Normalizing 'Jejemon' text using a modified mapping of the Soundex Algorithm

Joshua Karl T. Madronio and Vladimir Y. Mariano

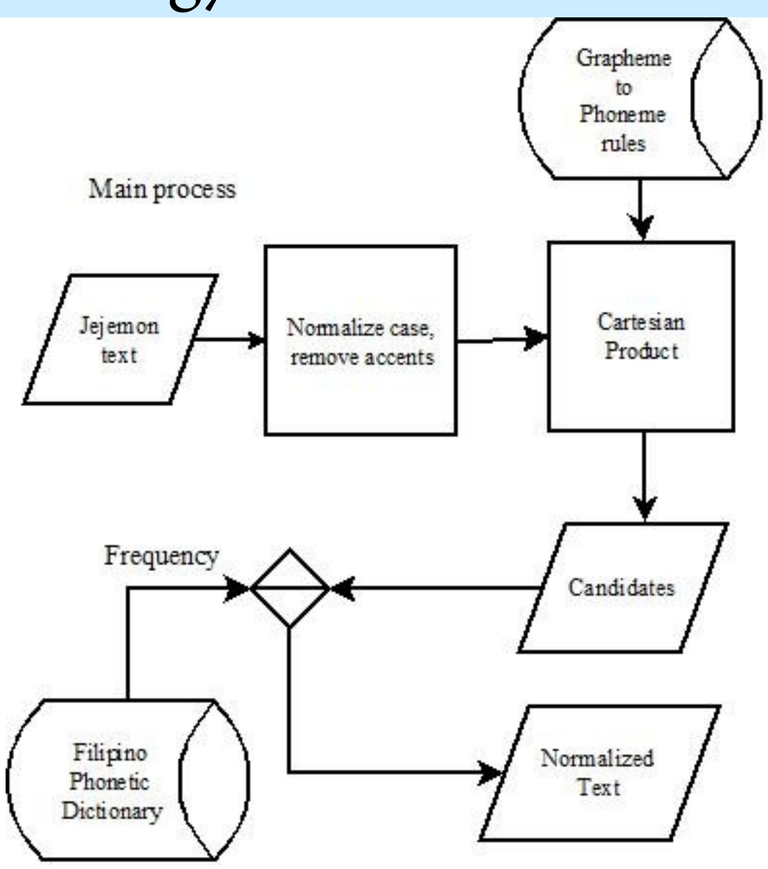
Introduction

Jejemon had been a recent culture fad in the Philippines. Although its impact on the humanities is still debatable, its presence on the Internet along with SMS lingo negatively affects automated systems that crawl them for input. What will happen, for example, to a keyword finder of an advertisement application if it encounters Jejemon and SMS lingo as its inputs? The output's quality will inevitably be noisy too; hence, the need for noisy text normalization arises.

There are three general metaphors in normalizing noisy text: (1) by treating the noisy text as a spelling error, (2) by simulating the text as another language and therefore, a translation task, and (3) by using speech synthesis algorithms [1]. The last approach is the youngest and perhaps the most promising. Pinto [2] used this approach, proposing specifically the Soundex phonetic algorithm. This study follows that suggestion but with modifications of the algorithm for Filipino and Jejemon since the Soundex is made for English.

Limitations of the study include dismissal of English words, out-of-vocabulary words, proper nouns, and grammatical errors. Severe or terminal Jejenese is also out of the scope of the study.

Methodology



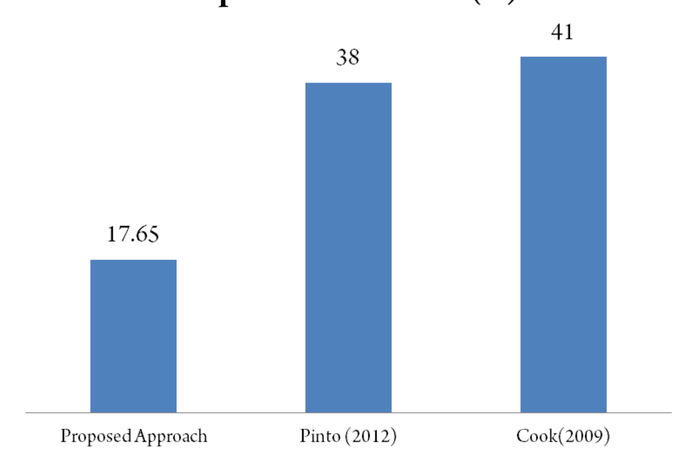
The grapheme-to-phoneme rules were manually extracted from an observed set of 251 Jejemon sentences gathered from various sites and forums. The Filipino Phonetic Dictionary with Soundex code, on the other hand, was derived from UP EEEI Digital Signal Processing Laboratory's BantayWika Text News Sub-Corpora which contains 260,705 sentences. Soundex rules for the dictionary were observed from a subset of this corpus.

[1] C. Kobus, F. Yvon, and G. Damnati, "Normalizing SMS: are two metaphors better than one?" in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, 2008, pp. 441-448.

[2] D. Pinto, D. Vilariño, 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.

[3] C. 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.

Results



The Filipino phonetic dictionary totaled 143,269 entries. The rules were successfully derived for both the dictionary and the grapheme-to-phoneme process. A simple evaluation technique was used for the proposed approach: computing word error rates (WER). Given a Jejemon sentence *Js*, and its normalized form *Ns*, count the number of normalized words in *Ns* divided by the total number of words. 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 [2], and [3] : 41% and a lexical similarity evaluation of 38%, respectively.

Conclusion & 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 study 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.

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

**limitation of the study*

About the Author

Joshua Karl T. Madronio is an undergraduate student of Computer Science at the prestigious ICS, UPLB. He is very thankful to God, to his family, to all of his wise gurus (from elementary and all), and to all of his friends. He likes to see computer applications to new (and preferably, weird) things because he believes in the genius of the machine.