Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 5475-5478*

Research Article

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

Ria Ambrocio Sagum[a,b]MCS

[a] Department of Computer Science, College of Computer and Information Sciences
[b]Research Management Office, Office of the Vice President for Research Extension and Development
Polytechnic University of the Philippines
Email:[a]rasagum@pup.edu.ph

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: The study developed a tool for identification of a Filipino Native Language given a textual data. The Filipino Language identified were Cebuano, Kapampangan and Pangasinan. It used Markov Chain Model for language modeling using bag of words (a total of 35,144 words for Cebuano, 14752 for Kapampangan, and 13969 of Pangasinan) from each language and maximum likelihood decision rule for the identification of the native language. The obtained model implementing Markov model, was applied in one hundred fifty text files with minimum length of ten words and maximum length of fifty words. The result of the evaluation shows the system's accuracy of 86.25% and an F-Score of 90.55%.

Keywords: Language Identification, NLI with Markov Chain Model, Language Processing

1. Introduction

The Philippines is an archipelago that consists of more than 7000 islands [1]. A factor why the Philippines is considered as multilingual country or country with different languages resulting in a different interpretation amongst the meaning of the word within the country. The language has different varieties for every region or provinces. These varieties share analogous elements diverge from another to different degrees. The divergent varieties are called Filipino dialects. In some cases, the varieties deviate from particular geographic region, social grouping, or historical era. The variation may go largely unnoticed or overlooked. Once this happen this will be the beginning of the misunderstanding of people as sometimes a word may have different meaning depending on the dialect. A problem that can happen also in translating a Filipino sentence. The dialect must be identified to identify the meaning of the word for an accurate translation. Language identification is one of the pre-processing units in natural language processing that can be used to do this task. This identification can be done 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 can be used for Filipino dialects to 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 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.

This study 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 will use Markov Chain for language modeling and maximum likelihood decision rule as a method for identifying the native language.

2.Related Works

2.1.Native Languages

The first language a person is exposed is that person's native language (mother tongue) [3], and is part of our personal, social, and cultural identity [4]. It is normal to have two or more native languages, a native bilingual or indeed multilingual. An example of these are people from India, Philippines, Malaysia, Singapore, and South Africa, who are known that their citizens are fluent in more than one language. A native language is said to be; based on the origin, the language learned by an individual first; based on internal identification, the languages an

individual is identified as speaker of; based on external identification, the languages an individual is identified as speaker of, by others; based on competence, the language one knows best; and based on function, the languages one uses most [5].

In the Philippines there are more than hundred languages spoken over different regions [6]. However, there are considered eight (8) major dialects (languages) and these are; Bikol, Cebuano, Hiligaynon or Ilonggo, Ilocano, Kapampangan, Pangasinan, Tagalog, and Waray.

From these major dialects three dialects were considered to be identified by the language identification tool assuming that these languages are more distinct from each other. These languages are:

1. Cebuano
2. Kapampangan
3. Pangasinan

A. 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 [7]. Different models can be used in language identification (LI) task. To date the model that is being used in LI is the per-language character frequency [8] also known as n-gram approach. Variants on this basic method include Bayesian Models for character sequence prediction [9], dot products of word frequency vectors [10] and information theoretic measure of document similarity [11][12]. Support vector machines (SVMs) and kernel methods were also applied to the task of language identification [13][14]. The common approaches in LI made use of "words", typically based in the naïve assumption that the language uses white space to delimit words.

B. Natural Language Identification (NLI)

This is a new area but beginning a research trend. Some of early researches were 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, generally their work has been focused exclusively on English language.

The area of NLI were 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 of some ways or another they are derived from their national language, they differ in their context or even in their spellings.

In Arabic native languages, the study of Sadat etal. presented a comparative study on dialect identification of Arabic language using social media texts; which is considered as a very hard and challenging task. It focused on the impact of the character n-gram Markov models and the Naïve Bayes classifiers using three n-gram models, unigram, bi-gram and tri-gram in the language identification. They have concluded that Naïve Bayes classifier performs better than the character n-gram Markov model for most Arabic dialects, and noticed that Naïve Bayes classifier base on character bi-gram model was more accurate than the other classifiers that are based on character uni-gram and tri-gram. Their study presented six Arabic dialect groups that can be distinguished using the Naïve Bayes classifier based on character n-gram model with a good performance [15].

In ASEAN region, ASEAN MT, a popular research that is a practical network-based service on ASEAN languages text translation stated that the significance of communication has increased gradually and will become extremely especially after 2015 once ASEAN community begins its integration [16].

3.Methodology

In Shannon's study, he proposed to use a Markov chain in creating a statistical model of the sequences of letters in a piece of English text [17]. This model is evidently being used in speech recognition, scientific computing applications including: the genemark algorithm for gene prediction, the Metropolis algorithm for measuring thermo dynamical properties, and Google's PageRank algorithm for Web Search [18].

