Detection of Intra-Sentential Code-Switching Points Using Word Bigram and Unigram Frequency Count

Arianna Clarisse R. Bacatan, Bryan Loren D. Castillo, Marjorie Janelle T. Majan, Verlia F. Palermo, and Ria A. Sagum

***Abstract*—Detecting code-switching points is important, especially with the increasing globalism and multilingualism. However, this is a challenging task, but with the help of computers and technology, this can be done easily. In this paper, an approach to effectively detect code-switching points in a Tagalog-English text input, especially those with alternating English and Tagalog words, is presented. The approach uses the frequency counts of word bigrams and unigrams from language models which were trained from an existing and available corpus. For the testing, 3 test data categories were used – twitter posts, conversations, and short stories. The test data were composed of a total of 3088 English and Tagalog words. The results show that the system's accuracy of properly identifying English and Tagalog words ranged from 81% - 95%, while the F-measure ranged from 72% - 95%. The research can be extended and improved using other n-grams, stemming, and searching algorithms.**

***Index Terms*—Code-switching point detection, intra-sentential code-switching, word bigram, word unigram.**

I. Introduction

According to study [1], a typical Filipino (those not living in a Tagalog-using area) grows up to speak at least three languages - the vernacular, English, and Filipino (Tagalog). Consequently, it is inevitable that code- switching occur in conversations among Filipinos. Code- switching is "a salient phenomenon and experience to most Filipinos" [2]. With this in mind, it is certain that those who live in the metro area of the country are not exempted from code-switching. In fact, conversations among typical Filipinos living around the Tagalog-using areas are composed of both English and Tagalog words. In Linguistics, this phenomenon is called Code-Switching. Detecting code-switching points is important, especially with the increasing globalism and multilingualism.

The effective detection of code-switching points can help in the development of intelligent systems with multilingual services such as traveling systems and automatic multilingual call centers. However, code-switching point detection is a challenging task, but with the help of computers and technology, this can be done easily.

There are a variety of approaches which can be used to identify code-switching points, such as affixation information, vocabulary list, alphabet n-gram, grapheme n-gram, and syllable structure [3]. The mentioned approaches can be grouped into two, one which uses n-grams and one which is dictionary-based. For dictionary-based approaches (e.g. affixation information, vocabulary list), words are matched and compared against a dictionary, while in n-gram based approaches (e.g. alphabet bigram, grapheme bigram, syllable structure), similarity measures and language models are used. Both types of approaches were proven to detect code-switching points, but according to the study [3], n-gram based approaches proved to yield more accurate results.

Choosing the right approach, along with some modifications and variations, code-switching points can be effectively and accurately detected, even in alternating English and Tagalog words.

Manuscript received January 2, 2014; revised March 8, 2014.

A. C. Bacatan, B. L. Castillo, M. J. Majan, and V. Palermo are with the University of Santo Tomas, Manila, Philippines (e-mail: ariannabacatan@gmail.com, bryan.loren.castillo@gmail.com, marjoriemajan@gmail.com, lian.palermo@gmail.com).

R. Sagum is with the Polytechnic University of the Philippines, Sta. Mesa, Manila and is also with the University of Santo Tomas, Manila (e-mail: riasagum31@yahoo.com).

II. Related Works

A. Code-Switching Point Detection

Code-Switching is the practice of moving back and forth between two or more languages, dialects, or registers. It is the use of more than one linguistic variety in a manner consistent with the syntax and phonology of each variety. More definitions of code-switching is presented in Table I [4]-[7].

TABLE I: Different Definitions of Code-Switching

Study	Definition
Hymes, D. (1962)	A common term for alternative use of two or more languages, varieties of a language or even speech styles
Hoffman, C. (1991)	The alternate use of two languages or linguistic varieties within the same utterance or during the same conversation
Myers-Scotton (1993)	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
Gross (2006)	A complex, skilled linguistic strategy used by bilingual speakers to convey important social meanings above and beyond the referential content of an utterance.

A code-switched sentence is consisted of two main parts, the primary language and the secondary language. The secondary language is usually manifested in the form of short expressions such as words and phrases [8]. According to [9], there are three kinds of code-switching:

1) Intersentential code-switching (e.g. "No, you can't use that learning stick. Bawal yan. Naku, Nika, ang mali mo talaga.") (Translated as: "No, you can't use that learning stick. That's not allowed. Oh no, Nika, you are wrong.")
2) Intrasentential code-switching (e.g. "Dapat ma-melt

yung ice.") (Translated as: "The ice should be melted.")

