Using Statistical Machine Translation for Style and Grammar Correction

Manolito Octaviano Jr
De La Salle University
2401 Taft Avenue
Manila, Philippines
manolito.octaviano@yahoo.com

Nathaniel Oco
National University
551 M.F. Jhocson St.
Sampaloc, Manila, Philippines
nathanoco@yahoo.com

Allan Borra
De La Salle University
2401 Taft Avenue
Manila, Philippines
allan.borra@dlsu.edu.ph

ABSTRACT
This paper presents the use of SMT approach for style and grammar checking. The principle behind the approach is to surface the common mistakes made by native speakers. Translation documents of English to Filipino were collected containing errors that were marked and corrected. Filipino sentences were encoded to produce a parallel corpus of incorrect and correct text. Then, parallel corpus was pre-processed and used as training data of SMT system to generate phrase table rules. Generated phrase table rules were manually analyzed to create rules. A total of 56 rules were generated. Finally, rules were integrated in LanguageTool - a rule-based style and grammar checker engine. The engine was tested using the default Tagalog resource of LanguageTool and yielded an accuracy of 78%. It was observed that the limited training data and the generated phrase table rules which is represented in tokens are the factors for the small number of generalized rules. For future works, a larger amount of training data will be added to describe more mistakes. Furthermore, part-of-speech tags will be introduced to generate more concrete rules.

Keywords
Machine translation, Statistical machine translation, Parallel corpus, Filipino, Style and Grammar correction

1. INTRODUCTION
The Philippine is an archipelago known to have 7, 107 islands with over 180 distinct languages. Among these languages, Tagalog is the basis of the Filipino language with 52 million native speakers[1]. It displays complex morphological structure [3] and was labeled as a language with "high degree of inflection" [4]. Due to the complex structure of the language, even native speakers are having a hard time in writing Filipino that follows the proper style and grammar rules. For instance, the use of the word *raw* and *daw* in sentence or the proper affixation like the prefix /pag-/ in the word *iingat* (e.g., *pagiingat* vs *pag-iingat* vs *pag iingat*). Taking these into account, there is a need to address this language issue.

Grammar checkers can be used to check sentence inconsistencies and variations. These are programs that determines syntactic correctness of a sentence by detecting sentence inconsistency and provides suggestion to address errors. Approaches of grammar checker are classified into three: syntax-based, statistics-based, and rule-based [11]. In syntax-based approach, texts are parsed and an input sentence is considered erroneous if the parsing failed. The statistics-based approach utilizes annotated corpus of part-of-speech and considers erroneous sentences based on computed probability scores. The rule-based approach uses a set of rules which matches against an input text which has at least been part-of-speech tagged. In the Philippines, several style and grammar checkers were developed such as extension for OpenOffice Writer [4], Panuring Pampanitikan [6], and Tagalog support of LanguageTool [12].

Given a limited linguistic resources, the use of the rule-based approach is more feasible. However, the rules are developed manually which is time-consuming and some rules are not the common mistakes made by native speakers. An alternative technique to develop rule faster is through statistical process. Generated rules from it can be added to the existing base of rules.

In this paper, we present the use of statistical machine translation (SMT) for generating rules for commonly committed mistakes made by native speakers. The idea is to learn the mistakes made by applying the SMT concept to a parallel corpus containing pairs of incorrect and correct texts.

This paper is organized as follows: Section 2 discusses the Filipino language, Section 3 gives related literature of the study, Section 4 describes the approach and resources utilized in the study, Section 5 reports the experiment and results, and Section 6 discusses the conclusion and future works.

2. FILIPINO LANGUAGE
The Filipino language exhibits different linguistic phenomena such as free-word order, code switching, complex morphological structure, and how words are spelled.

One linguistic feature of the Filipino language is the free word order of its sentence construction. Ramos classified sentence construction into *pagpapanaguri* and *pagtitiyak [14]*. Sentences that is in the form of *pagpapanaguri* follows the predicate-subject construction while *pagtitiyak* follows the subject-predicate and a lexical marker "ay" is usually present. For instance, the sentence "The kid grabbed a food from the store" can be translated while maintaining the context into several Filipino sentences as shown in Table 1.

[1] The 2010 Philippine Census data is taken from: Philippine Census, 2010. Table 11. Household Population by Ethnicity, Sex, and Region: 2013.

