Filipino Native Language Identification
Using Markov Chain Model and Maximum Likelihood Decision Rule

Jedidiah P. Garcia, Jay Leonarth F. Gecarane, Redentor A. Periabras, Paul Vincent D. L. Sablan
Department of Computer Science
College of Computer and Information Sciences
Polytechnic University of the Philippines
Manila, Philippines
{jedidiahgarcia2203, jaygecarane06, redperiabras, paulvincent.sablan} @gmail.com

Ria A. Sagum, MCS
Department of Computer Science
College of Computer and Information Sciences
Polytechnic University of the Philippines
Manila, Philippines
rasagum@pup.edu.ph

ABSTRACT
In this paper, the researchers determined the native language used on textual data under Cebuano, Kapampangan, and Pangasinan. The Filipino Native Language Identification used Markov Chain Model for language modeling and maximum likelihood decision rule for identifying the native language. Obtained model is applied on 150 text files with minimum length of 10 words and maximum length of 50 words. It shows that the accuracy of proposed system is 86.25% and with F – score of 90.55%. The results of the work can be used as pre–processing unit for higher order NLP practices particularly in Filipino languages.

KEYWORDS
Language Identification, *Filipino Native Language*, *Language Modeling*

I. INTRODUCTION

The Philippines is an archipelago that consists of more than 7,000 islands [1]. It is the reason why the Philippines is considered as multilingual country or country with different languages. It is also the reason of having a difficult correlation with each other and makes the country as one.

Filipino language is characterized by variation within every province. Varieties which share similar features diverge from one another to different degrees. Divergent varieties are often referred to as dialects. In some cases, the varieties may be distinct enough that some would consider them to be separate languages. In other cases, the varieties may be sufficiently similar to be considered merely characteristic of a particular geographic region, social grouping, or historical era. Sometimes speakers may be aware of dialect variation and be able to label a particular dialect with a name. The variation may go largely unnoticed or overlooked.

Language identification is one of the pre-processing unit in natural language processing. It is the task of determining an author's language through statistical computing and works by identifying patterns [2] . It became increasingly important, as more and more textual data is making its way all. Language identification for most of the national language of every country already exists, but as said, the Filipino language for example is characterized by variety. These variations may affect and cause failure with succeeding pre-processing units of NLP. Using a language model is one of the popular approaches to identify a language. Some of the known modeling techniques are through character n – gram, Markov models, naïve Bayes classifiers, support vector machines, and neural networks. N-gram is the base of language modeling. It is where words or letters are grouped into n from the start up to a document or word's termination while moving one step at a time. With a lot of training data, n-gram is able to identify unique characteristics of a language by identifying probability of a word or letter given n-1 words or letters. Knowing the dialects of the Filipino language, this may work, but the fact that these dialects are somehow based from Tagalog, the structure of words may still show potential similarities.

The project aims to create a native language identification tool that recognizes 3 of the 8 major dialects or languages in the Philippines. These languages are *Cebuano*, *Kapampangan* and *Pangasinense*. The Filipino Native Language Identification used Markov chain for language modeling and maximum likelihood decision rule as a method for identifying the native language.

II. RELATED LITERATURES

A. Native Languages

Native language, which was also known as *a first language* or *mother tongue* is a language that a person has been exposed to from birth [3]. The first language of a child is a part of the personal, social and cultural identity [4]. It brings the reflection and learning of successful social patterns of acting and speaking. One can have two or more native languages, thus being a native bilingual or indeed multilingual. These are usually people from *India*, *Philippines*, *Malaysia*, *Singapore* and

South Africa, where most of them speak more than one language.

A native language is said to be:
1. Based on the origin, the languages learned by an individual first;
2. Based on internal identification, the languages an individual is identified as speaker of;
3. Based on external identification, the languages an individual is identified as speaker of, by others;
4. Based on competence, the languages one knows best; and
5. Based on function, the languages one uses most [5].

In Philippines there are more than a hundred languages separately spoken all over the different regions [6] but there are 8 major dialects: *Bikol*, *Cebuano*, *Hiligaynon* or *Ilonggo*, *Ilocano*, *Kapampangan*, *Pangasinan*, *Tagalog*, and *Waray* [7]. These languages are used in different regions of the Philippines as shown in Figure 1.

