Chapter 1

INTRODUCTION

1.1 Background

The rapid dissemination of electronic communication devices such as e-mails, Short Messaging Systems (SMS), chatrooms, instant messaging programs and blogs has triggered the emergence of a new form of written texts called Chatspeak or Texting Language (TL). In this context, users have the tendency of using a non-standard form of language that disregards grammar, punctuation and spelling rules. With this condition, many people have questioned the future of literacy and the semantic ambiguity brought by this compressed non-standard form of language.

To effectively process this form of language, it is necessary to develop a robust language processing tool capable of bearing with the extreme form of "noise" they contain. Recovering a normalized text seems thus to be a necessary preprocessing step for many real-world Natural Language Processing (NLP) applications, such as *text-to-speech systems* that read out webpages, emails and blogs to the visually challenged people, *translation* which processes the conversion of sentences in one source language to another language and *text mining applications* such as *filtering*, *routing*, and *information retrieval*. In addition, since non-standard spellings and grammatical usage is now very common over the web, a *search engine for noisy text* (say blogs, chatlogs or emails) must be immune to the several texting language variants of a word. Thus, a decoder from texting language to standard language would be very beneficial in applications like *search engine tools* and *automatic correction tools.*

There are two main approaches in processing text normalization. The Rule based approach and the Statistical Machine Translation (SMT) approach. The Rule based approach is

the most commonly used method in normalizing text. It uses straight dictionary substitution, with no language model or any other procedure to help them disambiguate between possible words substitutions. A study of Raghunathan and Krawczyk (2008) proves that dictionary approach performs worse due to the fact that it cannot disambiguate between possible translation candidates for a given source language word. The translation of such system merely depends on the static mapping of the texting language word to its equivalent standard form.

The Statistical Machine Translation (SMT) approach on the other hand makes use of a language model and other procedures that enable them to disambiguate between possible substitutions of a word. SMT therefore accommodates the limitations of a straight dictionary approach.

Studies regarding text normalization using SMT for English, Chinese and French have been conducted. However, no study has ever been made for the text normalization of the Tagalog language. With these reasons, the proponents pursued this study in the hope that it would open possibilities for more research in the language and more importantly in its normalization using SMT

1.2 Problem Statement

The study sought to investigate the characteristics of the Tagalog texting language, use the Statistical Machine Translation approach to its normalization and integrate the translation to an IM client. Moreover, the study sought to answer the following questions:

1. What is the nature and type of compressions used in the Tagalog texting language?
2. What are the similarities and differences between the Tagalog texting language and other texting languages?

3. What model is suitable for the Tagalog texting language?
4. How can the Tagalog texting language be normalized using the SMT system?
5. How does the Tagalog texting language be normalized in an open-source IM client?

1.3 Objectives

The study intended to investigate the characteristics of The Tagalog texting language, use the Statistical Machine Translation approach to its normalization and integrate the translation to an IM client. Specifically, the study intended to accomplish the following objectives:

1. Know the nature and types of compressions used in the Tagalog texting language.
2. Compare and contrast the Tagalog texting language and other texting languages.
3. Find out which model suits the Tagalog texting language.
4. Explain how the Tagalog texting language be normalized using the SMT system.
5. Explain how the Tagalog texting language be normalized in an open-source IM client.

1.4 Significance

Recovering a normalized text is necessary because it can be used in many real-world Natural Language Processing (NLP) applications. One such application is *text-to-speech systems* that read out webpages, blogs and emails to the visually-impaired people. It is also important for *translation system* which processes the conversion of one source language to another language and *text mining applications* such as *filtering*, *routing*, and *information retrieval*.

Moreover, it is also a necessary tool for a *search engine* to be immune to the several texting language variants of a word such as those noisy texts found in blogs, chatlogs and emails.

The commonly used straight dictionary approach in the normalization of texting language has been proven inefficient in the decoding of word written in texting language for this might have more than one compression from its standard form (Henriquez, Hernandez, 2009)(Raghunathan, Krawczyk, 2008). Also, SMS is dynamic in nature and the language's vocabulary changes rapidly and is highly dependent on the region using it.

With these reasons, a study regarding Tagalog text normalization using a Statistical Machine Translation (SMT) approach will contribute significantly to the still-growing area of text normalization using SMT and would open possibilities for more research in the Tagalog language, more importantly, in its normalization using Statistical Machine Translation processing.

1.5 Scope and Limitations

The study covers the translation of the Tagalog texting language to its standard Tagalog form. A Statistical Machine Translation approach will be used in the implementation of the study.

It includes pure Tagalog text normalization alone. Thus, the normalization of texts containing Cebuano, English and/or other languages or texts containing a combination of two or more languages is beyond the scope of the study. Any input texts containing Cebuano, English and/or other language will be retained and will be displayed the same way to that of its original form.

As for the purpose of technical application, the translation system generated from the study will be applied to an open source Instant Messaging (IM) client.

1.6 Definition of Terms

Syntax **–** refers to how words can be combined together to make larger phrases such as sentences (Raghunathan et al.)

Semantics - deals with the meaning of the sentence (Raghunathan et al.)

Translation - the process of converting sentences in a source language to another while preserving the overall thought of the sentences.

noise **–** compressions made into the word, phrase or sentence

Source Text - the text input which needs to be translated

Target Text - the intended text translation of a given source text

Parallel Corpora - text files which contain source-target text pairs

Statistical Machine Translation - an approach to MT that is characterized by the use of machine learning methods (Lopez, 2008, p.2)

N-gram - sequences of n-words