3) Tagswitching (e.g. "Sige, (ok) I'll buy it.") (Translated as: "Alright, (okay) I'll buy it.")

There have been a lot of researches regarding code-switching. These researches focus mainly on topics such as speech recognition, language identification, text-to-speech synthesis and code-switching speech database creation [8].

B. Dictionary-Based CSPD

The study [10] improved the accuracy rate of a dictionary-based approach for automatic code switching point detection. The study successfully addressed the problems, intra-word code-switching and common words, which caused the mentioned approach to have a 79.96% accuracy only. Their devised pattern matching refinements (PMRs) were common word exclusion, common word identification, and common n-gram pruning, and it showed an accuracy of 94.51%.

C. N-gram Based CSPD

The study [11], detected code-switching points by using frequency counts of word bigrams trained from an existing and available corpus. The study was able to correctly identify English and Tagalog words with accuracy rates that ranged from 52.53% to 90.61%. The research also lead to the conclusion that the system, though it has a high range of accuracy, cannot properly identify code-switches in alternating English and Tagalog words.

III. CSPD Using Word Bigram and Unigram Frquency Count

Fig. 1 shows the simplified CSPD System Architecture which includes the Training and Code-Switching Point Detection (CSPD) Module, and the output file.

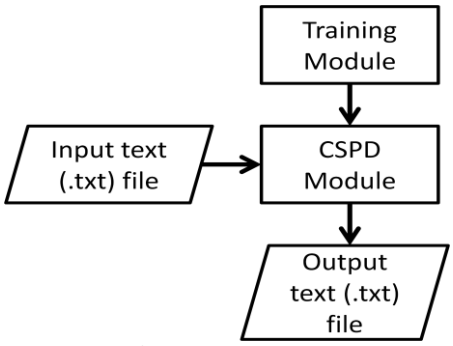

Fig. 1. Simplified CSPD system architecture.

A. Training Module

The system uses two language models, one word bigram language model and one word unigram language model. The models were created using the open-source n-grams tool by Z. Yu. The tool is packaged in C++ and uses ternary search tree instead of a hashing table for faster n-gram frequency counting. The proponents then used Cygwin, a Linux-like environment which runs on Windows, to run the n-grams tool. The word bigram model was trained using 500,000 sentences from the 1 meelyun corpus whereas the word unigram model was trained using 1,000 commonly used English phrases and sentences.

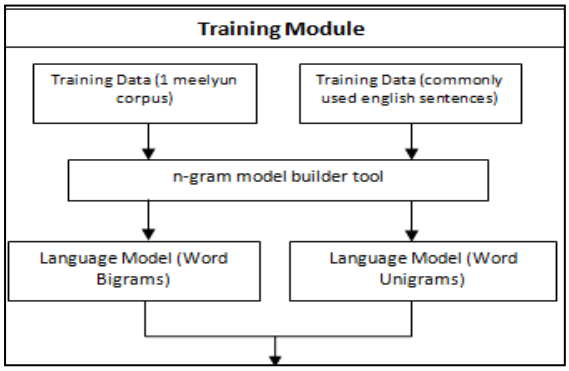

Fig. 2. The training module.

Fig. 2 shows the training module wherein two models are present, a word bigram language model and a word unigram language model.

B. Code-Switching Module

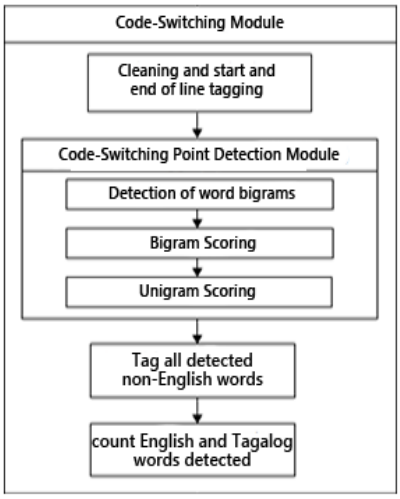

Fig. 3. The code-switching module.

Fig. 3 shows the Code-Switching Module. This includes the cleaning and tagging, Code-Switching Point Detection Module, tagging of all non-English words, and the output which consists of the count of English and Tagalog words detected.

1) Cleaning and start and end of line tagging

All sentence-boundary and concept-boundary punctuation marks (e.g. '.', ',', '!', '?', etc.) from the input are removed, and every start and end of the line are tagged (<s>, </s>).