Although there are 8 varieties, we picked following languages to be identified in our language identification tool assuming that these languages are more distinct from each other.

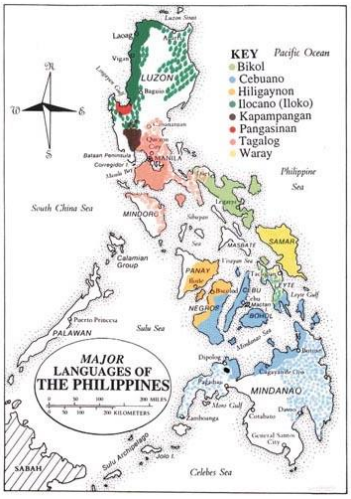

Figure 1. The Major Languages of the Philippines

1. Cebuano

The *Cebuano* language, often colloquially referred to by most of its speakers simply as *Bisaya* ("Visayan"; not to be confused with other Visayan languages), is an Austronesian regional language spoken in the Philippines by about 20 million people, mostly in Central Visayas, eastern Negros Island Region, western parts of Eastern Visayas and most parts of Mindanao, most of whom belong to the Visayan ethnic group [8].

Cebuano Language uses "ABAKADA" as their alphabet consisting 15 consonants and 3 vowels. Vowels i/e and o/u are interchangeable meaning that there is no difference in pronunciation between the two of them. The structure of words are likely the same with Tagalog words excluding the use of foreign letters like c (hard c is replaced by k then soft c is replaced by s), f(replaced by p), j(replaced by dy), q (replaced by kw) ,v(replaced by b) ,x (replaced by ks) and z except with the use of foreign words [9].

Cebuano is not a truly-written language, which means that generally, a Cebuano word is spelled just as it is pronounced. Each syllable and vowel is pronounced separately and distinctly. One Cebuano syllable will consist of either a vowel (V), a vowel with a consonant (CV) / (VC), or a vowel between two consonants (CVC). This makes reading in Cebuano much easier than English because the words can be easily broken into syllables [10].

2. Kapampangan

Kapampangan is an Austronesian language spoken in Indung *Kapampangan* by the ethnic group known as Bangsang *Kapampangan* or *Kapampangan* people. It is located in the northern island of Luzon in the Philippines. The language Kapampangan is also called as Amanung Sisuan which means "breastfed" or "nurtured language" [11].

Standard *Kapampangan* has 21 phonemes, 15 consonants and five vowels. Like other Philippine languages, *Kapampangan* is a predicate-initial language. The predicate is followed by pronouns and/or adverbs and optionally followed by one or more phrases [12]. *Kapampangan* seem to be more liberal in converting nouns to verbs as compared to Tagalog and other Philippine Languages. *Kapampangan* also seem to have no problem in turning adjectives into verbs. *Kapampangan* speaker can be spotted even if he is speaking Tagalog the moment he verbalizes and adjective [11].

3. Pangasinan

Pangasinan language is one of the major languages of the Philippines. It is commonly used in the province of *Pangasinan* and also understood in some places in Benguet and Nueva Ecija and Aeta of Zambales [13]. The *Pangasinense* belongs to the Malayo – Polynesian languages. It is similar to other languages such as *Indonesian* in Indonesia, *Malaysian* in Malaysia, *Hawaiian* in Hawaii and *Malagasy* in Madagascar. The *Pangasinense* is closely related to four small Southern Cordillera languages including *Ibaloy*, *Karaw*, *Kalanguya* and *Ilongot* [14].

Pangasinense is an agglutinating language [15]. It is capable of forming morphologically complex words by stringing together sequences of affixes before and/or after a root morpheme [16]. Hence, it is also capable of undergoing morphophonemic processes and manifests evidence of morphophonemic changes by sequencing of morphemes and phonemes.

The *Pangasinense* is rich in adjectives which are capable of making various morphophonemic changes including assimilation and reduplication. The morphophonologically processed adjectives are being recognized by the native speakers which helps in the registry of *Pangasinense* lexicon [17].

