7th IEEE International Conference Humanoid, Nanotechnology, Information Technology
Communication and Control, Environment and Management (HNICEM)
The Institute of Electrical and Electronics Engineers Inc. (IEEE) – Philippine Section
12-16 November 2013 Hotel Centro, Puerto Princesa, Palawan, Philippines

Philippine Component of the Network-based ASEAN Language Translation Public Service

Nicco Nocon[1], Nathaniel Oco[2], Joel Ilao[1], Rachel Edita Roxas[2]
[1]De La Salle University
[2]National University
noconoccin@gmail.com, nathanoco@yahoo.com, joel.ilao@delasalle.ph, rachel_roxas2001@yahoo.com

Abstract— **Communication between different nations is essential. Languages which are foreign to another impose difficulty in understanding. For this problem to be resolved, options are limited to learning the language, having a dictionary as a guide, or making use of a translator. This paper discusses the development of ASEANMT-Phil, a phrase-based statistical machine translator, to be utilized as a tool beneficial for assisting ASEAN countries. The data used for training and testing came from Wikipedia articles comprising of 124,979 and 1,000 sentence pairs, respectively. ASEANMT-Phil was experimented on different settings producing the BLEU score of 32.71 for Filipino-English and 31.15 for English-Filipino. Future Directions for the translator includes the following: improvement of data through changing or adding the domain or size; implementing an additional approach; and utilizing a larger dictionary to the approach.**

Index-Terms—statistical machine translation, ASEAN integration, Moses, Network-based ASEAN Language Translation Public Service

I. INTRODUCTION

Interaction and communication among ASEAN countries are important, especially to tourists of each country. A tool such as a multi-lingual machine translator is beneficial for assisting ASEAN countries, making communication easier. ASEANMT-Phil, the Philippine component of the ASEAN Machine Translation Project, is a phrase-based statistical machine translator developed to connect various languages within the ASEAN countries, through the use of the English language as pivot. ASEANMT-Phil's main task is to translate words from Filipino to English and vice-versa. To perform its task, ASEANMT-Phil utilizes Moses, a statistical machine translation engine capable of translating a given language pair by automatically generating translation models through training, for which both training and testing used Wikipedia articles as data. Throughout this paper, the development of ASEANMT-Phil is discussed, given the following sections: related works, ASEANMT-Phil's design, results and discussions, and conclusion and future directions.

II. RELATED WORK

In building a translator for Filipino-English and vice-versa, different approaches can be applied for it to be accomplished. Approaches include: LFG-based [1], Template-based [5], Rule-based [9], [10], Statistical Machine Translation or SMT [3], [8], and many more. Each of the approaches' highest scoring performances based on BLEU [6] scores is displayed on TABLE I. The statistical machine translation approach performed best among other works and having said that, for ASEANMT-Phil, we chose to use the same approach. Additionally, not only approaches differ in creating translators. Resources used can also vary, instances include choice of data such as movie subtitles and articles [4], which can be in different languages such as Indonesian, Vietnamese [7], European [3], English and Filipino [1], [5], [10], and more.

TABLE I. BLEU PERFORMANCE SCORES

Approach	Language Corpus	BLEU
LFG-based	English and Filipino	-
Template-based	English and Filipino	-
Simard et al.'s Rule-based	Europarl	31.11
Tan et al.s's Rule-based	English and Filipino	05.73
Koehn et al.'s SMT	Europarl	35.54
Schlippe et al.'s SMT	French (Text Normalization)	96.00
Larasati's SMT	English and Indonesian	30.14

III. ASEANMT-PHIL

ASEANMT-Phil is a phrase-based statistical machine translator which utilizes Moses for its training and testing. Data used were articles gathered from Wikipedia, comprising of 124,979 sentence pairs for training and 1,000 for testing. It was developed through the processes seen in Figure 1.