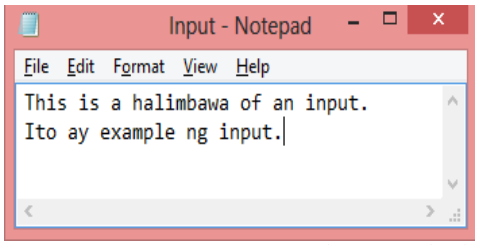

Fig. 4. Sample input text (translated as: "this is an example of an input").

Fig. 4 shows the sample input. The input consists of a combination of English and Tagalog words.

2) CSPD (code-switching point detection)

Word bigrams are extracted from the text input, scored based on the bigram language model, and then all bigrams whose scores are below the threshold are scored again based

on the unigram language model. Afterwards, words whose scores are below the threshold are considered as Tagalog words, and then they are tagged. This module implements (1) and (2).

3) Tagging of detected non-English words

All words whose scores are below the threshold, will be tagged (<tag>, </tag>).

4) Counting of detected English and Tagalog words

The detected Tagalog words are counted by using a counter which is incremented by one when <tag> is read from the arraylist of words and ends in </tag>. For counting the English words, the counter for English words is incremented when the word is not tagged.

C. Output

The output is a text file where all detected Tagalog words are tagged. The total number of words, together with the total number of bigrams, number of English words, and number of Tagalog words are also part of the output. Additionally, the list of Tagalog words detected is also part of the output.

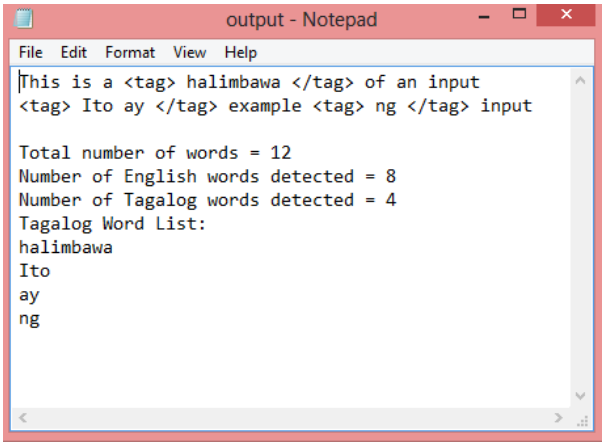

Fig. 5. Sample output text.

Fig. 5 shows the sample output with the tagged Tagalog words.

IV. Equations and Scoring Process

A. Word Bigram Scoring

- Default score of words is 0
- Threshold is set to 1, as it produced the highest accuracy rates according to [11]
- Let $B = \{B_1, \ldots, B_n\}$ be the set of n bigrams for the input, and ordered as they are seen in the sentences
- Let B_f be a member of B, where f is a value from 1 to n
- Iteration:

1) If B_f and B_{f+1} is frequent (i.e. its frequency is higher than 0), a score of 2 is added to the first and second word in B_f, and for the second word in B_{f+1}
2) If B_f is frequent while B_{f+1} is infrequent, 1 is added to the first and second words in B_f ,while the score of B_{f+1} is decremented by 1
3) If B_f is infrequent and B_{f+1} is frequent, the score of the first word in B_f is decremented by 1, while 1 is added to the scores of the first and second words in B_{f+1}
4) If both B_f and B_{f+1} are infrequent, (2) will take place

B. Word Unigram Scoring

- 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

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Words}}$$

where:

- True Positives is the total number of correctly identified English words.
- True Negatives is the total number of correctly identified non-English words.
- Total Words is the total number of words in the input.

$$\text{Word Error Rate} = \frac{WT}{T} \qquad \text{Precision} = \frac{CT}{T}$$

$$\text{Recall} = \frac{CT}{NT} \qquad \text{F-measure} = \frac{2PR}{P+R}$$

where:

- T is the number of supposed to be tagged words (expected number of tags)
- CT is the number of correct tags
- WT is the number of wrong tags
- NT is the actual number of words tagged (actual result)
- P is the Precision
- R is the Recall

V. Testing

The system was tested using various text inputs containing both English and Tagalog words. Different types of input were used: Texts from social-networking sites like Twitter, texts containing dialogues/conversations, and short Stories. (Shown in Table II-Table IV). The following results are computed using (3) and (4):

TABLE II: Twitter Test Results

Test	Accuracy	WER	P	R	F-Measure
1	84.31%	8.57%	0.857	0.909	88.24%
2	85%	17.07%	0.878	0.837	85.71%
3	76.09%	3.76%	0.714	0.95	81.55%
4	81%	10.17%	0.780	0.902	83.64%
5	79.17%	5.77%	0.769	0.909	83.33%
6	80.81%	17.78%	0.756	0.810	78.16%

