DICE'S COEFFICIENT ON TRIGRAM PROFILES AS METRIC FOR LANGUAGE SIMILARITY

Nathaniel Oco, Leif Romeritch Syliongka, Rachel Edita Roxas, Joel Ilao
College of Computer Studies
De La Salle University
Philippines
{nathan.oco,leif.syliongka,rachel.roxas,joel.ilao}@delasalle.ph

***Abstract*— In this study, we present Dice's coefficient on trigram profiles as metric for language similarity. As testbed, we focused on eight Philippine languages. No known language similarity value for these languages exists. Documents containing transcribed audio recordings, news articles, religious and literary texts were taken from an online corpus and used as training data. Character trigram profiles were then generated using an n-gram generator and language similarity was computed. The results were matched against those reported in the literature and against the language family tree. To evaluate the metric, it was applied to five languages with known similarity values. The results were then compared with an existing lexical similarity metric. The average difference is 27%. Analyses of the results reveal that phonetic spelling play an important role in language similarity. As future work, the metric can be used on phonetic transcriptions.**

Keywords—language similarity; Philippine languages; closely-related languages; Dice's coefficient; trigram profiles

I. INTRODUCTION

Language similarity, as the name suggests, refers to the degree of correlation between two languages. Current metrics for language similarity [1], [2] often include cognate analysis, which requires expert knowledge and is tedious. In this study, a metric for language similarity that relies only on text documents as training data is presented. Documents were taken from PALITO [3] and character trigrams were generated using Apache Nutch[1]. The language similarity was computed using Dice's coefficient on trigram profiles.

As testbed, we focused on eight Philippine languages with high number of native speakers. Philippine languages are part of the Malayo-Polynesian languages, which is a subgroup of the Austronesian language family. The languages covered in this study are shown in Table I. These languages were classified in *Ethnologue*, a catalog of the world's living languages, under *Wider Communication Status*[2]: "The languages are used in work and mass media without official status to transcend language differences across a region." Currently, no known similarity value for these languages exists.

[1] Apache Nutch: http://nutch.apache.org/
[2] Philippine Languages Status: http://www.ethnologue.com/country/PH/status

TABLE I. LANGUAGES AND NUMBER OF NATIVE SPEAKERS

Language	Ethnologue Language Code	Number of Native Speakers[3]
Bikol	bcl	4,580,000
Cebuano	ceb	20,030,000
Hiligaynon	hil	5,770,000
Ilocano	ilo	6,920,000
Pampangan	pam	2,900,000
Pangasinan	pag	1,540,000
Tagalog	tgl	21,500,000
Waray-Waray	war	2,560,000

This paper is structured as follows: section 2 discusses related works on language similarity; section 3 covers the similarity metric; evaluation in section 4; and conclusion of our work in section 5.

II. RELATED WORK

Language similarity is a topic that has been studied extensively. Linguists use *cognates* to determine language similarity. Cognates are words having the same etymological origins. McMahon and McMahon [2] applied *phylogenetic analysis* to determine which languages are similar. This maps cognates and other language features (e.g. verb position, relative position of verb and object) in a table and languages with similar features are grouped together. Languages who share the same properties could probably come from the same subgroup of languages. Knowledge about the features is required to properly group similar languages. Downey and colleagues [1] developed a similarity metric using ALINE [4]. In this related work, cognates are identified using an algorithm that matches phonemes. Phonemes must be phonetically transcribed for this to work. In both cases, expert knowledge is required and there is human involvement.

III. METRIC FOR LANGUAGE SIMILARITY

We present a metric for language similarity that does not rely on cognates and expert knowledge. The following steps are involved: gathering volumes of text as training data; creating *language models* using the data collected; and using Dice's coefficient on trigram profiles for the similarity score of

[3] Philippines in Figures: http://www.census.gov.ph/sites/default/files/2013%20PIF.pdf

the language models. These language models are smaller representations of a language, encoded in terms of character sequences and their frequency count. The framework for the metric was taken from a previous work on language identification [5]. In this earlier work, the intersections of language models were counted and its effects on language identification accuracy were determined.

A. N-grams

Language models are expressed in terms of *n-grams*, which is an *n*-character slice of a word [6]. Deriving the formal definition of longest common subsequence [7], the standard formulation for an n-gram is as follows: given a string $X = \{x_1...x_k\}$, character sequences $Z = \{z_1...z_n\}$ is an n-gram of X if there exist a strictly incrementing sequence $i_1...i_n$ of indices of X such that for all $j = 1...n$, $X_{i_j} = Z_j$.