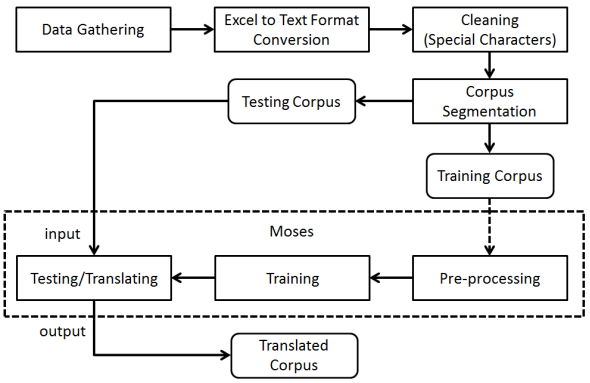

Figure 1. ASEANMT-Phil Framework

Several experiments were done to acquire the best Moses setting for the corpus to have a high BLEU score while retaining quality of the translations. Before each experiment, the training corpus was run through three Perl scripts which performs tokenizing, true-casing, and cleaning the corpus. In tokenization, spaces are inserted between words and punctuations. In true-casing, words in the corpus are transformed into their most likely cases. In cleaning, the sentences with lengths outside the given range are deleted.

The settings experimented on are divided into two types: the training (T) settings and decoder/translation (D) settings. The original setting for training is as follows: language model (LM) order of 3, max phrase length of 7, and cleaning range of 1-80; while for the decoder is distortion limit (DL) 6 and max phrase length of 20. The changes implemented on each experiment are shown in TABLE II. Additional experiments were conducted through post-processing: changing the testing data and system's output into lowercase, and utilization of 7,232 word entry dictionary. In addition to the processing, the corpora were de-tokenized, converting special tokens (e.g. & and ') into their corresponding characters.

IV. Results and Discussion

In deciding the setting for ASEANMT-Phil, the best for each training and decoder setting was combined to form the overall best. Comparing the experiments' scores (see TABLE III. for the BLEU scores) for training, Experiment 4 has the highest score with maximum phrase length 10 as the setting. However, this setting stores a lot in the memory with little difference to the original's score. Because of this, it has been decided to set the training setting to default. On the other hand, Experiment 6 has the highest among the decoder settings. Unfortunately, it exposes an issue with the setting of zero for distortion limit. No distortion limit means there is no reordering in translations, which is unsuitable to Filipino-English or vice-versa translations because the two languages do not have the same grammar rules. For this reason, the next which is Experiment 9 was chosen and applied to the original. Performing the post-processing, the score was increased after lowercasing and lowered on utilization of the dictionary, making the lowercased Experiment 9 the final setting with the BLEU of 32.71 for ASEANMT-Phil's Filipino-English, matching the previous choice Experiment 6. Also performing the same setting into its inverse English-Filipino (see TABLE IV. for the BLEU scores), the BLEU score of ASEANMT-Phil English-Filipino is 31.15.

TABLE II. List of Experiments Conducted

Experiment	Changes
1	LM order: 7
2	(T) Max phrase length: 5
3	(T) Max phrase length: 3
4	(T) Max phrase length: 10
5	Cleaning range: 1-40
6	DL: 0
7	DL: 3
8	DL: -1 (unlimited)
9	Added command: -monotone-at-punctuation or -mp
10	Added command: -minimum-bayes-risk or -mbr
11	(D) Max phrase length: 10
12	(D) Max phrase length: 7
13	(D) Max phrase length: 5
14	(D) Max phrase length: 3

TABLE III. BLEU Score Performance (Filipino to English)

Label	BLEU
Baseline	12.74
Original	32.02
Experiment 1	32.01
Experiment 2	31.97
Experiment 3	31.73
Experiment 4	32.04
Experiment 5	31.13
Experiment 6	32.71
Experiment 7	32.43
Experiment 8	29.18
Experiment 9	32.54
Experiment 10	32.13
Experiment 11	32.02
Experiment 12	32.02
Experiment 13	31.99
Experiment 14	31.71
Applied Experiment 9	32.54
Lowercase	32.71
Dictionary	32.48
ASEANMT-Phil F-E	32.71

TABLE IV. BLEU SCORE PERFORMANCE (ENGLISH TO FILIPINO)

Label	BLEU
Original	30.65
Applied Experiment 9	30.98
Lowercase	31.15
Dictionary	25.19
ASEANMT-Phil E-F	31.15

The performance of ASEANMT-Phil is a third of a hundred with more improvements needed, even though mostly were already solved. Improvements start in detecting the faults of the system, which can be seen on the data, approach, and even some on the tool used. On the data, issues were found at both training and testing data – with inclusion to the system's output.

