FICOBU: Filipino WordNet Construction Using Decision Tree and Language Modeling

Ria Ambrocio Sagum, Aldrin D. Ramos, and Monique T. Llanes

***Abstract*—The paper discusses the approach in creating a Filipino WordNet. A semi-supervised learning approach using Decision Tree and Language Modeling. This will take advantage on the information found on the web. It will help future NLP researchers in Filipino language. The approach uses words from a dictionary as preliminary data and as seed for the search engine to start crawling the WWW. To decide if the word is part of Filipino language, the word will first undergo in Code-Switching Points Module (CSPD). CSPD scores the word by using the frequency counts of word bigrams and unigrams from language models which were trained from an existing and available corpus. After scoring, Filipino Stemmer will get the stem of the word and examine if the stem word is part of the said language. Once the words were scored and stemmed, the archive will evaluate if the word is Filipino. To test the accuracy of the system, we collected different articles around the web and then grouped it into two groups — Plain Filipino and Bilingual. The result shows the F-measure for Plain Filipino Category range between 65.65% - 96.85% with an average of 85.64% while for Bilingual range between 60% - 100% with an average of 88.17%.**

***Index Terms*—Corpus building, information retrieval, data and web mining, lexicography.**

I. Introduction

Technology has played a big role in the development of various industries, it has changed the banking sector, changed education, changed the agricultural industry, changed the entertainment word, and in has restructured many businesses [1]. The huge impact of technology in the lives of people had been very helpful in giving an easier access to life's difficulties. Computer Science researchers have produced artificially intelligent systems which highly think like a human like Siri. It is a built-in "intelligent assistant" that enables users of Apple devices to speak natural language voice commands to operate the mobile device and its apps [2]. The voice operated assistant uses Natural Language Processing to answer questions, respond to commands, and aid as the user need it [3].

Like the artificially intelligent Siri, many systems use different methods and tools to make a machine recognize and understand human language. One of these is using NLTK (Natural Language Toolkit). NLTK is a platform that deals with natural language, it uses corpora as one of its components to deal with human language [4]. WordNet is one of the corpora it uses. Today, there are 72 existing WordNets in the world available for use. [5] Having the knowledge that the Filipino language is not on the list, accomplishing a WordNet for the Filipino Language provides future NLP researchers a tool in order to do studies. In this study, the researchers aim to create a corpus builder of the Filipino language in order to help in the improvement of linguistic resources for the Philippine National Language.

In 2007, WordNets have already been developed using around forty languages around the world but despite its numerous applications, there is still no WordNet developed for the Filipino language [6]. According to The Global WordNet Association, there are currently 72 existing WordNets in the world available for use [5]. WordNet is a large lexical database for the English Language [7]. As of the moment, there are existing attempts of creating a Filipino WordNet but its classifications, and number of words are still unknown [8] [6].

The motivation for the creation of Filipino WordNet is to provide a solid base of formal linguistic information that could subsequently be used for pertinent language technology applications as outlined by [9]. These include information retrieval and extraction, particularly in concept identification in natural language and in query expansion, language teaching, translation applications, and in parameterizable information systems which allowed personal searching of documents based on users' interest [8].

The name "FiCoBu" was based on the previous unpublished research, "Filipino Corpus Builder". Almost same objectives as the current research, it aimed to build a corpus of the Filipino language for the use of NLP researchers. The system had an accuracy rate of 69.93%. [10]. The said study mostly worked on the collection of data. This study aims to create a WordNet for Filipino language that will support future NLP works that uses Filipino Language using certain algorithms. Improvements were made such as well-structured database.

II. Related Works

Information Retrieval is ultimately an issue of determining which documents in a corpus should be retrieved to satisfy a user's information need which is represented by a query, and contains search term(s), in addition to some information such as the relatively importance. Thus, the retrieval decision is possessed by finding the similarity between the query terms with the index terms appearing in the document [11]. The meaning of an unknown word can often be inferred from its

Manuscript received May 15, 2018; revised October 24, 2018.

R. A. Sagum is with the Department of Computer Science, College of Computer and Information Sciences (CCIS), Polytechnic University of the Philippines, Philippines (e-mail: rasagum@pup.edu.ph).