For n of size one, it is called a unigram, size two is called a bigram, and size three is called a trigram. For $n \geq$ four, these are referred to by the value of n (e.g. 4-gram or four-gram). As an example, the list of trigrams that can be generated from the word "*trigram*" are {*_tr,tri,rig,igr,gra,ram,am_*}. An underscore signifies the beginning and end of a word and are part of the trigram. The maximum number of possible n-grams is indicated in Table II. The number of possible combinations increases as n increases.

TABLE II. THE NUMBER OF POSSIBLE COMBINATIONS

Size of *n*	Possible Combinations[a]
1	27^1
2	28^2
3 and above	$28^2 \times 27^{n-2}$

[a] The Philippine alphabet has 27 letters: the 26 letters from the English alphabet and 'ñ'

N-grams of size three are used in this study. A language model composed of trigrams is called a *trigram profile*. Higher values for n cannot cover single-letter words (e.g. "a") while lower values are not enough to represent the unique character sequences of a language. Also, the number of generated n-grams for lower values of n is low, as seen in Table II. Trigrams are enough to represent the unique character sequences of a language while covering single-letter words (e.g. _a_).

B. Trigram Profiles

Documents – containing transcribed audio recordings, news articles, religious and literary texts – covering all eight languages were collected from PALITO [3], an online repository of documents and sign language videos. The sentences underwent cleaning – document and HTML tags were removed. All text documents in PALITO have been examined and verified by linguists, experts, and native speakers of the languages. Table III shows the number of words used per language corpus. Each language corpus covers approximately 205,000 to 284,000 words.

TABLE III. SIZE OF THE TRAINING DATA PER LANGUAGE

Language	Corpus
Bikol	254,000
Cebuano	271,000
Hiligaynon	274,000
Ilocano	258,000
Pampangan	205,000
Pangasinan	261,000
Tagalog	284,000
Waray-Waray	278,000

Trigram profiles were generated by feeding the documents to Apache Nutch. It generated the list of trigrams present in the document and counted their frequency. Fig. 1 shows a scatter plot of the different language models. The long tail starts after rank 100. For this reason, only the top 100 trigrams were used and those with low frequency count were discarded. Low frequency count can be attributed to noise data (e.g. presence of foreign words).

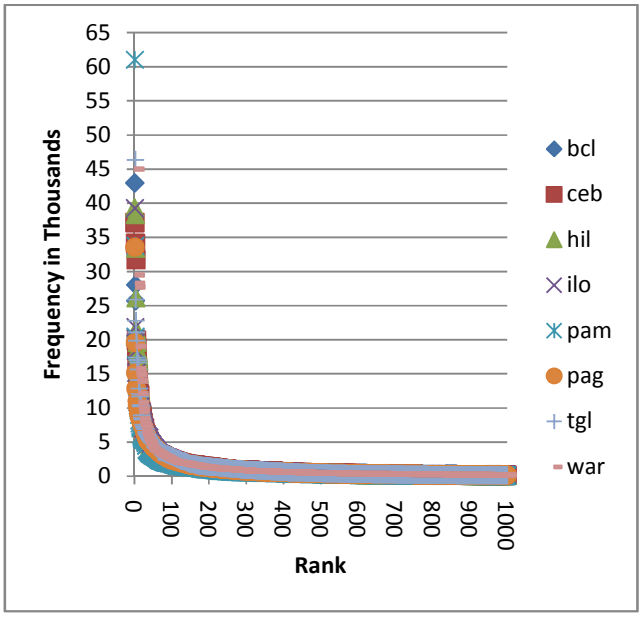

Fig. 1. Scatter plot of the language models showing a power law probability distribution

C. Dice's Coefficient on Trigram Profiles

The process of computing language similarity involves counting the number of trigrams between two languages. This is done using Dice's coefficient [8] defined in (1), where X and Y represent two different trigram profiles. Following this process, a Dice's coefficient matrix was generated, shown in Table IV. The higher the value, the more similar the languages are. In an earlier work [5], only the number of intersecting trigrams was taken.