Philippine Computing Journal Dedicated Issue on Natural Language Processing, pages 31–36
Vol. XI No. 1 August 2016

Table 1. Sentence construction in Filipino

Pagtitiyak	
Sentence 1:	Ang bata ay kumuha ng pagkain sa tindahan.
Pagpapanaguri	
Sentence 1:	Kumuha ang bata ng pagkain sa tindahan.
Sentence 2:	Kumuha ng pagkain ang bata sa tindahan.
Sentence 3:	Kumuha ng pagkain sa tindahan ang bata.

The presence of code-switching in sentence construction is also a linguistic feature of the Filipino language. Myers-Scotton and Ury defined code-switching as "the use of two or more linguistics varieties in the same interaction or conversation" [15]. The code-switching (CS) in the nation can be classified into (1) intra-sentential CS and (2) intra-word CS. The intra-sentential CS is the interchanging words and clauses between Filipino and English language. On the other hand, intra-word CS is the use of Filipino affixes and morphological rules to an English word. Table 2 shows example of the code-switching in Filipino.

Table 2. Types of code-switching in the Philippines

Type	Example	Translation
Intra-sentential	Ang cool nila.	They are cool.
Intra-word	Mag-drive	To drive

Aside from the linguistic features of Filipino in sentence construction, the complex morphological phenomena of the language are remarkable. These features are affixes, partial reduplication, and full reduplication.

Filipino words may undergo different types of affixes: prefixation, a word may have added one syllable like /mag-/ or as many as 7 syllables like /ikinapagpapaka-/; suffixation, a word may have attached one from the four defined suffixes of the language which is /-an/ or /-in/ that normally attached to words that ends with a vowel and /-han/, or /-hin/, that usually added to consonant-ending words. infixation, a word may undergo either from the two defined infix of the language, /-in/ or /-um-/; circumfixation, combinations of prefix, infix, and suffix. In addition, phoneme change of /r/ and /d/ in prefixation while /o/ and/u/ in suffixation may occur in certain words of the language.

The reduplication in Filipino word may be either partial wherein the part of the word stem is used to form a new word or full in which the entire word stem is repeated. Usually, full reduplication contains duplicated syllables, affixes, and sometimes hyphenation at the same time. Table 3 shows the different forms of the root word "sagot".

Table 3. Different forms of word "sagot"

Morphological phenomena	Word
Prefix	Pagsagot
	Nagsagot
	Nasagot
Suffix	Sagutan
	Sagutin
Infix	Sinagot
	Sumagot
Circumfix	Pinagsagutan
	Nagsagutan
Partial reduplication	Sasagutan
Full reduplication	Sagut-sagutan

Another linguistic phenomenon is the rule that Filipino used in spelling out words "*Kung ano ang bigkas, siyang sulat"* [13]. This rule is usually applied in transforming English loanwords to its corresponding Filipinized version. Table 4 shows example of Filipinized form of English words.

Table 4. Filipinized words

English loanwords	Filipinized version
Company	Kumpanya
Scholarship	Iskolarship
Computer	Kompyuter
Record	Record

3. RELATED LITERATURE

3.1 Rule-based Grammar Checker

The rule-based approach of grammar checker offers advantages as compare to syntax-based and statistics-based approach. The problem with the syntax-based approach is the required comprehensive grammar which covers all types of texts one wants to check [11]. On the other hand, statistics-based approach requires flexible corpus in its training process [12]. With the rule-based approach, rules can be built incrementally and modified according to the needs of the user. Thus, created rules can easily be tested. In addition, the approach can precisely locate the inconsistency in the sentence and provide the corresponding correction.

The rule-based approach needs to create enormously large number of rules in order to cover more errors. However, rules are developed manually which is time-consuming and costly. A study [9] described different approaches of rule-creation by means of automatic and semi-automatic development. It was noted the use of machine learning algorithms to acquire rules such as the use of statistical machine translation.

3.2 Statistical Machine Translation

Machine translation is the process of computer-aided translation from one language to another. One of the approaches of translation is done through statistical process called statistical machine translation (SMT). It uses bilingual text or parallel corpus to learn the language patterns. Equation 1 shows the well-known equation of SMT [2].

Equation 1. Statistical machine translation equation

$$e = \arg\max P(e \mid f) = \arg\max P(e)P(f \mid e)$$

In the context of translating incorrect to correct form, *e* represents the target language sentence (correct) and *f* represents the source language sentence (incorrect). The *P(e)* represents the language model probability of target language corpus (English language model) while the *P(f|e)* represents the translation model probability of the sentence-aligned parallel corpus.