A. D. Ramos and M. T. Llanes are with Polytechnic University of the Philippines, Philippines.

doi: 10.18178/ijmlc.2019.9.1.772

context [12]. Bootstrapping semantics from text is one of the greatest challenges in natural language learning. It has been argued that similarity plays an important role in word acquisition [13]. Identifying similar words is an initial step in learning the definition of a word. To identify word, WordNet can be used.

There are different methods and tools existing to make a machine understand a natural language. One of these is using a WordNet. WordNet is essential to many NLP applications. There are continuous attempts done by many researchers to create a WordNet for Filipino language, but still not performed.

A. WordNet

WordNet is a large lexical database of English language nouns, verbs, adjectives and adverbs that are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. There are many relations among words in WordNet. The synset noun is connected to synsets hypernyms, hyponyms, coordinate terms, meronym, and holonym by means of semantic relations; and the synset verb to hypernym, tropnym, entailment, and coordinate terms. The main relation among words in WordNet is synonymy, as between the words shut and close or car and automobile. Synonyms--words that denote the same concept and are interchangeable in many contexts--are grouped into unordered sets (synsets). The majority of the WordNet's relations connect words from the same part of speech (POS). Thus, WordNet really consists of four sub-nets, one each for nouns, verbs, adjectives and adverbs, with few cross-POS pointers. [14].

B. Filipino WordNet

The Filipino word-net project (FilWordNet). Borra and his team [8] are currently conducting linguistics research on the evolution of Tagalog grammar among metropolitan residents of the Philippines in which they plan to use FilWordNet in performing manual markup of Filipino corpora. FilWordNet will be a basis for a stemmer/lemmatizer that will use FilWordNet to prevent overly "greedy" removal of affixes from words. FilWordNet will also provide a basis for work in developing a named entity recognition system.

C. Decision Tree Learning and N-Gram

Looking for an algorithm to be used in the study, researchers chose to use Decision Tree Learning. This is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Leaned trees can also be represented as sets of if-then rules to improve human readability. These learning methods are among the most popular of inductive inference algorithms and have been successfully applied to a broad range of tasks from learning to diagnose medical cases to learning to assess credit risk of loan applicants. [15]

In addition to that, an n-gram model will be used in the development of the system. It is a type of probabilistic language model for predicting the next item in such a sequence in the form of a $(n-1)$–order Markov model. [16] N-grams are used for a variety of different task. For example, when developing a language model, n-grams are used to develop not just unigram models but also bigram and trigram models. Google and Microsoft have developed web scale n-gram models that can be used in a variety of tasks such as spelling correction, word breaking and text summarization. Another use of n-grams is for developing features for supervised Machine Learning models such as SVMs, MaxEnt models, Naive Bayes, etc. [17]

D. Language Modeling

Language models were originally developed for the problem of speech recognition; they still play a central role in modern speech recognition systems. The parameter estimation techniques that were originally developed for language modelling are useful in many other contexts, such as the tagging and parsing problems. [18] An approach to effectively detect code-switching points in a Tagalog-English text input, especially those with alternating English and Tagalog words used frequency counts of word bigrams and unigrams from language models which were trained from an existing and available corpus. [19] Code-switching is defined as the use of two or more languages in the same conversation, usually within the same conversational turn, or even within the same sentence of that turn. [20]

III. Methodology

Fig. 1 shows the block diagram of the designed system. There are two major blocks used in the system. First is the Corpus Builder Block that deals with collecting articles and building a corpus. The second block is the API Block which deals in using the created corpus to recognize Filipino words in an article.

