COMP90049 LEXICAL NORMALISATION REPORT

DOES NUMBER OF VOWELS AFFECT THE PRECISION OF PHONETIC MATCHING ALGORITHMS WHEN USED TO CORRECT SPELLING ERRORS?

Introduction

Background Information & Literature Review

Mis-spelling words can be a ubiquitous bane for many social media users. As such, spelling correction algorithms have been developed to predict the word that the user was intending to spell such that the misspelled word can be automatically corrected. One of the many techniques for correcting misspelled words is phonetic matching. Phonetic matching uses how the misspelled word sounds to predict what the user's intended word is (Zobel & Dart 1996).

Amongst all phonetic matching algorithm, Soundex is the most widely known. Soundex was "developed by Odell and Russell, and patented in 1918. It uses codes based on the sound of each letter to translate a string into a canonical form of at most four characters, preserving the first letter" (Zobel & Dart 1996). The following steps outline the procedures involved in translating a word into its sound code:

1. With the exception of the first letter, all letters are grouped according to how they sound and the word is translated into a sound code.
   KYGNNE becomes **K05220**
2. Consecutive consonants are removed
   K05220 becomes **K0520**
3. Vowels are removed
   K0520 becomes **K52**
4. The resultant code is truncated to four characters if it has more than four characters
   K52 remains the same because it has less than four characters.

The translated sound code is then compared with all translated sound codes of words in a dictionary and all words that match the sound code of the misspelled word are considered to have a sound similar to the misspelled word.

Phonix+ is an improved version of Soundex. It groups vowels and client consonants the same way Soundex does, but other consonants are grouped with an improved grouping scheme. The resultant code is not truncated to four characters, and sound codes of words in the dictionary with a minimum edit distance to the sound code of the misspelled word are considered to have a similar sound to the misspelled word (Zobel & Dart 1996).

In a journal article published in *Computers in Genealogy* in 1998, Christian claims that the sound of vowels in English varies so greatly from word to word that they cannot be objectively grouped into sound codes (Christian 1998). Should this claim be true, then words with more vowels will contain more sound variations. Phonetic matching removes vowels from their eventual sound code, and with these vowels removed, these sound variations will therefore not be captured in the sound code. This will then result in the predictions procured by phonetic matching to be less effective.

Goals of the Experiment

In this experiment, we will investigate if there is a correlation between the number of vowels in a word (with exception of the first letter) and the precision of the Soundex and Phonix+ algorithm when used to correct spelling errors. If a negative correlation is observed, we will also investigate if the precision of Phonix+ is significantly better than the precision of Soundex for words containing a higher number of vowels (with the exception of the first letter). The reason for comparing both algorithm's precisions for words with higher number of vowels (with the

exception of the first letter) is because if the number of vowels does reduce the precision of both phonetic matching algorithm, we would like to investigate if Phonix+ can better mitigate the un-captured sound variations caused by the removal of vowels than Soundex.

Both algorithms will be implemented to correct the spelling of the following data sets corresponding to short messages obtained from the social media platform, Twitter (Baldwin et al. 2015),

- A list of words containing 10322 misspelled words on Twitter
- A dictionary of 370099 elements containing the possible words that may be the correct spelling of the misspelled word
- A list of words that corresponds to the correct spelling of the each misspelled word (10322 elements)

The Experiment

Hypotheses

The first hypothesis of this experiment will be in support the claim by Christian, and therefore:

1. **There is a negative correlation between the number of vowels in a word (with exception of the first letter) and the precision of Soundex and Phonix+ when it is used to correct a spelling error.**
2. **The precision of Phonix+ will be double or more than that of Soundex for words with three or more vowels (with the exception of the first letter).**

Implementation and the Evaluation Metric

The data set consists of misspelled words that contain a different number of vowels. Analysing the data on python returns the number of words for a specific vowel count (excluding the first letter):

Vowel count (excluding first letter)	Number of words
No vowels	1695
1 vowel	2315
2 vowels	2879
3 vowels	862
4 vowels	271
5 vowels	59
6 vowels	8
7 vowels	1

It can be observed that the number of misspelled words in each vowel count group drops significantly after four vowel count. For a more accurate result, the experiment will only consider misspelled words that have no vowels, one vowel, two vowels, three vowels, and four vowels.

