**ACKNOWLEDGEMENT**

**TABLE OF CONTENTS**

**ABSTRACT**

Jejemon had been a recent culture fad in the Philippines and although its impact on the Filipino youth is debatable, a fact is that Jejemon texts are present all over the Internet. This can be problematic since there are many automatic systems that crawl the Internet regularly: website indexers and categorizers, ads and search engines, filters, and many more. If such systems are given noisy input, the probability for a good output decreases, and hence the need for noisy text normalization. This paper presents an approach that was proposed by [1] in the normalization of English and Spanish SMS texts. The Soundex algorithm was modified to suit Filipino. By computing word error rates, the approach proves to be relatively accurate at 82.35%.

**I. INTRODUCTION**

A. Significance of the Study

Jejemon texting, or using Jejenese, is the intentional distortion of the written language to make the text shorter but harder to read and understand [2]. This phenomenon started in short messaging service (SMS) and the necessity to express oneself within the usual 150-character limit. The sub-culture later spread on Twitter [3] and then across various blogs, forums, and social networking sites. Today, there are over a million self-proclaimed Jejemons [4] that have their own language, written text, and even fashion [5]. It is also notable that the previous purpose of the distortion deviated from economic and space restrictions [3] to solely leisure. Instead of converting, for example, the phrase “*I love you*” to “*I luv u*,” the Jejenese for this is now “*1 L0v3 yH0u*.” This intentional distortion has become so popular among the youth that the former secretary of the Department of Education, Mona Valisno, expressed opposition to it. She worries that by “resorting to wrong practice, students’ outcome will also be wrong” [6].

On the Internet community, Jejemon text has a counterpart known as *1337 sp34k* (“leet speak”) which dates back to the early 1980s [7]. Although leet speak and Jejenese both distort the spelling of words, they differ in many aspects. First, leet speak came from elitism (leet from the word elite) on the Internet [8] while Jejenese was born out of the necessity of shortening SMS texts. Computer programmers were probably the first *1337 h4xx0rz* (leet hackers) and leet speak was made to look like programming languages because of the frequent occurrence of numbers [8]. Second, leet speak is used to resemble the spoken language. It uses font change (bigger font size means shouting), color change (red means angry), emoticons (<3, means heart or love), and non-lexical speech sounds (“*hhmmmrrphh!*” can mean dismay) that aid in expressing speech [8]. Jejenese, on the contrary, obscure the expressive meaning of text when it is spoken. Last, leet speak is used deliberately for evading filters, securing passwords, gaming, and hacking computers [7]. This third characteristic of leet speak is unconsciously shared by Jejemon texts, too. It therefore provides the same inferior quality of input to automated systems across the Internet like website indexers and categorizers, data miners, topic generators, and many others.

In addition to a few local news articles and reports, Jejemon was included by two scholarly investigations about cyber culture: a report by [9] and by [10]. The most extensive research done on the sub-culture is perhaps, only those among in the humanities. Alay (in Indonesia), and leet speak, the closest subjects to Jejemon, are also unexplored topics in computer science but not unknown to psychology and humanities. The situation leaves SMS normalization as the most related area of research to Jejemon text normalization. For a while, the challenge of normalizing SMS texts got the attention of the NLP community. Three approaches are dominant in solving this problem: (1) by using spelling correction algorithms, (2) by treating SMS texts as a translation task, and (3) by using speech synthesis algorithms [11]. The spelling correction method is really problematic due to the great amount of noise it brings to the algorithm. To solve this, noisy channel models are built using concepts from probability. The next approach is to treat SMS texts as an entirely new language that needs to be translated to its normal form. However, SMS texts are characterized by great variability to its syntax [11] so the design of the translation machine must be complex and robust enough to handle all possible variations of a single word (*kamusta*, *kmuxta*, *kmustaH*, *muxta*, and so on). The last approach, which is by using speech synthesis algorithms, is the youngest but perhaps the most promising. [1] proposed using the speech synthesis approach, particularly the Soundex algorithm, in normalizing not only SMS texts but other noisy inputs. This study followed that suggestion in the normalization of Jejenese.

Experts categorize Jejemon into three types: mild, moderate, and severe or terminal [4]. This study attempted to normalize Jejenese effectively up to the moderate degree. It also focused on the root language of Jejenese which is Filipino. Therefore, out-of-vocabulary words (names of people, streets, URLs, numbers) and the occasional English words in Jejenese was not given that much attention. Correcting the grammatical structure of the input text is also out the scope of this study.