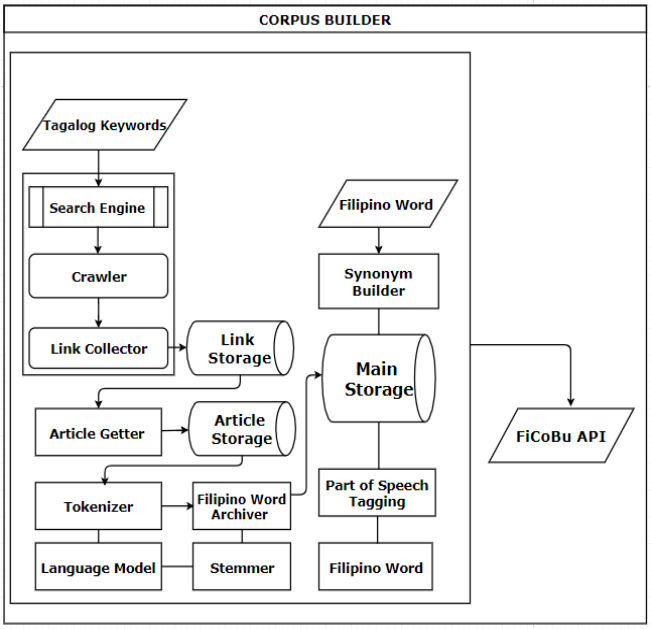

Fig. 1. System architecture of FiCoBu.

A. Preliminary

The researchers used an online dictionary as an initial Tagalog keyword. These keywords are set to be the first piece to make the crawler do its work. The links of the dictionaries are:

- http://www.katig.com/tagalog.html

- http://tagalog.pinoydictionary.com

B. Corpus Builder

1) Crawler

This part consists of a web crawler that will browse the World Wide Web to gather Filipino articles.

2) Stemmer

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form—generally a written word form. [11] In Filipino language there are several ways of putting an affix to a word such as unlapi(prefix), hulapi (suffix), gitlapi (infix), kabilaan (circumfix), and reduplication which could be partial or full reduplication. [21]

a) Prefix

Prefix is a type of affix which is added before a rootword to form new words. Some valid affixes are pinag-, pina-, nag-, pa-, um-, mag-, ma- [22]

Example:
Masaya = ma +saya
Magmahal = mag+mahal
pakain = pa + kain

b) Suffix

Suffix is a type of affix which is added after a root word to form a new word. Some of the valid suffixes are –an, -in, -hin.

Example:
Abangan = abang + an
Anihin = ani + hin
Alamin = alam + in

c) Infix

Infix is a type of affix written within the root word and for a new word. Some valid infixes are –um-, -in-.

d) Circumfix

When a word has two or more type of affixes it is called circumfix.

Example:
Pakainin = pa + kain + in
Mamahalin = ma + mahal + in
Paglinisin = pag + linis + in

e) Reduplication

Reduplication in Filipino could be partial or full. In partial reduplication a certain syllable in a word is duplicated. And full reduplication occurs when the root word was totally duplicated to form a new word.

3) Tokenizer

This process will get the content per page collected by the crawler and will use as input in the language model module.

4) Language model

In training the language model, the proponents used Filipino Bible for the bigram model and different Filipino sentences, songs, poems, and articles for the unigram model. In creating n-gram model, the proponents used the open-sourced program Patterns found in CodeProject. Because of limitation of pattern, for the unigram model the proponents created a system for creating unigram model.

In CSPD Module, there are rules to be followed to score the article and get the Filipino words. As of [19], the rules of scoring are as follows:

- Default score of words is 0
- Threshold is set to 1, as it produced the highest accuracy rates according to [11]
- Let B = {B1, …, Bn } be the set of n bigrams for the input, and ordered as they are seen in the sentences

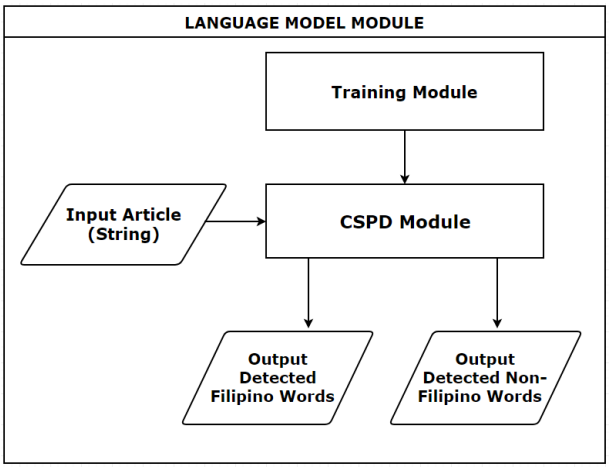

Fig. 2. Language model design.