A. *Training Data*

In the training data, there were several special characters (e.g. é, á, ü, ™, €), where some were detected by the tokenization.perl script, stopping the training. However, without deleting or replacing these characters, the translations may be affected, that the special character is needed to be with a certain phrase before it can be used as a translation. So, through the use of regular expressions, these special characters were deleted and replaced, solving future issues in tokenization and translation. Another issue was there were numerous misaligned sentences present on the data caused by the consolidation of text files. When the cleaning script was used, after detecting the first occurrence of the misalignment, the rest of the entries also suffered. To resolve it, pairs were tracked down and deleted from the data set to prevent training errors and to preserve other aligned pairs.

B. *Testing Data*

Regarding the testing data, issues are concerning the performed translation or training. Casing, spaces, out-of-vocabulary (OOV) words, function words, over-translations, faulty reordering, and un-matched choices of words between human and system translations are factors where both entries differ on several cases.

Seen on post-processing are the transformation of both texts through the words' cases, and the results imply that two words are only determined as the same if they have the exact same cases, striking the sense of accuracy.

Another issue found is the OOVs, words not found by the translator in the phrase-table. This issue can be tracked back to the training data, where it lacks the certain words to be placed in the phrase-table. The effects of these OOVs are displayed on the system's output where some words are retained instead of getting translated, resulting sentences to be in a form of code-switching. OOVs can also exist in the output even though a variation of a word is inside the phrase table. For instance is the word, *tutulog* 'will sleep', even though there is an entry of *tulog* 'sleep' in the phrase table, it could still be unknown to the translator. With the high number of training data, the OOVs in the system's translations are still high, which means either the data is still not enough or the data is not fit for a higher accuracy translation.

Function words such as *the*, *both*, *which*, *some*, etc., can be misused in translations for both English-Filipino and vice-versa, especially on the words found at the beginning of the sentence. Usually *the* can be a start for a sentence, however if the reference does not start with *the* and the candidate (system's output) translated *the*, it would take a toll on the system's accuracy. Probable cause for the issue is that sentences in the testing data are connected unlike the translations which each line is independent or considered always as a new one. Example provided is the sentence, "The this classification based on…" is the translated output instead of the ideal, "this classification based on…" or a different although plausible example "The guqin is a very quiet instrument" and "guqin is a very quiet instrument". These examples, show slight difference to both the reference and the candidate, but because of this, the system's accuracy drops.

"basis of *kasiyahan*. the" 'basis of happiness. the' is an example of faulty reordering. It is caused by the default reordering or distortion limit with the value of 6. Meaning, the word chosen within a phrase can skip or move at the most of 6 words. Reducing the limit caused not only the accuracy to improve, but also the time consumed for translations. Though, as stated before, it is not ideal in bilingual translations. But there is a secondary solution for this which is Experiment 9. In experiment 9, monotone-at-punctuation stops the usage of reordering only when there is punctuation. With this setting, reordering will not go out of hand especially when the character is the ending punctuation.

On un-matched choices of words, it is between the difference of words used by the human translator (reference) and the generated translations made by the system (candidate). Instances of which either the human or the system translator chose the given words are shown in the following: words like *numerous* and *several*; *inaccurate* and *not accurate*; *pre-history* and *early history*; *pagkabuo* can either be translated as *creation* or *formation*; *pangunahing* for either *basic* or *major*; and more. This issue is also the result of the Filipino language's complexity. There are numerous words where a translation for a Filipino word can be equivalent to several variations. Although this issue is one of the reasons of why the accuracy of ASEANMT-Phil is low, we still analyzed the sentences and chose the one with the most properly structured setting or potential.