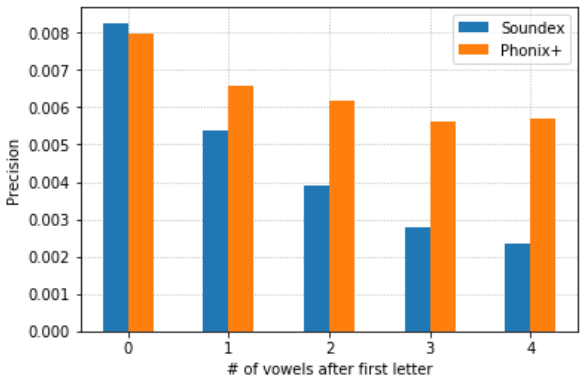
All words in the dictionary and the list of misspelled words are converted into their corresponding sound codes using both Soundex and Phonix+. The number of vowels (excluding the first letter) in each word are also counted in the list of misspelled words. The misspelled word's sound code is then compared with all sound codes of the dictionary to procure the number of returned results and the number of returned result that is the correct spelling of the word (relevant). As such, the number of returned result that is relevant will be either 1 or 0, because, in this experiment, there is only one or no correct spelling of a misspelled word.

The number of returned results and the number of returned result that is relevant are used to compute the precision of Soundex for each of the vowel count group. Precision is a way of quantifying the effectiveness of an approximate matching algorithm and it is the percentage of results returned by the

algorithm that is relevant to the user out of the total number of returned results.

$$Precision = \frac{Number\ of\ returned\ results\ that\ are\ relevant}{Number\ of\ returned\ results}$$

Results and Analysis



The above shows the precisions of Soundex and Phonix+ when it is used to correct the spelling of words with different vowel counts (excluding the first letter). The results reveal that the precision of Soundex reduces exponentially as the vowel count of the misspelled words get increases. The precision of Phonix+ also shows a reduction in precision as the vowel count of the misspelled words get increases, though the reduction is not as drastic as that of Soundex. As such, a negative correlation can be observed between the number of vowels in a word (excluding the first letter) and the precision of Soundex and Phonix+.

Vowel count (excluding first letter)	Precision of Soundex	Precision of Phonix+	Ratio
No vowels	0.826%	0.797%	0.96
1 vowel	0.536%	0.657%	1.22
2 vowels	0.390%	0.618%	1.58
3 vowels	0.280%	0.562%	2.01
4 vowels	0.235%	0.569%	2.42

In the above table, $Ratio = \frac{Precision\ of\ Phonix+}{Precision\ of\ Soundex}$

For words containing three or more vowels, it can be observed that the precision of Phonix+ is more than double of the precision of Soundex. This is a significant difference in precision! Both hypotheses can, therefore, be accepted.

Conclusion

From the results of the experiment, we can conclude that the more vowels present in a misspelled word will cause a reduction the precision of the Soundex and Phonix+ algorithm when it is used to correct spelling errors in the above data set. This conclusion reinforces the claim by Christian that vowels in the English language varies greatly in sound, and therefore cannot be objectively categorised into a sound code as most consonants can. The reduction in the precision of Phonix+ is however far less drastic than that of Soundex. As such, when using phonetic matching algorithms to correct spelling errors, it is perhaps best to use Phonix+ for words that contain three or more vowels (with the exception of the first letter)

Bibilography

Zobel, Justin and Philip Dart. (1996). Phonetic String Matching: Lessons from Information Retrieval. In Proceedings of the Eighteenth International ACM SIGIR Conference on Research and Development in Information Retrieval. Z¨urich, Switzerland. pp. 166–173.

Christian, P., 1998. Soundex-can it be improved?. Computers in Genealogy, 6, pp.215-221.

Baldwin, Timothy, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu (2015) Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. In Proceedings of the ACL 2015 Workshop on Noisy User-generated Text, Beijing, China, pp. 126–135.