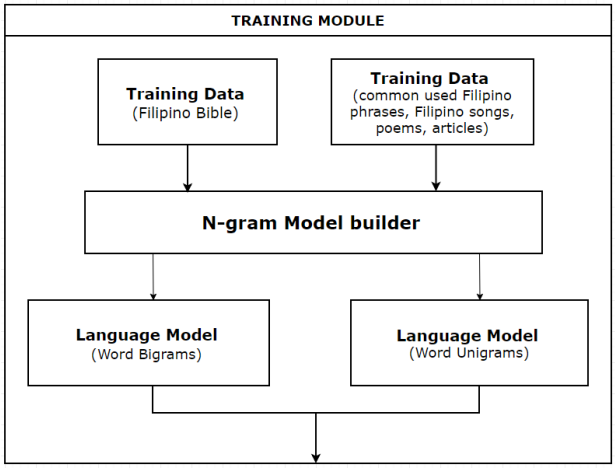

Fig. 3. Training module of language model module.

- Let Bf be a member of B, where f is a value from 1 to n
- Iteration:
  1) If Bf and Bf+1 is frequent (i.e. its frequency is higher than 0), a score of 2 is added to the first and second word in Bf, and for the second word in Bf+1
  2) If Bf is frequent while Bf+1 is infrequent, 1 is added to the first and second words in Bf, while the score of Bf+1 is decremented by 1
  3) If Bf is infrequent and Bf+1 is frequent, the score of the first word in Bf is decremented by 1, while 1 is added to the scores of the first and second words in Bf+1
  4) If both Bf and Bf+1 are infrequent, (2) will take place

For Word Unigram Scoring, these are the rules:

- Words passed here will have a default score of 0
- If a word is frequent in reference to the unigram model, 1 is added to its score
- If a word is infrequent in reference to the unigram model, its score will be left as 0

5) Part of speech

In this process, part of speech tagger is manually typed in a text file and will be placed in a POST Builder or will search in online dictionary.

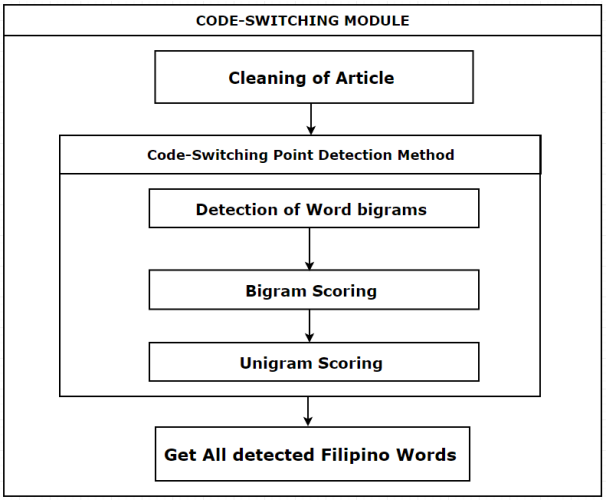

Fig. 4. CSPD module.

6) Synonym builder

Synonym Builder uses an online dictionary to get the synonym of a word.

C. API

The methods in API are as follows:

- isFilipino – determines if the word is a Filipino or not
- .Lemma – will return an object to get the Lemma of the word
- .Sense – will return an object to get the Sense of the word. The Sense includes the part of speech, and example.
- .Synonyms – will return an object to get the synonyms of the word.

IV. Discussion of Results

A. F-Measure

The performance of the system will be measured through the use of the Harmonic Mean, or F-measure. The F-measure is the weighted average of the values of the Precision and Recall. By multiplying the values by 2 and dividing it by the sum of the Precision and Recall, we can get the harmonic mean of the software. A high F1 score will imply a good performance of the software.

B. Recall

Recall will be measured by dividing the total number of Filipino words retrieved by the system, by the total number of Filipino words in the text.

$$R = N_{(\text{retrieved Filipino words})}/N_{(\text{total Filipino words})}$$

C. Precision

Precision will be measured by dividing the total number of correct words tagged by the system, by the sum of the correct words and incorrect words tagged by the system.

$$P = N_{correct} / N_{correct} + N_{incorrect}$$

D. Rating System

To interpret the performance of the system, Table 1 will be the basis in rating it.

TABLE I: Rating System for the Parameters: Precision, Recall, and F-Measure [23]

Computed Value	Verbal Interpretation
97% - 100%	Excellent
93% - 96-99%	Superior
89% - 92.99%	Very Good
85% - 88.99%	Good
80% - 84.99%	Satisfactory
75% - 79.99%	Fair
70%-74.99%	Pass
Below 70%	Fail

Table II shows the summary of evaluation of 500 different articles. This table shows the total Recall, Precision, and the F-Measure of the evaluation. There are total of 25,618 number of Filipino words from the 500 text files used in testing.