Two of the translation models can be used are word-based and phrase-based model. The word-based model generates word translation and source word that can be map to multiple target words while phrase-based model generates translation of sequences of words. A study [8] explains why phrase-based model outperforms the word-based model.

There are different works applying SMT for correcting error. The study of Brockett [1] utilized the SMT technique in correcting countable/uncountable nouns, a POS which confuses people learning English as a second language (ESL). Using artificial errors in sentences as their training data, the SMT approach was able to beat the Microsoft Word 2003 grammar checker even though it produced a higher inaccurate number of corrections. Similar approach was applied to correct Japanese learners using revision logs of a language learning website as data [10]. They claimed that segmenting sentences into character-wise model outperforms the word-wise model. A work of Ehsan and Faili [5] wherein SMT and rule-based approaches in correcting grammars and spellings were compared, showed that both are complementary to each other, resulting into a hybrid system which could achieve better results for correction.

4. METHODOLOGY

The methodology of the research is shown in Figure 1. Student submission of Wikipedia translation were used as training data of an SMT system. After the training process, it produces a phrase table rules that were analyzed to surface the frequent mistakes made by students. Identified mistakes were generalized to develop rules for a rule-based style and grammar checker. Then, the grammar checker was tested.

4.1 Data Collection

Table 5. Encoded Text

Source text	Translated text
mga tinipon na gawa	koleksyon na gawa
Tinangal	Tinanggal
nanalo siya sa higit sa 100 na larong singles.	nanalo siya nang higit sa 100 na larong singles.
kasing haba	kasing-haba
Painiwala	Paniniwala

Translations documents of English to Filipino Wikipedia submitted by students were collected from Filipino department of De La Salle University (DLSU). There are Filipino sentences containing errors that were marked and corrected by faculty members such as spelling and grammar error. These were encoded to produce a parallel corpus of Filipino incorrect and correct text. There are sentences that were marked but do not contain correction. These sentences are in the form of *pagtitiyak*. However, the style deviation of the university in translation only considered the *pagpapanaguri*. In addition, this is the sentence construction of the Tagalog that used in communication. Thus, these were encoded in the form that do not use the lexical marker "ay":

Sentence: Ang bata ay kumuha ng pagkain sa tindahan.

Encoded: Kumuha ang bata ng pagkain sa tindahan.

A total of 2,889 pairs of incorrect and correct text documents were encoded to produce a corpus. Table 5 shows a sample encoded documents from the parallel corpus. The encoded corpus is used as training data in identifying the common mistakes. The source text column refers to the encoded text that is erroneous while the translated text column refers to the corresponding correct form.

4.2 SMT Application

The Moses SMT system created by the group of Koehn [7] is utilized. It is an open-source toolkit that implements statistical approach to machine translation. The encoded parallel corpus of incorrect and correct is used as training data of Moses.

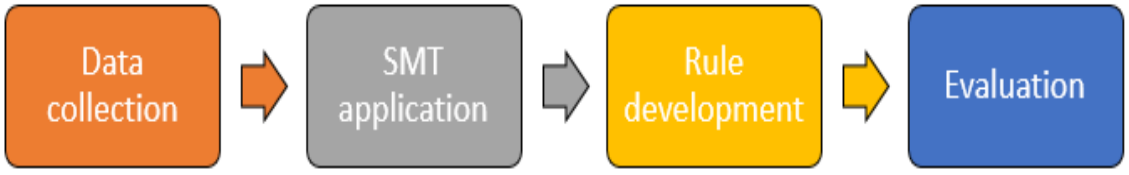

Figure 1. Rule creation process of the system

First, the parallel corpus was pre-processed using the built-in Moses command scripts *tokenizer.perl* and *clear-corpus-n.perl*. The *tokinizer.perl* is used to separate words and punctuations by injecting spaces in the middle. On the other hand, *clear-corpus-n.perl* is used to remove long or empty texts in the corpus. Then, phrase table rules were generated from the training process using the default setting with Giza++ for word alignment and SRILM Toolkit for language modeling in training process. Table 6 shows a subset of generated phrase table rule. The *phrase table scores* refer to the following:

- inverse phrase translation probability *f(f|e)*
- inverse lexical weighting *lex(f|e)*
- direct phrase translation probability *f(e|f)*
- direct lexical weighting *lex(e|f)*

The scores of phrase translation probability are the phrase-to-phrase probability model while lexical weighting operates the word alignments within pairs of phrase. For the *alignment* column, it shows the word alignment between source text and translated text.