Jejenese cannot be avoided [5] more so on the online community. Normalization of Jejenese before feeding it to such systems is one good thing to do in alleviating the problem that it brings to automated systems on the Internet. Developers and programmers of these systems will benefit from the research directly, and consequently, consumers of the Internet. Hence, this study may be one of the solutions to the many difficulties that Jejenese poses to the Internet.

B. Objectives of the Study

The general objective of the study is to develop an effective way of normalizing Jejemon text using grapheme-to-phoneme rules and a modified Soundex mapping. Specifically, this study will attempt:

1. to develop a pronunciation dictionary, modify the Soundex algorithm for pronunciations, and define rules in extracting phonemes from the Jejemon text; and

2. to assess the efficacy of the developed approach by computing word error rates.

C. Date and Place of the Study

This study was conducted from October 2012 to February 2013. Preparation of the necessary databases, computing environment, and other preliminaries were finished by the end of October 2012. The main task of developing the mechanism started after the preliminaries and ended before February 2013. Assessment and testing lasted for a month beginning February 2013. A home-based computer proved to be enough for the whole study.

**II. REVIEW OF RELATED LITERATURE**

Jejemon is a local phenomenon in the Philippines; therefore, the task of studying the culture is remitted mainly to the Filipino scientific community. To date, no research in computer science addresses the issue of Jejenese use on the Internet. However, there are certain studies in Natural Language Processing (NLP) about processing of noisy data (or non-standard words) like spelling errors and SMS texts. The latter is closely related to Jejemon text normalization because of their same origin which is from mobile text messaging systems. Various models, mechanisms, and algorithms have been developed over the years to answer this problem. The progress, however, is slow because of a number of difficulties: small data sets [15], out-of-vocabulary words [15], non-standard rules in rule-based approaches [11] [14], disadvantages that certain approaches present [11], absence of current research on context application [14], and the young developmental stages of noisy input corpora like the SMS corpus [16].

A. Resources for the research

Having said that Jejemon is a local phenomenon in the Philippines, certain language resources that are important to NLP like grammar, morphological information, lexicon, and corpora [17] are also localized. Although grammar and morphological information are easily available because of the extensive research made on theoretical linguistics, resources for computational linguists like digitized lexicons and corpora are limited. Pioneering works have been attempted by the Center for Language Technologies, De La Salle University (CeLT, DLSU) to answer this deficiency.

In 2008, CeLT launched Palito, an online repository of mostly literary texts. The main difference of Palito over other corpora is that it tries to build one from an online community. Filipinos from all over the world can upload digitized text, wait for an approval, and that text will be incorporated to Palito. Currently, it has eight languages (Tagalog, Waray-waray, Bikolano, Ilokano, Hiligaynon, Cebuano, Kapampangan, and Pangasinense) plus the Filipino Sign Language; an advantageous feature because most of local NLP researches focus only on Tagalog [17] applied to machine translation.

From this initial work, researchers expanded the corpora by adding entries from other resources. One of the most notable is the corpora made by [13] which contains 9.104 million words for monolingual Filipino, and 14.6 thousand pairs of sentences (580.2 thousand words) for bilingual Filipino-English. Because a community-driven approach can be slow, what they did was to tap the web, majority from news sites, for good-quality texts. The machine translator which used the corpora [13] scored 29.24 in the BLEU metric.

Other researches for language resources are now being conducted not only in the Philippines, but also in neighboring countries like the study made by [18] entitled *The Use of Indonesian Speech Corpora for Developing Filipino Continuous Speech Recognition System*. As a minority language, Filipino needs these kinds of resources for further advancement. Fortunately, progress is now taking place in the Filipino NLP community.

B. SMS normalization

SMS texts have a close proximity to Jejenese because of their same origin. However, unlike SMS texts, Jejenese deviated from the previous purpose of space and economic restrictions [3] to purely leisure. Table 1 illustrates this difference:

|  |  |
| --- | --- |
| **Normal** | *Hello po! Kamusta na kayo diyan?* |
| **SMS language** | *Helo po! Kmsta n kau jan?* |
| **Jejenese** | *3ll0w p0H! Kmxt4H n4h k4u zh4n??* |

*Table 1. Difference of SMS language to Jejenese*

Instead of shortening, Jejenese adds unnecessary letters and symbols that make words, on the average, even longer. The SMS language can be considered as a subset of mild Jejenese characterized by the occasional sprinklings of extra characters, primarily *h*, and misplaced capital cases. Still, the noise of Jejenese when compared to SMS texts is greater.

Using the spelling correction metaphor, the baseline BLEU for the correction is at best at 0.06.