$$Dice's\ Coefficient = 2(X \cap Y)/(X + Y) \qquad (1)$$

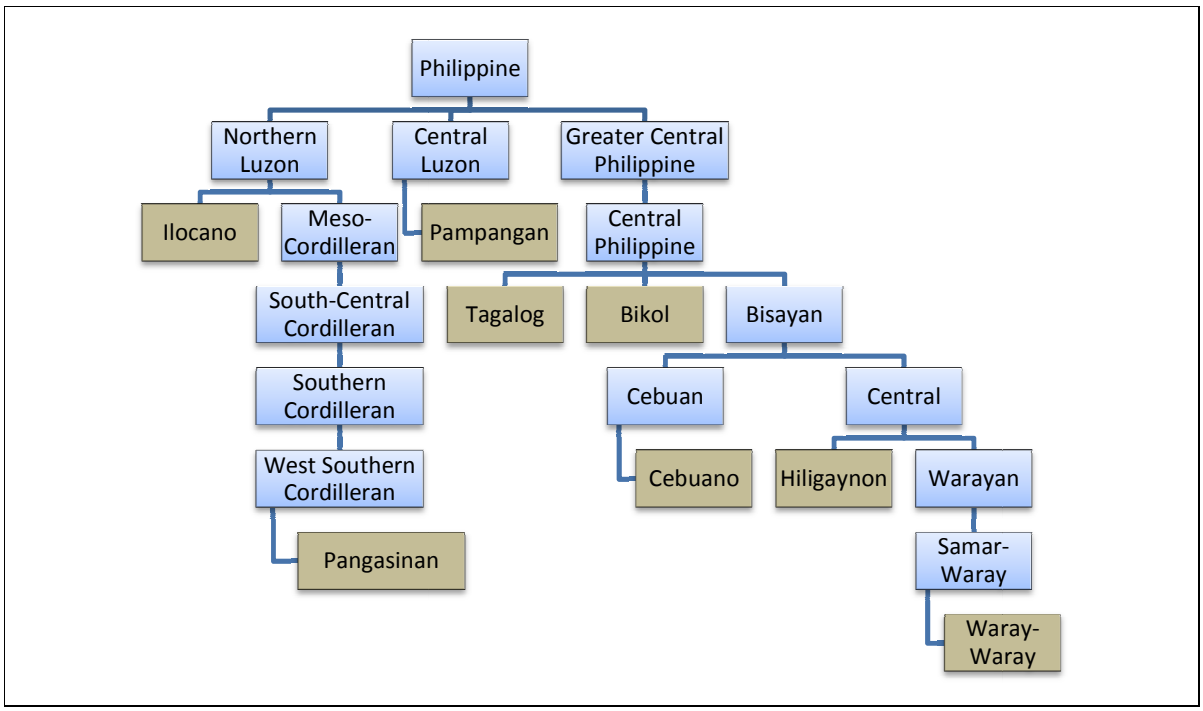

Fig. 2. Language family tree in Ethnologue[4]

TABLE IV. DICE'S COEFFICIENT MATRIX

Language Code	bcl	ceb	hil	ilo	pam	pag	tgl	war
bcl	1.00	0.67	0.62	0.48	0.46	0.52	0.62	0.61
ceb	0.67	1.00	0.75	0.52	0.48	0.55	0.65	0.64
hil	0.62	0.75	1.00	0.51	0.41	0.54	0.67	0.70
ilo	0.48	0.52	0.51	1.00	0.46	0.61	0.55	0.49
pam	0.46	0.48	0.41	0.46	1.00	0.47	0.50	0.41
pag	0.52	0.55	0.54	0.61	0.47	1.00	0.56	0.52
tgl	0.62	0.65	0.67	0.55	0.50	0.56	1.00	0.64
war	0.61	0.64	0.70	0.49	0.41	0.52	0.64	1.00

The Dice's coefficient matrix was matched against the language family tree in Ethnologue (Fig. 2). Analyses show that language pairs with Dice's coefficient values greater than 60 are under the same subgroup. Ilocano and Pangasinan are under the Northern Luzon subgroup of languages while Tagalog, Bikol, Cebuano, Hiligaynon, and Waray-Waray are under the Central Philippine subgroup of languages. This matches what is reported in the literature, which looked at geographical locations and historical records. According to Fortunato [9], Tagalog, Bikol, and Visayan languages (i.e. Cebuano, Hiligaynon, and Waray-Waray) are closely-related while according to Anderson and Anderson [10], Ilocano and Pangasinan are closely-related. The language pairs identified as closely-related languages are shown in Table V. We set the threshold for similarity as the floor function of 0.61, the lowest Dice's coefficient in Table V. Language pairs with lower Dice's coefficient than the threshold 0.60 are not considered closely-related languages.

