Normalizing ‘Jejemon’ text using a modified mapping of the Soundex Algorithm

Joshua Karl T. Madronio and Vladimir Y. Mariano

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 fairly accurate at 82.35%.**

Index Terms— **Soundex algorithm, text normalization, Jejemon**

I. Introduction

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 (“hhmmmmrrphh!” 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

Presented to the Faculty of the Institute of Computer Science, University of the Philippines Los Baños in partial fulfilment of the requirements for the Degree of Bachelor of Science in Computer Science

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.

II. Materials and Methods

The normalization of Jejemon texts was divided into three general sub-processes: (1) definition of grapheme-to-phoneme rules, (2) building of a Filipino phonetic dictionary with Soundex code, and (3) the main process which is the normalization. The first two sub-processes are preliminaries of the third process. After the normalization, all Jejenese words J_w in the input is expected to be normalized to their equivalent valid words N_w in a one-to-one correspondence.

A. Accumulation and definition of grapheme-to-phoneme rules

The rule-based approach to phoneme conversion was implemented (vis-à-vis the dictionary-based) because the syntax of Jejemon texts varies greatly. The rules were compiled from existing literature, from Jejemon.com, and from hand-crafted rules. The general format for the rules is as follows: [12], in his article *A Brief Guide to Filipino Pronunciation*, reported that there are five vowels (*a, e, i, o, u*) and 23 consonants (*b, c, d, f, g, h, j, k, l, m, n, ñ, ng, p, q, r, s, t, v, w, x, y, z*) in Filipino. The consonants *c, f, j, ñ, q, v, x,* and *z* are primarily used in loaned words from other languages like English and Spanish, and in proper nouns like names of people and streets. He also noted a few "irregularities" like the words *ng* and *mga*, *diy* for the /j/ sound, *siy* for the /sh/ sound, and *au* for the /o/ sound. There is also *tiy* for the /ch/ sound which Morrow missed. Other than these, the rest of the Filipino alphabet is pronounced the way it is written. As for Jejemon mappings, preliminary rules were gathered from Jejemon.com. The site offers an informal list of these mappings which they call as *Jejebeth* or the Jejemon alphabet. The letters, in alphabetical order, are as follows: 4 b c D 3 f 6 h 1 j k 7 m N 0 p Q r 5 t u V w x Y z. It is noteworthy that this alphabet is just a suggestion and that it is not the standard letters to use in Jejenese. A set of about 300 Jejemon sentences was gathered further from the Internet to broaden this base list. After that, the rules were compiled.

ID	Character (Grapheme)	Candidates (Phoneme)
001	a	`D, 2`
002	b	`4, 4D, 42, 43`
...	...	...
n	n_{char}	n_{p1}, n_{p2}, n_{pn}

TABLE I
Format of the rules list

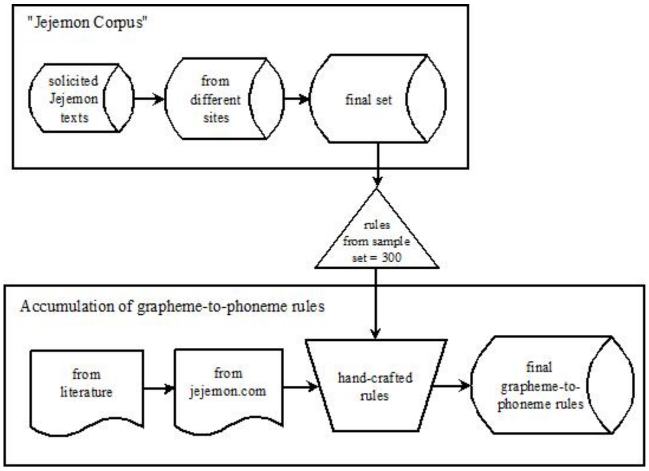

Fig. 1. The framework used for gathering and defining grapheme-to-phoneme rules

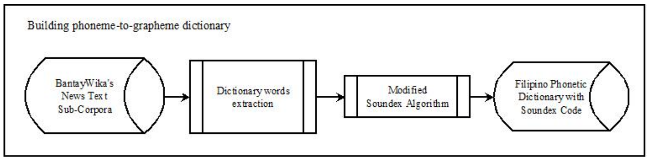
Building phoneme-to-grapheme dictionary
BantayWika's News Text Sub-Corpora
Dictionary words extraction
Modified Soundex Algorithm
Filipino Phonetic Dictionary with Soundex Code