B. Statistical Language Model

It is a model that specifies a priori probability of a particular word sequence in the language of interest. Given an alphabet or inventory of units Σ and a sequence of W = w1,w2,…,wn Є Σ, a language model can be used to compute the probability of W based on parameters previously estimated from a training set. Most commonly the inventory Σ (also called vocabulary) is the list of unique words encountered in the training data [18]. Usually used as training data is corpora – a collection of writings, conversations, speeches, etc. that is used to study and describe language.

C. Language Identification

This is the task of identifying the language used by an author in his or her contexts. The first known approach in language identification is text categorization [19] but the best-known model is per-language character frequency [20] also known as n-gram approach. Variants on this basic method include Bayesian Models for character sequence prediction [21], dot products of word frequency vectors [22] and information theoretic measure of document similarity [23] [24]. Support vector machines(SVMs) and kernel methods we're also applied to the task of language identification [25] [26]. Notice that these approaches make use of notions of such as "word", typically based on the naïve assumption that the language uses white space to delimit words.

D. Native Language Identification

NLI is a fairly recent, but rapidly growing area of research. While some early research was conducted in the early 2000s, most work has only appeared in the last few years. This surge of interest coupled with the inaugural shared task in 2013 has resulted in NLI becoming a well-established NLP task. The NLI Shared Task in 2013 was attended by 29 teams from the NLP and SLA areas. While there exist a large body of literature produced in the last decade, almost all of this work has focused exclusively on English.

NLI are already implemented in other countries such as Kingdom of Saudi Arabia and Finland. Just like the national languages, native ones are also needed to be identified because for some ways or another although they are derived from their national language, they will always differ in their context or even in their spellings.

In Arabic native languages, Sadat presented a comparative study on dialect identification of Arabic language using social media texts; which is considered as a very hard and challenging task. They studied the impact of the character n-gram Markov models and the Naive Bayes classifiers using three n-gram models, unigram, bi-gram and tri-gram. Their results showed that the Naive Bayes classifier performs better than the character n-gram Markov model for most Arabic dialects. Furthermore, the Naive Bayes classifier based on character bi-gram model was more accurate than other classifiers that are based on character uni-gram and tri-gram. Last, their study showed that the six Arabic dialect groups could be distinguished using the Naive Bayes classifier based on character n-gram model with a very good performance [27].

In ASEAN region, one popular research is the ASEAN MT, a practical network-based service on ASEAN languages text translation. They believe that the significance of communication has increased gradually and will become extreme especially after 2015 when the ASEAN Community begins [28]. Native language identification would be a great help to this kind of project.

III. METHODOLOGY

Shannon proposed to use a Markov chain to create a statistical model of the sequences of letters in a piece of English text [29]. Markov chains are now widely used in speech recognition, scientific computing applications including: the genemark algorithm for gene prediction, the Metropolis algorithm for measuring thermodynamical properties, and Google's PageRank algorithm for Web Search [30].

A Markov model predicts that each letter in the alphabet occurs with a fixed probability. We can create a markov model from a specific piece of text by counting the number of occurrences of each letter in that text, and using these counts as probabilities.

For example, we have words ***alun***, ***apot***, and ***apon*** which are *Kapampangan*. We can compute the probability of the occurrences of each letter in those words. Letter *a* has two possible ensuing letter which is *l* and *p* but it happens that *p* is ensuing twice than *l*. From there we can say that the probability that *l* will come next after *a* is *1/3* and the probability that *p* will come next after *a* is *2/3*. The sample model is shown in Figure 2.