TABLE II: Summary of 500 Different Articles Evaluated Using FiCoBu

File Type	NE	Recall	Precision
Plain Filipino	15,377	85.13%	86.29%
Bilingual	10,241	86.68%	88.68%
Total	25,618	85.75%	87.32%

The table also shows that there is lesser unretrieved Filipino words in Bilingual files (15.86%, 1,408 out of 8877) than in Plain files (17.42%, 2,281 out of 13091). Having an 88.67% F-Measure, this implies that FiCoBu is at its best in retrieving Filipino words in Bilingual files compared to 85.70% F-Measure in Plain files. And according to the rating system from chapter 3, 86.53% F-measure indicates that the system is good.

V. Conclusions

The main objective of this research is to create a WordNet for Filipino language by collecting Filipino articles through web-crawling using certain algorithms. Here is the summary of all the findings obtained from the results of the study after conducting a series of implementations on the FiCoBu:

The degree of accuracy of the system in detecting Filipino if article is:

Plain Filipino

- Precision – The FiCoBu gained an overall score of 86.29% out of 250 files. This denotes that out of 250 Plain Filipino files tested, 86.29% of the total words were retrieved correctly by the system. The Precision of the system in retrieving Filipino words in Plain Filipino is good.
- Recall – The FiCoBu gained an overall score of 85.13% out of 250 files. This denotes that 85.13% of the total Filipino words were retrieved by the system. The Recall of the system in retrieving Filipino words in Plain Filipino is good.
- F-Measure – Using the formula for computing F-Measure, the overall computed score for FiCoBu is 85.70%. The F-Measure determines the performance of the system. This means that when testing Plain Filipino, 85.70% is the accuracy rate of the system in retrieving Filipino words

correctly. And based on the rating system by [Ebo ña et al. 2013], the system is good.

Bilingual

- Precision – The FiCoBu gained an overall score of 88.68% out of 250 files. This denotes that out of 250 Bilingual files tested, 88.68% of all words were retrieved correctly by the system. The Precision of the system in retrieving Filipino words in Plain Filipino is almost very good.
- Recall – The FiCoBu gained an overall score of 86.68% out of 250 testing files. This denotes that 86.68% of all the Filipino words were retrieved by the system. The Recall of the system in retrieving Filipino words in Plain Filipino is good.
- F-Measure – Using the formula for computing F-Measure, the overall computed score for FiCoBu is 88.67%. The F-Measure determines the performance of the system. This means that when testing Plain Filipino, 88.67% is the accuracy rate of the system in retrieving Filipino words correctly. And based on the rating system by [Ebo ña et al. 2013], the system is almost very good.

The system has difficulties in detecting uncommon or deep words and words having many affixes. Since the researches have collected articles containing deep Filipino words that are not yet on the system's database, FiCoBu cannot detect it as a Filipino word.

VI. Recommendations

Based on the obtained findings, the researchers recommend:

1) Crawl more Filipino websites in order to gather more Filipino words to be added on the database.
2) Improvements for training data for the language model can produce more accurate detection of Filipino words.
3) The research can be extended and improved in addition of different synset relation like meronymy, holonymy, etc, and adding of etymology which makes a complete WordNet.
4) Improvements in the system's method specifically in the automation of getting the part of speech and synonyms that do not depend on a dictionary.

References

[1] R. Karehka. (2012). The impact of technology on our lives today. Use of Technology. [Online]. Available: http://www.useoftechnology.com/impact-technology-life-today/

[2] S. Forrest. "What is Siri? A word definition from the Webopedia computer dictionary," Webopedia: Online Tech Dictionary for IT Professionals. [Online]. Available: http://www.webopedia.com/TERM/S/siri.html

[3] S. Erica and S. Steve, "Talking to siri: Learning the language of Apple's intelligent assistant," Que Publishing, Indianapolis, 2013.

[4] B. Steven, L. Edward, and K. Ewan, *Natural Language Processing with Python*, O'Reilly Media Inc., 2009.