Fig. 2. The framework used for building the Filipino phonetic dictionary

B. Building of a Filipino phonetic dictionary in Soundex code

Aside from the syntax, the pronunciation of Jejemon also deviates, like in the case of the Jejenese *xa* for the word *sa*, or the Jejenese *aqOuh* for the word *ako*. The pronunciation dictionary contains entries of words and their Soundexed counterpart so that a Soundex query will return all the possible words given a certain pronunciation. The entries in the phonetic dictionary were obtained from the BantayWika News Text Sub-Corpora that were built by tapping from the web [13]. The dictionary words and their respective frequencies were then extracted. As proposed by [1], one column of the duplicated entries was converted to the modified Soundex code for noisy queries to the dictionary. The headers of the dictionary are in Table 2 while Figure 2 illustrates the framework followed for the building of the Filipino phonetic dictionary.

ID	Word	Soundex Code	Frequency
001	ako	`D63`	2345
002	akin	`D627`	5674
...	...	...	...
n	$word_n$	$soundex_n$	f

TABLE II
Headers of the phonetic dictionary

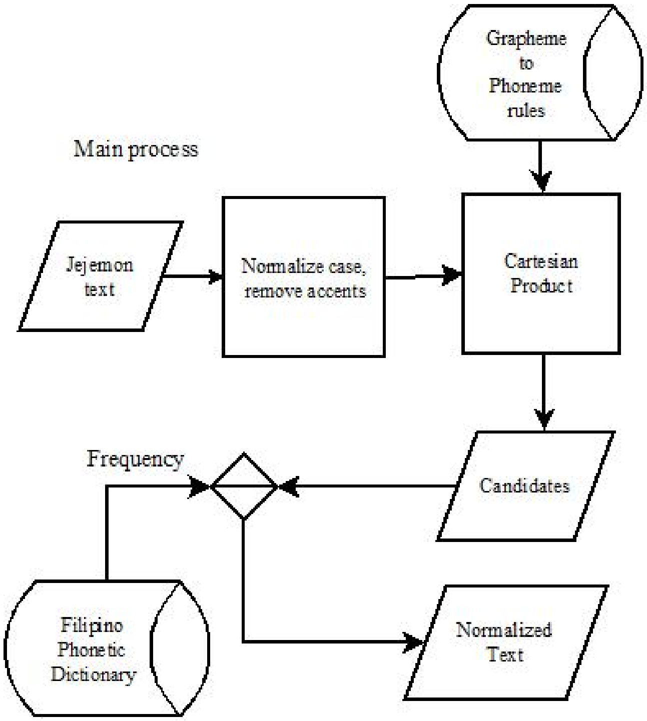
Grapheme to Phoneme rules
Main process
Jejemon text
Normalize case, remove accents
Cartesian Product
Frequency
Candidates
Filipino Phonetic Dictionary
Normalized Text

Fig. 3. The framework used the normalization

C. Normalization

After preparing the rules and the phonetic dictionary, the normalization process can now proceed. First, the Jejemon text must be stripped off of excess repetitive symbols (the *s* and *!* in *el0WsSs!!!*). However, the process must be wary of truly repetitive letters as the case of the letter *a* in the word *maaaring*. After the text has been stripped, it will be converted to its Soundex equivalent. The candidates words given this Soundex code (pronunciation) will then be extracted from the compiled rules using a Cartesian product of the phonemes given by

$$R_1 \times R_2 \times \ldots \times R_n = (r_1 r_2 \ldots r_n), (r_1 r_2 \ldots r_n), \ldots \quad (1)$$

where *R* is a rule and *r* is a Soundex code. For each of these candidates, a Soundex query from the dictionary is performed to get the candidates and their respective frequencies. The results will then be sorted according to its frequency and the heaviest candidate is returned.

III. Results and Discussion

A. Accumulated rules

A total of 251 Jejemon sentences that were roughly assessed as mild (SMS language or text lingo) and moderate were accumulated from various sites including social networks, forums, and the comments section of a news portal; while a total of 206,705 sentences were gathered in the BantayWika Text News Sub-Corpora. The Soundex rules observed from the corpus for the dictionary entries are listed on Table 3.