C. *Dictionary*

The dictionary when applied to ASEANMT-Phil was intended to decrease the OOVs and increase the BLEU score. However, it turned out to degrade ASEANMT-Phil's performance. The reason is related to interlingual homographs, where a word exists in both languages such as *at*, *may* and *man*. With this, the interlingual homographs create a conflict which results the dictionary into making mistakes in

translations, generating *and*, *have*, and *also*, respectively. In addition, the dictionary did not have much of an effect due to the lack of entries wherein it did not cover the different tenses of the words present in the corpus and words were too deep to be within the dictionary's scope. An example is the word *pahinga* 'rest' where it is present in the dictionary, but its variant *magpapahinga* 'will rest' is nonexistent. Branching from the same example, another issue is that the dictionary is only word-based. When searching for translations, the phrase "will rest" will only be searched separately, meaning one for *will* and another for *rest*.

Although not all of the issues are in the dictionary, the testing corpus also has its issues only connected to the dictionary. Majority of these are regarding accidental misspellings and again some which are too deep for the dictionary to result into an OOV. Samples like *gayupaman* instead of *gayunpaman* 'however' and *bagongriles* instead of "*bagong riles*" 'new rails/tracks' are the translator's accidental word errors. Words such as *letrato* instead of *litrato* 'photo' are another sample resulted from the different styles of writers due to Filipino's complexity.

D. Others

Aside from the given issues, there are others that are in need to be pointed out such as the time and memory consumption of training and translations. For a corpus with around 130k sentence pairs, time was consumed at an approximate of 3 to 6 hours. The solution for the time consumption was to utilize a setting in Moses where multiple cores could be used to perform the training faster. On the other hand, the memory consumption of translations for the same number of sentence pairs, 2GB of RAM was not enough. Changing workstations with 4GB of RAM was the action we took to accommodate the translations needed.

V. CONCLUSION

ASEANMT-Phil, a bi-directional English-Filipino phrase-based statistical machine translator was built using Wikipedia articles, and had gone through pre-processing, experimental, and post-processing stages to come up with the most highest BLEU scores while maintaining the quality of the translations. For English-Filipino, the translator performed with 32.71 BLEU, while Filipino-English got 31.15. Several issues were encountered and some were resolved. Issues left are the following: OOVs and un-matching words between the human and the system's outputs. Even though ASEANMT-Phil has 124,979 training pairs the BLEU score is still a third of BLEU's highest possible score. For future directions, the increase of training data, a switch or addition of domains of training data, implementing another approach as pre- or post-processing, and utilizing a larger dictionary to the approach are recommended.

ACKNOWLEDGMENT

ASEANMT-Phil is supported in part by the Commission on Higher Education through the funded program "AseanMT-Phil (the Philippine Component of the Asean Machine Translation Project): A Hybrid Bi-directional English-Filipino Statistical Machine Translation System".

REFERENCES

[1] Borra, A., Chan, E., Lim, C., Tan, R., & Tong, M. (2007). LFG-Based Machine Translation Engine for English and Filipino. In Proceedings of the 4th National Natural Language Processing Research Symposium.

[2] Koehn, P. (2003). EUROPARL: A Parallel Corpus for Statistical Machine Translation. In MT Summit.

[3] Koehn, P., Och, F., & Marcu, D. (2003). Statistical Phrase-based Translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association of Computational Linguistics on Human Language Technology.

[4] Larasati, S. (2012). Handling Indonesian Clitics: A Dataset Comparison for an Indonesian-English Statistical Machine Translation System. In Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation.

[5] Ong, E., Go, K., Nuñez, V., Morga, M., & Veto, F. (2007). Template-Based English-Filipino Machine Translation System. In Proceedings of the 4th National Natural Language Processing Research Symposium.

[6] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

[7] Pham, L., Tran, V., & Nguyen, V. (2013). Vietnamese Text Accent Restoration with Statistical Machine Translation. In Proceedings of the The 27th Pacific Asia Conference on Language, Information and Computation.

[8] Schlippe, T., Zhu, C., Gebhardt, J., & Schultz, T. (2010). Text Normalization based on Statistical Machine Translation and Internet User Support. In INTERSPEECH.

[9] Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (2007). Rule-based Translation with Statistical Phrase-based Post-editing. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.

[10] Tan, M., Hong, B., Alcantara, D., Perez, A., & Tan. L. (2006). Learning Translation Rules for a Bidirectional English-Filipino Machine Translator . In Proceedings of the The 20th Pacific Asia Conference on Language, Information and Computation.