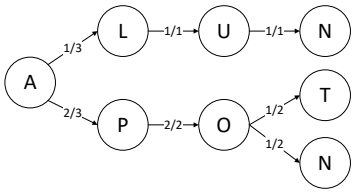

Figure 2. Markov Chain

Markov chains of alphabetical letters' initial probabilities *q(x)* and transition probabilities *p(x,y)* for a language model is interpreted as:

$$q(x) = \frac{number\ of\ occurences\ of\ x\ as\ the\ first\ letter}{number\ of\ words}$$

$$p(x,y) = \frac{number\ of\ pairs(x,y)}{\sum_{z\ \in\ letter\ set} number\ of\ pairs(x,z)}$$

In creating our language model, we used bag-of-words for each language to determine the probability of the occurrences of each letter in tri-gram basis. There we're a total of 35144 words for *Cebuano*, 14752 for *Kapampangan*, and 13969 for

Pangasinan. Given these bag-of-words, it is first preprocessed to remove all special, common characters, and punctuation marks such as commas, columns, semi-columns, quotes, stops, exclamation marks, question marks, signs, etc. The next step is to convert all the characters into lowercases. The initial and transition probabilities are then calculated.

Identification process starts with the modeling the input like what we did in language modeling to obtain the word set X. Next is to start reading all the language models and the letters set obtained from the training session. For each language model, calculate the probability of the word set X using this formula,

$$\log[P(\mathbf{X}=\mathbf{x}\mid\lambda)]=\sum_{i=1}^{M} n_i \log q(i)+\sum_{i=1}^{M}\sum_{j=1}^{M} n_{ij} \log p(i,j)$$

where *M* is the number of alphabetical letters. The language model with best evaluation will be the identification of the unknown language string.

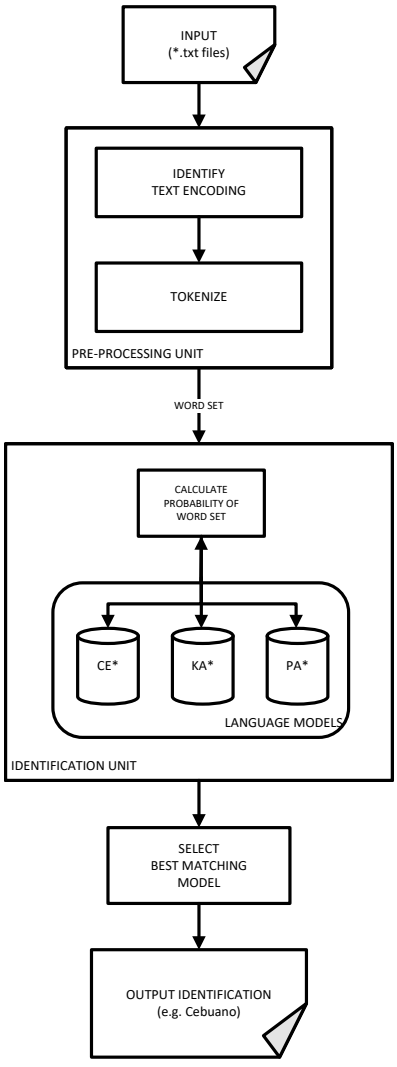

Figure 3. System Architecture

IV. RESULTS

The researchers gathered 150 testing data with 10, 20, 30, 40 and 50 words for each language. These testing data were excerpts from news articles, poems, stories and other literary forms in *Cebuano, Kapampangan,* and *Pangasinan.* By using the tool, the researchers came up with these results:

No. of Words	Pangasinan		Kapampangan		Cebuano	
	F	Accuracy	F	Accuracy	F	Accuracy
10	75	60	71.43	69.23	62.5	50
20	100	100	100	100	100	100
30	100	100	95.24	91.67	94.74	90
40	100	100	86.96	81.25	82.35	70
50	100	100	95.24	91.67	94.74	90
Average	95	92	89.77	86.76	86.87	80

Table 1. Testing Results

The result shows that *Pangasinan* attained highest accuracy over the other two languages. In testing data with minimum of 10 words, the three languages got a low accuracy. The accuracy of identification with the short texts can be significantly improved by creating *n-gram* model for certain length of text. A short text is not commonly to have most frequent *n–grams* in the language. The model should consist of infrequent *n-gam*s [31]. In testing data with minimum of 20 words, the three languages got a perfect accuracy. In testing data of minimum of 30 and 40 words, the *Pangasinan* got a perfect accuracy while the other two languages also got a high accuracy. In last testing data with minimum of 50 words, the *Pangasinan* got a perfect accuracy while the other two languages got the same accuracy attained of testing data in 30 words. The *Pangasinan* got 92% accuracy which attained the highest accuracy while *Kapampangan* and *Cebuano* got 86.76% and 80% accuracy respectively. The result shows that the *Pangasinan* always got a perfect accuracy from the length of 20 to 50 words while the other two languages got a variant accuracy. The imbalance number of training data surpassed the performance of the language model to identify the native language which results of variant accuracy.