Table 6. Sample phrase table rule

Source	Translated	Phrase table scores	Alignment
lengguwaha	lengguwahe	1 1 1 1	0-0
maket	merkado	1 0.5 1 0.5	0-0
may impeksyon ay	may impeksyon na	1 0.5 1 0.05	0-0 1-1 2-2

4.3 Rule Development

The generated phrase table rules were manually analyzed to learn the frequent mistakes. After mistakes were identified, these were generalized to create rules and apply to the LanguageTool - an open-source style and grammar checker [11]. The LanguageTool uses rules stored in XML file containing the error patterns that is used to identify the errors in a sentence and give the suggested correction. These patterns can be represented in terms of tokens, regular expression or part-of-speech tag. Listing 1 shows a sample rule file that displays the 3 basic elements of each rule: pattern to be matched, messages or suggestion, and example. If an input sentence matches a declared pattern, it will notify the user and give the suggested correction.

```
<pattern case_sensitive="no" mark_from="0">
    <token> computer </token>
</pattern>
<message>
    <suggestion> kompyuter </suggestion> ba ng iyong nais?
</message>
<short> typographical error </short>
<example correction=" kompyuter " type="incorrect">
    Gumamit sila ng <marker> computer </marker> kanina.
</example>
<example type="correct">
    Gumamit sila ng <marker> kompyuter </marker> kanina.
</example>
```

Listing 1. Sample rule file

5. RESULTS AND DISCUSSION

5.1 Test Sentences

The system was tested using the sentences from the default Tagalog resource of LanguageTool [12]. The test sentences were assessed by an expert in the language and classified 44 error-free and 39 erroneous sentences. These sentences were categorized into 7 as shown in Table 7.

Table 7. Categories of test sentences

Category	Incorrect (*input*)	Correct
Correct	-	Sa DLSU rin ako nag-aral.
Style deviation	Siya ay isang madasaling bata.	Isang madasaling bata siya.
Number agreement	Magaganda akong bata.	Maganda akong bata.
Word repetition	Pinalakad ng ng abogado si Maria.	Pinalakad ng abogado si Maria.
Affix usage	Nagbunot ng ngipin ang dentist	Bumunot ng ngipin ang dentista.
Ligature usage	Isa siyang mababa na bata.	Isa siyang mababang bata..
Sound and letter change	Tinalo ka daw ng kaklase mo.	Tinalo ka raw ng kaklase mo.

5.2 Implementation in LanguageTool

A total of 56 rules were generated. These rules can be categorized into 4 type as shown in Table 8. Then, rules were implemented in LanguageTool and test sentences were fed.

Table 8. Categories of generalized rules

Category	Sentence	Error	Correction
Sound and letter change	Tinalo ka daw ng kaklase mo.	daw	raw
Affix usage	Siya ang pinaka matabang bata.	pinaka matabang	pinakamatabang
Typographical error	Gumamit ang bata ng computer.	computer	kompyuter
Style deviation	Tayo ay kumain.	ay	---

The LanguageTool detects and suggests a correction only when a sentence has style or grammar error. A total of 21 out of the 39 incorrect sentences were properly detected. A total of 44 out of the 44 error-free sentences were properly detected. The accuracy of the system scored 78%. Figure 2 shows an image of the LanguageTool simulating detection and suggestion on a sentence.

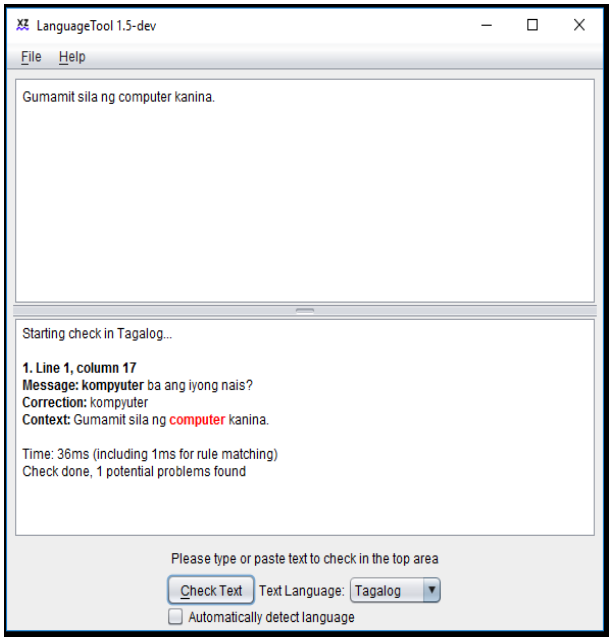

Figure 2. Screenshoot of LanguageTool

Analysis shows that the limited training data is a factor the small number of generalized rules. The SMT system is dependent on the size of training data to produce a phrase table rules that will be used for generalizing rules. Therefore, a larger corpus is needed to describe more mistakes.