[4] http://www.ethnologue.com/subgroups/philippine

TABLE V. LIST OF CLOSELY-RELATED LANGUAGES

Language Pair	Dice's Coefficient on Trigram Profiles
Hiligaynon – Cebuano	0.75
Hiligaynon – Waray-Waray	0.70
Hiligaynon – Tagalog	0.67
Cebuano – Bikol	0.67
Cebuano – Tagalog	0.65
Cebuano – Waray-Waray	0.64
Tagalog – Waray-Waray	0.64
Bikol – Hiligaynon	0.62
Bikol – Tagalog	0.62
Ilocano – Pangasinan	0.61
Waray-Waray – Bikol	0.61

IV. EVALUATION

We evaluate the proposed metric by comparing it against a known lexical similarity metric[5]. However, since no existing values exist for Philippine languages, we tested it on five languages with high number of native speakers – English (eng), French (fre), German (ger), Italian (ita), and Spanish (spa). Table VI shows the Dice's coefficient matrix and Table VII shows the Ethnologue lexical similarity matrix. Table VIII shows a table comparing both metrics. Dice's coefficients for the English – French and French – German language pairs are comparable to Ethnologue's. However, the same cannot be stated for English – German, French – Italian, French – Spanish and Italian – Spanish language pairs. The average difference for all six language pairs is 27%.

[5] Ethnologue provides a lexical similarity metric for different languages.

TABLE VI. DICE'S COEFFICIENT VALUES

Language Code	eng	fre	ger	ita	spa
eng	1.00	0.36	0.26	0.32	0.29
fre	0.36	1.00	0.19	0.42	0.41
ger	0.26	0.19	1.00	0.21	0.17
ita	0.32	0.42	0.21	1.00	0.54
Spa	0.29	0.41	0.17	0.54	1.00

TABLE VII. ETHNOLOGUE LEXICAL SIMILARITY

Language Code	eng	fre	ger	ita	Spa
eng	1.00	0.27	0.60	N/A	N/A
fre	0.27	1.00	0.29	0.89	0.75
ger	0.60	0.29	1.00	N/A	N/A
ita	N/A	0.89	N/A	1.00	0.82
Spa	N/A	0.75	N/A	0.82	1.00

TABLE VIII. COMPARISON BETWEEN DICE'S COEFFICIENT ON TRIGRAM PROFILES AND ETHNOLOGUE LEXICAL SIMILARITY

Language Pair	Proposed Metric	Ethnologue Lexical Similarity	Difference
English-French	0.36	0.27	0.09
English-German	0.26	0.60	0.34
French-German	0.19	0.29	0.10
French-Italian	0.42	0.89	0.47
French-Spanish	0.41	0.75	0.34
Italian-Spanish	0.54	0.82	0.28

Analyses of the results reveal one weakness with the proposed metric – phonetics. Dice's coefficient on trigram profiles does not take into consideration the phonetic spelling of words. Languages whose pronunciation keys are different scored lower than those with similar pronunciation keys.

V. CONCLUSION

It has been shown in this research that Dice's coefficient on trigram profiles can be used as metric for language similarity, especially for languages with the same pronunciation keys. As future work, the proposed metric can be used on documents containing phonetic transcription, and on corpora from different genres or domains.

VI. REFERENCES

[1] S.S. Downey, B. Hallmark, M.P. Cox, P. Norquest, and J.S. Lansing, "Computational feature-sensitive reconstruction of language relationships: developing the ALINE distance for comparative historical linguistic reconstruction," J. Qualitative Linguistics, vol. 15, 2008, pp. 340–369.

[2] A. McMahon and R. McMahon, Language classification by the numbers. Oxford, UK: Oxford University Press, 2005.

[3] S.N. Dita, R.E. Roxas, and P. Inventado, "Building online corpora of Philippine languages," in Proc. PACLIC, 2009.

[4] G. Kondrak, "A new algorithm for the alignment of phonetic sequences," in Proc. NAACL, 2000.

[5] N. Oco, J. Ilao, R.E. Roxas, and L.R. Syliongka, "Measuring language similarity using trigrams: limitations of language identification," in Proc. ICRTIT, 2013.

[6] D.M.D. Dimalen and R.E. Roxas, "AutoCor: a query based automatic acquisition of corpora of closely-related languages," in Proc. PACLIC, 2007.

[7] G. Kondrak, "N-gram similarity and distance," in Proc. SPIRE, 2005.

[8] L.R. Dice, "Measures of the amount of ecologic association between species," Ecology, vol. 26 no. 3, 1945, pp. 297-302.

[9] T.F. Fortunato. Mga pangunahing etnoling-guistikong grupo sa Pilipinas. Manila, Philippines: De La Salle University Press, 1993.

[10] V.B. Anderson and J.N. Anderson, "Pangasinan – an endangered language? Retrospect and prospect," Philippine Studies, vol. 55 no. 1, 2007, pp. 116-144.