V. CONLUSIONS

Language identification tool using a markov chain language model mainly depends on the language model. The researchers assumed that the more training data they had for a certain language model will bring up high accuracy in identifying inputs. But the results of *Cebuano* which has the biggest training data did not support their sentiment, it's accuracy is 9.38% behind than the other two languages.

At first glance, you will surely think that it is the *Cebuano* language model's fault but the researchers think that it is not. For example, you are about to go to college and your preferred course is accountancy. If you will ask someone to tell you about accountancy, which among an accountancy and engineering student are you going to ask? An accountancy student is more knowledgeable about your preferred course so he or she will give you more reliable answer about it compared to the engineering student. The same scenario happens to this tool. Since the training data of *Cebuano* is twice bigger than the other two language models, the evaluation function for the maximum likelihood returns a more sophisticated result. *Kapampangan* model gives a higher evaluation than *Cebuano* model for some of the inputs in *Cebuano* language because *Kapampangan* model has lower knowledge to give it a smarter evaluation.

The researchers therefore conclude that the balance of the training data is a factor to attain fair maximum likelihood evaluation and get a more accurate result for *Cebuano*.

The researchers also concluded that the language models should also contain infrequent n - gram in identifying native language from short texts to increase the identification accuracy.

VI. RECOMMENDATIONS

The following recommendations will be helpful for future enthusiasts on working with Filipino Native Identification:

1. Add features such as wordnet that will help to gather more data training data.
2. Balance the weights or the number of words in the training data of the models to avoid biased likelihood computation.
3. Apply the approaches used by the researchers in this study to other Filipino native languages.
4. Compare the feature and structure of words of different Filipino native languages to support the evaluation results of the future researches related to Filipino native language identification.
5. Cluster and separate native languages with closely related word structure.

VII. REFERENCES

[1] Geography of the Philippines, "Wikipedia," 21 February 2017. [Online]. Available: https://en.wikipedia.org/wiki/Geography_of_the_Philippines. [Accessed March 2 2017].

[2] Native-language identification, "Wikipedia," 26 April 2016. [Online]. Available: https://en.wikipedia.org/wiki/Native-language_identification. [Accessed 2 March 2017].

[3] L. Bloomfield, Language, Motilal Banarsidass Publ, 1935.

[4] T. Hirst, "The Importance of Maintaining a Childs First Language".

[5] "First Language," Wikipedia, 1 February 2017. [Online]. Available: https://en.wikipedia.org/wiki/First_language. [Accessed 8 February 2017].

[6] P. M. Belvez, "Varieties of Filipino," National Commision for Culture and the Arts, 30 April 2015. [Online]. Available: http://ncca.gov.ph/subcommissions/subcommission-on-cultural-disseminationscd/language-and-translation/varieties-of-filipino/. [Accessed 8 February 2017].

[7] "Major Languages of the Philippines," CSUN.edu, [Online]. Available: http://www.csun.edu/~lan56728/majorlanguages.htm. [Accessed 8 February 2017].

[8] "Cebuano Language," Wikipedia, 2011. [Online]. Available: http://en.wikipedia.org/wiki/Cebuano_language. [Accessed 30 January 2017].

[9] T. Mariking, "Learning Cebuano," 7 November 2005. [Online]. Available: http://www.leaningcebuano.com/files/CebuanoStudy Notes.pdf. [Accessed 30 January 2017].

[10] "Living Cebu," 30 May 2002. [Online]. Available: http://www.livingincebu.com/pdf/cebuano/cebuano_language_objectives.pdf. [Accessed 30 January 2017].

[11] M. R. M. Pangilinan, "An Introduction to Kapampangan Language," Tokyo, 2014.