Another observed factor is the generated phrase table rules. The SMT system generates the phrase table rules based on the processed training data, which are currently represented in tokens and not part-of-speech tags. Thus, creating rules that utilized tags is not possible because phrase table rules are plainly tokens. This is the same reason why the system was unable to give suggestion to sentences with error style. Utilizing tags should be considered in generating rules that tokens is not applicable.

Lastly, it was observed that the rule for *sound and letter change* category should further be improved. For instance, the sentence "Ikaw daw ang panalo." was marked as correct by the system. However, the letter /w/ of the word *ikaw* sounds a vowel. The LanguageTool must detect and give suggestion to it – "Ikaw raw ang panalo".

6. CONCLUSION

This paper presents the use of SMT approach as an alternative technique to generate rules for a grammar checker. Parallel corpus from the DLSU's Filipino department were encoded and used as training data of SMT system to generate phrase table rules. Then, the phrase table rules were manually analyzed to produce rules. A total of 56 rules were generated. Finally, these were integrated in LanguageTool. The engine was tested using the default Tagalog resource of LanguageTool and yielded an accuracy of 78%. It was observed that the limited training data and the generated phrase table rules which is represented in tokens are the factors for the small number of generalized rules. For future works, a larger amount of training data will be added to describe more mistakes. Furthermore, part-of-speech tags will be introduced to generate more concrete rules and improved existing rules.

7. ACKNOWLEDGMENTS

This research work is supported by the Department of Science and Technology, Philippines as part of the "Interdisciplinary Signal Processing for Pinoys: Software Applications for Education (ISIP:SAFE)" program. We thank Ethel Ong, Nicco Nocon, Matthew Go, Alena Sipalay, Jose Francisco Tabia, Angelo Pastoriza, and Camille Billones for being instrumental in this research.

8. REFERENCES

[1] Brockett, C., Dolan, W. B., & Gamon, M. (2006, July). Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 249-256). Association for Computational Linguistics.

[2] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation.*Computational linguistics*, *19*(2), 263-311.

[3] Cheng, C., Roxas, R., Borra, A. B., Lim, N. R. L., Ong, E. C., & See, S. L. (2008). e-Wika: Digitalization of Philippine Language. In *DLSU-Osaka Workshop*.

[4] Dimalen, E. D., & Dimalen, D. M. D. (2007). An OpenOffice Spelling and Grammar Checker Add-in Using an Open Source External Engine as Resource Manager and Parser. In *OpenOffice. org Conference, Barcelona*.

[5] Ehsan, N., & Faili, H. (2013). Grammatical and context-sensitive error correction using a statistical machine translation framework. *Software: Practice and Experience*, *43*(2), 187-206.

[6] Jasa, M., Palisoc, J, & Villa, M. 2007. Panuring Pampanitikan (PanPam): A Sentence Syntax and Semantic Based Grammar Checker for Filipino. Undergraduate Thesis. De La Salle University, Manila.

[7] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... & Dyer, C. (2007, June). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions* (pp. 177-180). Association for Computational Linguistics.

[8] Koehn, P., Och, F. J., & Marcu, D. (2003, May). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 48-54). Association for Computational Linguistics.

[9] Miłkowski, M. (2012). Automating rule generation for grammar checkers.*arXiv preprint arXiv:1211.6887*.

[10] Mizumoto, T., Komachi, M., Nagata, M., & Matsumoto, Y. (2011, November). Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *IJCNLP* (pp. 147-155).

[11] Naber, D. (2003). A rule-based style and grammar checker.

[12] Oco, N., & Borra, A. (2011). A grammar checker for Tagalog using LanguageTool. *Asian Language Resources collocated with IJCNLP 2011*, 2.

[13] Ortograpiyang Pambansa. 2013. Komisyon sa Wikang Pilipino.

[14] Ramos, T. 1971. Makabagong Bararila ng Pilipino. Rex Book Store.

[15] Scotton, C. M., & Ury, W. (1977). Bilingual strategies: The social functions of code-switching. *International Journal of the sociology of language*,*1977*(13), 5-20.