C. Grapheme-to-phoneme rules extraction methods

Extraction of rules for getting possible phonemes from a given grapheme can be divided into manual (hand-derived), automatic, and semi-automatic methods. The manual method (Bonaventura et al., 1998) derives grapheme-to-phoneme rules solely from a linguistic expert, or it will suffice to have, according to Divay (1997), a fairly educated native speaker of the language in discussion. However, the expert needs to set extensive rules, making this approach “labor-intensive and [therefore,] expensive” (Sproat et al., 1998). Next is the automatic method (Gilloux, 1991) which does not need any human assistance. It extracts rules using machine learning and a reasonable quantity of training data. The last method is the most exploited in the past years due to its advantages and minimal error rate. Semi-automatic approaches to rules derivation combine the manual and automatic methods in various ways. Studies by Jansche (2001); Jiampojarman, Cherry, and Kondrak (2010); Kim, Lee, and Lee (2002); Rama, Singh, and Kolachina (2009); Rentzepopoulos, and Kokkinakis (1996); Bosch, and Daelemans (1993); Seneff (2007); and Dwyer, and Kondrak (2009) have used the semi-automatic method in grapheme-to-phoneme conversion.

**LITERATURE CITED**

[1] D. Pinto, D. V. no, Y. Alemán, H. Gómez, N. Loya, and H. Jiménez-Salazar, “The Soundex Phonetic algorithm revisited for SMS text representation,” in *Text, Speech and Dialogue*, P. Sojka et al., Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[2] A. N. Monitor. Philippines: DepEd chief reiterates call to stop jejemon. [Online]. Available: http://search.proquest.com/docview/1027476205?accountid=141440

[3] E. Guillermo. (2010, June) In praise of Jejemon. The Philippine Daily Inquirer. [Online]. Available: http://globalnation.inquirer.net/viewpoints/viewpoints/view/20100614-275524/In-praise-of-Jejemon

[4] J. S. Luis and R. P. (Director). The Jejemon phenomenon: What do experts say? aired April 30, 2013 24 Oras. News Video.

[5] H. Marcoleta. (2010, Apr.) Jejemons: the new ‘jologs’. The Philippine Daily Inquirer. [Online]. Available: http://lifestyle.inquirer.net/2bu/2bu/view/20100424266068/gtJejemons-The-new-jologs

[6] A. N. Monitor. Philippines: Deped: No need to ban jejemon. [Online]. Available: http://search.proquest.com/docview/1027478815?accountid=141440

[7] M. Perea, J. D. nabeitia, and M. Carreiras, “Observation: R34D1NG W0RD5 W1TH NUMB3R5,” Journal of Experimental Psychology: Human Perception and Performance, vol. 34, no. 1, pp. 237–241, 2008.

[8] T. R. LeBlanc, “”Is There A Translator in Teh House?”: Cultural and discourse analysis of a virtual speech community on an internet message board,” Master’s thesis, Louisiana State University, Louisiana, USA, 2005.

[9] Z. Rimay, “Cybercultural communication,” Master’s thesis, Budapest University of Technology and Economics, Budapest, 2010.

[10] M. Fullmer. (2010, June) Literacy in the Facebook era. [Online]. Available: http://waraylanguage.org/literacy-facebook-era.pdf

[11] C. Kobus, F. Yvon, and G. Damnati, “Normalizing SMS: are two metaphors better than one?” in Proceedings of the 22nd International Conference on Computationa Linguistics (Colling 2008). Manchester, 2008, pp. 441–448.

[12] P. Morrow. (2003, May) A brief guide to Filipino pronunciation. [Online]. Available: http://www.mts.net/ pmorrow/ﬁlpro.htm

[13] J. Ilao and R. Guevara. Mining Filipino-English corpora from the web. Retrieved on August 31, 2012 from http://sentrofilipino.upd.edu.ph/bantaywika/ISMAC\_Ilao\_Guevara.pdf

[14] P. Cook and S. Stevenson, “An unsupervised model for text message normalization,” in Proceedings of the Workshop on Computational Approaches to Linguistic Creativity (CALC ’09). Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 71–78.

[15] A. Aw, M. Zhang, J. Xiao, and J. Su, A phrase-based statistical model for SMS text normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions.* Association for Computational Linguistics, Sydney. 2006, 33-40

[16] Fairon and Paumier \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

[17] Roxas R., and Borra A. (2000) Computational linguistics research on Philippine languages. In Proceedings of the *38th Annual Meeting on Association for Computational Linguistics* (ACL '00). Association for Computational Linguistics, Stroudsburg, PA, USA, 1-2. DOI=10.3115/1075218.1075292 http://dx.doi.org/10.3115/1075218.1075292

[18] Sakti\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*