[12] M. R. Pangilinan and K. Hiroaki, "Motivations for Pamamakmul Amanu "Word Swallowing in Kapampangan"," in *Cross-Linguistic Perspective on the Information Structure of the Austronesian Languages*, Tokyo, Japan, 2013.

[13] "Pangasinan Language," Wikipedia, 2011. [Online]. Available: http://en.wikipedia.org/wiki/Pangasinan_language.

[Accessed 30 January 2017].

[14] R. S. Himes, "The souther Cordilleran group of Philippine languages," *Oceanic Linguistics*, vol. 37, no. 1, pp. 120-77, 1998.

[15] R. G. Gordon, "Ethnologue: Languages of the World, Fifteenth edition," 29 August 2009. [Online]. Available: http://www.ethnologue.com/web.asp. [Accessed 30 January 2017].

[16] C. J. Hall, "An Introduction to Language and Linguistics: Breaking the Language Spell," New York, 2005.

[17] M. A. B. Austria, "Assimilation and Reduplication in Pangasinan Adjectives: A Morphophonemic Analysis," SlideShare, 2012. [Online]. Available: http://www.slideshare.net/shinathrun/assimilation-and-reduplication-in-pangasinan-adjectives. [Accessed 31 January 2017].

[18] D. M. Bikel and I. Zitouni, Multilingual Natural Language Processing Applications, From Theory to Practice, IBM Press.

[19] E. M. Gold, "Language Identification in the limit," *Information and Control,* pp. 447-474, 1967.

[20] W. B. Cavnar and J. M. Trenkle, "N-gram based categorization," in *Proceedings of the Third Symposium on Document Analysis and Information Retrieval*, Las Vegas,, 1994.

[21] T. Dunning, "Statistical Identification of the Language," Computing Research Laboratory, 1994.

[22] M. Darnashek, "Gauging Similarity with N-grams: Language-independent categorization of text," *Science*, pp. 267:843-848, 1995.

[23] J. A. Aslam and F. Meredith, "An information-threoretic measure for document similarity," in *Proceedings of 26th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, Canada, 2003.

[24] B. Martins and M. J. Silva, "Language Identification in web pages," in *Proceedings of the 2005 ACM symposium on Applied Computing*, Santa Fe, USA.

[25] O. Teytaud and R. Jalam, "Kernel-based text categorization," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2001)*, Washington DC, USA, 2001.

[26] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins, "Text Classification using string kernels," *Journal of Machine Learning Research,* no. 2, pp. 419-444, 2002.

[27] F. Sadat, F. Kazemi and A. Farzindar, "Automatic Identification of Arabic Language Varieties and Dialects in Social Media," 2004.

[28] "Network-based ASEAN Languages Translation Public Service," ASEAN MT, [Online]. Available: http://aseanmt.org. [Accessed 2 February 2017].

[29] C. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, pp. 379-423, 623-656, October 1948.

[30] "Markov Model of Natural Language," [Online]. Available: https://www.cs.princeton.edu/courses/archive/spring11/cos126/assignments/markov.html. [Accessed 2017 January 2017].

[31] T. Vatanen, J. Vayrynen and S. Virpioja, "Language Identification of Short Text Segments with N-gram Models".

[32] E. W. Weisstein, "Maximum Likelihood," MathWorld - A Wolfram Web Resource, [Online]. Available: http://mathworld.wolfram.com/MaximumLikelihood.html. [Accessed 8 February 2017].

[33] A. K. Singh and G. Jagadeesh, "Identification of Languages and Encodings in Multilingual Document".

[34] D. Tran and D. Sharma, "Markov Models for Written Language Identification".

[35] "Wikipedia," 21 February 2017. [Online]. Available: https://en.wikipedia.org/wiki/Geography_of_the_Philippines. [Accessed 2 March 2017].

[36] "Wikipedia," 26 April 2016. [Online]. Available: https://en.wikipedia.org/wiki/Native-language_identification. [Accessed 2 March 2017].

[37] T. Vatanen, J. Vayrynen and S. Virpioja, "Language Identification of Short Text Segments with N-gram Models".