

Figure 4.5: Comparison of BLEU Evaluation Score of Unique Incremental APE Phases

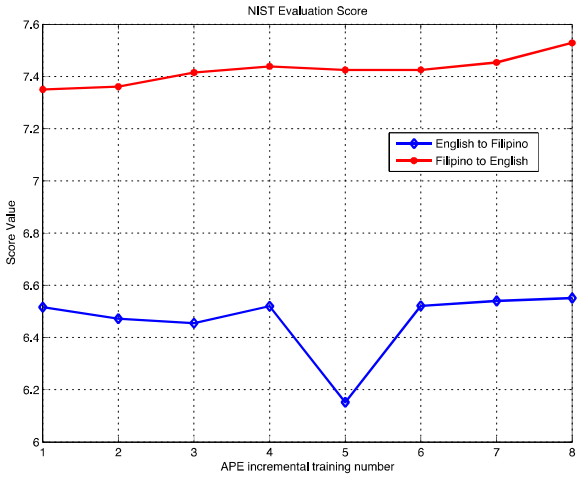

Figure 4.6: Comparison of NIST Evaluation Score of Unique Incremental APE Phases

The performance trends of the APE systems trained with unique feedback data, as shown in Figure 4.5 and Figure 4.6, are steadily increasing and are more stable. Although it appears that training with unique feedback data is better, the performance difference may be due to the way the test sentences were selected, which is purely random and did not consider the biases towards more frequently translated sentences or phrases captured using crowdsourced feedback, and on which the APE systems were incrementally trained.

11. CONCLUSION AND FUTURE WORK

In the implementation of FEBSMT, a feedback system was added to a statistical machine translation system to make updates more dynamic and responsive to quality human feedback. The four evaluation metrics namely, BLEU, NIST, METEOR, and TER were used. The system was implemented bi-directionally and both were iteratively run until convergence rates of translation scores are observed. The machine translation quality of the APEs at the onset is higher than their respective baseline evaluation scores. However, the evaluation scores soon reached its peak before decreasing gradually. This suggests that the feedback significantly affected the probability scores of the Language Model and the phrase tables, and thus affected the translations of the baseline system that are correct to begin with. It would, however, be interesting to empirically investigate the corresponding trends if the training and feedback data be made much larger by letting the system run in the long term. Furthermore, we observed that the Filipino- to-English translation has a higher machine translation quality overall, compared to the English-to-Filipino translation.

For the post-editing, the source of feedback may use the concept of crowdsourcing, wherein FEBSMT will be made available online for humans to use and provide feedback. There will be more users and the translation system will be tested thoroughly. Deploying the translation system will bring more sources of feedback and a better opportunity for the system to improve its translation. There can also be an added feature for verifying the sources of feedback and filtering out of the noisy feedback to avoid negative effects on the translation system.

12. REFERENCES

[1] Banerjee, S., Lavie, A.2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.

[2] Doddington, G. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Proceedings of the Human Language Technology Conference, California, 128-132.

[3] Koehn, P. 2014. MOSES. Statistical Machine Translation System, User Manual and Code Guide, Cambridge University Press.

[4] Nocon, N., Oco, N., Ilao, J., Roxas, R.2013. Philippine Component of the Network-based ASEAN Language Translation Public Service. 7th IEEE Conference Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management

[5] Papipeni, K., Roukos, S., Ward, T., and Zhu, W.J. 2002.BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, 311-318.

[6] Simard, M., and Foster, G.2013. PEPr: Post-Edit Propagation Using Phrase-based Statistical Machine Translation. Proceedings of the XIV Machine Translation Summit, Canada, 191-198.

[7] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.2006. A Study of Translation Edit Rate with Targeted Human Annotation.

[8] United Nations.nd. Millenium Development Goals And Beyond 2015. Goal 8: Develop A Global Partnership For Development. http://http://www.un.org/millenniumgoals/global.shtml