[5] W. Anja. (2016). Wordnets in the world. The Global WordNet Association. [Online]. Available: http://globalwordnet.org/wordnets-in-the-world/

[6] T. P. P. Paul and L. N. Rose, "Towards a Filipino WordNet," presented at 4th National Natural Language Processing Research Symposium Proceedings.

[7] M. George, "WordNet: A lexical database for English," *Communications of the ACM*, 1995.

[8] B. Allan *et al*., *Introducing Filipino WordNet*, Manila 2010.

[9] J. Morato *et al*., "WordNet applications," in *Proc. the Second Global WordNet Conference (GWC-2004)*, Brno, Czech Republic, 2004.

[10] S. Ria *et al*., *Filipino Corpus Builder*, 2015.

[11] S. Ria *et al*., *Stemmer*, Manila, 2014.

[12] L., Dekang, "Automatic retrieval and clustering of similar words," presented at 17th Internationational Conference on Computational Linguistics.

[13] D. Gentner, "Why nouns are learned before verbs: Linguistic relativity versus natural partitioning," *Language Development: Language, Thought, and Culture*, vol. 2, 1982, pp. 301-334.

[14] F. Christiane, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, 1998.

[15] M. Tom, *Machine Learning*, McGraw-Hill, 1997.

[16] J. Dan and M. Christopher, *Natural Language Processing*, 2015.

[17] G. Kavita, "Text mining, analytics & more: What are n-grams?" *Text Mining & Analytics*, 2014.

[18] C. Michael, *Language Modeling*, Columbia University, New York, 2013.

[19] B. Arianna *et al*., "Detection of intra-sentential code-switching points using word bigram and unigram frequency count," *International Journal of Computer and Communication Engineering*, vol. 3, no. 3, May 2014, pp. 184-188.

[20] M.-S. Carol, *Social Motivations for Codeswitching: Evidence from Africa*, Clarendon Press, 1995.

[21] P. Nocher, *Aralin sa Filipino*, 2012.

[22] J. Hopcroft and J. Ullman, *Introduction to Automata Theory, Languages and Computation*, Jemma, Inc, 1979.

[23] E. Karen *et al*., "Named-entity recognizer (NER) for Filipino novel excerpts using maximum entropy approach," *Journal of Industrial and Intelligent Information*, vol. 1, no. 1, March 2013, pp. 63-67

Ria Ambrocio Sagum was born in Laguna, Philippines on August 31, 1969. She took up bachelor of computer data processing management from the Polytechnic University of the Philippines and Professional Education at the Eulogio Amang Rodriguez Institute of Science and Technology. She received her master's degree in Computer Science from the De La Salle University in 2012. She is pursuing her postgraduate studies and is taking Doctorate in Information Technology at De La Salle University. She is an Associate Professor and Research Coordinator at the Department of Computer Science, College of Computer and Information Sciences, Polytechnic University of the Philippines in Sta. Mesa, Manila. Her specialization is in the field of Natural Language Processing. Ms. Sagum has been a presenter at different conferences both in International and National level. She is a member of different professional associations including ACM-CSTA and a board member of the Computing Society of the Philippines- Natural Language Processing Special Interest Group.

A**ldrin Ramos** grew up in Gapan City, Nueva Ecija, Philippines and graduated from Polytechnic University of the Philippines with a bachelor's degree of computer science specialized in research. He is currently working as a web developer, responsible in creating solutions, delivering content and supports client. He finds himself enjoying sharing to his friends and colleagues what he's been studying and had been learned. One of his officemates once told," I am proud to say that the work he did to this company is a very crucial part of its growth. We learned and gained a lot improvement from the techniques he shares and advices when we're lost to a problem that may seem no answer. I am lucky to be his student and very grateful to everything he'd done", and this marked as one of his achievements in life. When he's not glued in front of the computer, he spends his time sketching watching, reading, or doing photography.

Monique Llanes, was born in Antipolo City, Rizal, Philippines, is the daughter of Nestor and Leonora Llanes. She is a programmer analyst at H2 Software Consulting Services Inc. and was deployed at Concerted Management Corporation. She is also a passionate freelance web designer. She has been designing various fan sites and websites since 2007. She took up Bachelor of Science in Computer Science at the Polytechnic University of the Philippines in Sta. Mesa, Manila.