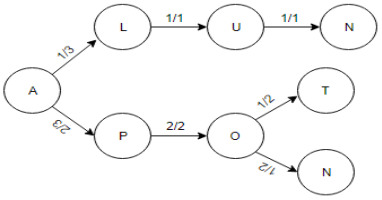

Figure 1: MC model

This model predicts that each letter in the alphabet occurs with a fixed probability. A model can be created from a specific piece of text that counts the number of occurrences of each letter in that text and using these counts as its probabilities.

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

MC of alphabetical letter's initial probabilities q(x) and transition probabilities p(x,y) for a language model is interpreted as: q(x) is the number of occurrence of x as the first letter/ number of words and p(x,y) is the number of pairs (x,y)/ the sum of all the element of letter and number of pairs (x,z).

In creating the system's language model, bag-of-words for each language were used to determine the probability of the occurrences of each letter in tri-gram basis. A total of 35144 words for Cebuano, 14757 for Kampampangan, and 13969 for Pangasinan were used. These bag-of-words went through pre-processing stage wherein special, common characters, and punctuation marks such commas, columns, semi-columns, quotes, stops, exclamation marks, question marks, sign, etc were removed. After which all characters were converted to its lowercase. The initial and transition probabilities are then calculated.

Identification process starts with the modelling the input language to obtain the word set X. Next thing it will do is to start reading all the language models and the letters set obtained from the training session. For each language model, it will calculate the probability of the word set X as :

$$\log[P(\mathrm{X}=\mathrm{x}|\lambda)]\sum_{i=1}^{\mathrm{M}}\mathrm{n}\iota\, logq(i)+\sum_{i=0}^{M}\sum_{i=0}^{M}n\iota\ \ logp(i,j)$$

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

System Design

The system starts with identifying the text encoding used for the input data and splitting it into words. The extracted word set will be compared to each language model to compute for the likelihood evaluation. The language model that returns the highest evaluation score will be declared as the language identification of the input data.

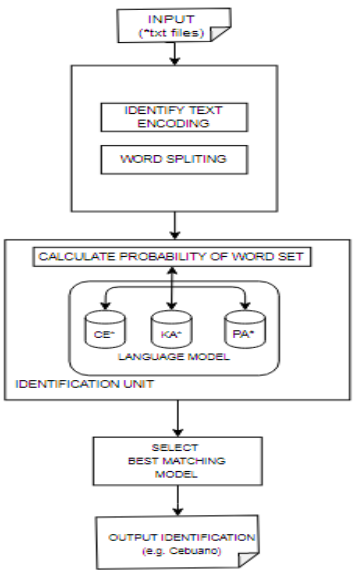

Figure 2: Filipino Identifier

4.Discussion of Results

There were one hundred fifty testing data gathered 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, Kampampangan,* and *Pangasinan.* Table 1 gives the summary of the results of the accuracy of the system.

The result shows that *Pangasinan* attained highest accuracy over the other two languages. In testing data with minimum 10 words, the three languages got a low accuracy. The identification accuracy with the short text can be significantly improved by increasing the number n – grams since that the n – gram approaches stru8ggle with. A short text is not commonly to have most frequent n – grams in the language. The model should consist of infrequent n – gams (Vatanen, Väyrynen, & Jaakko & Virpioja, Sami, p. 2010). In testing data with minimum of 20 words, the tree 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 *Kampampangan* and *Cebuano* got 86.76% and 80% accuracy respectively. The result shows that the *Pangasinan* always got the 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.

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	95.94	91.67	94.74	90
Average	95	92	89.77	86.76	86.87	80

Table 1. Testing Results

5.Conclusions

Language identification tool using a markov chain language model mainly depends on the language model. The assumption done was that given more training data the highest the accuracy will be. But with the results of Cebuano language identification, which has the biggest training data, did not prove that the assumption is correct, its accuracy is 9.38% and it is behind from the two other languages.

At first one surely thinks that it is the Cebuano language model's fault, but the researcher 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. Kampampangan model gives a higher evaluation than Cebuano model for some of the inputs in Cebuano language because Kampampangan model has lower knowledge to give it a smarter evaluation.

It was therefore concluded that the balance of the training data is a factor to attain fair maximum likelihood evaluation and get a more accurate result for Cebuano, and that the language models should also contain infrequent n – gram in identifying native language from short texts to increase the identification accuracy.

6.Recommendations

The following recommendation will be helpful for future researchers on working with Filipino Native Identification: (a) Add features such as wordnet that well help to gather more data training, (b) to balance the weights or the number of words in the training data of the models to avoid biased likelihood computation, (c) apply the approaches used by the researchers in this study to other Filipino native languages, (d) compare the feature and structure of words of different Filipino native languages to support the evaluation results of the future researchers related to Filipino native language identification, and (e) cluster and separate native languages with closely related word structure.