On the other hand, the reversed mapping/rules observed from the Jejemon sentences that was used for the candidates selection are as listed on Table 4.

Soundex Code	Characters
0	*hw + spacing characters*
1	*0123456789 + non-spacing characters*
2	*eiy*
3	*ou*
4	*bv*
5	*fp*
6	*ckq*
7	*mn*
8	*dt*
9	*gj*
A	*l*
B	*r*
C	*sxz*
D	*a*

TABLE III
Derived Soundex rules from BantayWika Text News Sub-Corpora

B. The Filipino Phonetic Dictionary

After these rules were compiled, the Filipino Phonetic Dictionary was made. For every unique word and a corresponding frequency in the BantayWika Text News-Subcorpora, a dictionary entry was alloted. Table 2 illustrates the header of this dictionary. The number of entries in the phonetic dictionary totalled 143,269.

C. Normalization

A simple evaluation technique was used for the proposed approach: computing word error rates (WER). Given a Jejemon sentence J_s, and its normalized form N_s, count the number of normalized words in N_s. WERs were computed for a test set of 100 Jejemon sentences and averaged. The WER average was found out to be only 17.65%, which is significantly lower than those reported by [14], and [1] : 41% and a lexical similarity evaluation of 38%, respectively. Table 5 in the Appendices lists the common problems encountered by the developed approach.

IV. Conclusion and Future Work

Modifying the Soundex algorithm for Filipino and Jejemon had been effective in the normalization of Jejemon sentences. Although the approach cannot be directly compared to SMS normalization for other languages, an accuracy rate of 82.35% can be said to be good enough given the limited resources.

This paper opens a handful of new possibilities for research because: (1) the developed approach has plenty of room for improvement, (2) Filipino SMS-text normalization is still an 'under-explored' topic in Computer Science, and (3) there might be more efficient but less tedious approaches in solving the same problems.

Problems encountered by the approach also open new areas for research. As an example, word separation for Filipino might be the solution to noisy concatenated words.

V. Appendices

A. The Derived Soundex rules from the gathered Jejemon Sentences

Character	Soundex candidates
a	*D*
b	*4, 42, 43, 4D*
c	*6, C2, 62, 6D*
d	*8, 82, 83, 8D*
e	*2*
f	*5, 25, 52, 53, 5D*
g	*9, 92, 93, 9D*
h	*0, 02, 03, 0D, null*
i	*2, D2*
j	*9, 92, 93, 9D*
k	*6, 36, 62, 63, 6D*
l	*A, 2A, A2, A3, AD*
m	*7, 27, 72, 73, 7D*
n	*7, 27, 72, 73, 7D*
o	*3*
p	*5, 52, 53, 5D*
q	*6, 63*
r	*B, 2B, B2, B3, BD*
s	*C, 2C, C2, C3, CD*
t	*8, 82, 83, 8D*
u	*3, 23*
v	*4, 42, 43, 4D*
w	*0, 02, 03, 0D*
x	*C, 2C, 26C, C2, C3, CD*
y	*2, 02, 22, 23, 2D*
0	*1, 3*
1	*1, 2, A, 0D7, 372*
2	*1, 83*
3	*1, 8B2, 2*
4	*1, 53B, D*
5	*1, C, C2, C3, CD*
6	*1, 1, C2C, 9, 92, 93, 9D*
7	*1, A, A2, A3, AD*
8	*1, 28, 4, 42, 43, 4D*
9	*1, 9, 92, 93, 9D*
Spacing Chars	*0, null*
!	*0, 2, null*
/	*0, A, null*
at	*1, D, D8*
number	*1, 73742B*
and	*1, null*
dollar	*1, A, A2, A3, AD*
bar	*1, D78*
+	*1, 8, 82, 83, 8D*
Unspecified	*1, null*

TABLE IV

Derived Soundex rules from the gathered Jejemon sentences

B. Problems encountered by the approach

Problem	Example
'Jejenized' English words	*gud nYtz!*
Syllable truncation	*wa* for *wala*
Letter truncation	*ndi* for *hindi*
OOV words	*IV-A*
Proper nouns	*Batong Malake St.,*
Word concatenation	*palng* for *pa lang*
Dictionary noise	entries like *mu*

TABLE V

Problems encountered by the approach

References

[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/20100424-266068/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.

[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.