TABLE III: Conversations Test Results

Test	Accuracy	WER	P	R	F-measure
1	92.06%	13.64%	0.909	0.870	88.89%
2	94.67%	0%	0.911	1	95.35%
3	89.87%	18.52%	0.889	0.828	85.71%
4	86.84%	4%	0.64	0.941	76.19%
5	86.67%	10%	0.833	0.892	86.21%
6	89.87%	25%	0.917	0.785	84.62%

TABLE IV: SHORT STORIES TEST RESULTS

Test	Accuracy	WER	P	R	F-Measure
1	85.36%	74.42%	0.953	0.547	69.49%
2	83.27%	56.34%	0.915	0.597	72.22%
3	81.89%	65.75%	0.973	0.597	73.57%
4	88.68%	27.59%	0.931	0.771	84.38%
5	83.78%	25%	0.833	0.769	80%
6	90.35%	24.24%	0.909	0.789	84.51%

Test results show that the use of word bigram and unigram models yielded outputs ranging between 75-85% accuracy (Twitter posts), 85-95% accuracy (conversations) and 80-90% accuracy (short stories). However, as can be seen in the Word Error Rate column, while the texts from Twitter and conversations achieved a low WER, the short stories which comprises twice the number of words of the other categories, showed a high WER. The over-tagging of Tagalog words was identified as the main cause of this result. Also, there are instances where some rare English words were not detected as English and were assumed as Tagalog words. Another cause of the tagging errors is the detection of short-lettered Tagalog words as English.

VI. CONCLUSION AND FUTURE WORK

This study used Word Bigram and Unigram Frequency Counts in identifying and tagging Tagalog words in a sentence. The addition of a word unigram model for the word scoring was proven to be successful in addressing the problem of accurately detecting code-switches, especially between alternating English and Tagalog words. As a result, this study was able to achieve a highest accuracy rate of 94.67%. However, unless a way to speed up the processing time is developed, models created from a large corpus (e.g. character n-grams from a large training data) are not suggested.

The following are the recommendations of the developers in order to improve the capabilities of the system as well as the algorithms and methods involved in detecting intra-word sentential code-switching points:

- As a future work, the speed of comparing frequencies could be improved.
- Usage of a bigger corpus for the unigram language model.
- Addition of Word Stemming to the system process and corpus could also be tested and explored.
- The improvement regarding the system's capability in handling long texts is also recommended.
- Also, the addition of character n-gram and other higher n-gram levels can be tested and explored.
- Named-Entity Recognition can be applied to solve the problem of classifying proper nouns present in the text input.
- This work can also be applied to other bilingual languages, involving English and a language other than Tagalog.
- This study can also be applied and connected with studies involving the process of speech-to-text conversion so that it could be used in detecting and tagging Tagalog-English code-switched audio conversations.

VII. SUMMARY

The research titled "Detection of Intra-Sentential Code-Switching Points Using Word Bigram and Unigram Frequency Count" successfully developed a code-switching point detection system which can effectively process bilingual text inputs composed of English and Tagalog words.

The detection of the code-switches was done based on the frequency count of the input in reference to the language models. Two language models were used, a word bigram and a word unigram language model. Once the input is processed by the system, it undergoes bigram scoring and unigram scoring. All words whose scores are below the threshold are considered as non-English, ergo Tagalog in this study, and are tagged.

Afterwards, an output is produced where Tagalog words are tagged. The total number of words, total number of detected English words, total number of detected Tagalog words, and a Tagalog word list is also part of the output.

ACKNOWLEDGMENTS

The proponents would like to thank Mr. Nathaniel Oco for being instrumental in this study and to the Almighty God, for giving the proponents wisdom, strength, and guidance throughout the completion of this research.

REFERENCES

[1] A. Gonzales, "The language planning situation in the Philippines," *Journal of Multilingual and Multicultural Development*, vol. 19, no. 5-6, pp. 487-525, 1998.
[2] M. L. S. Bautista, "Code-switching studies in the Philippines," *International Journal of the Sociology of Language*, vol. 88, issue 1, pp. 19-32, 1991.
[3] T. P. Tan and Y. L. Yeong, "Applying grapheme, word, and syllable information for language identification in code switching sentences," in *Proc. International Conference on Asian Language Processing*, 2011, pp. 111-114.
[4] S. Gross, "Code-switching," in *Encyclopedia of Language and Linguistics*, K. Brown, Ed., Elsevier, 2006, pp. 508-511.
[5] C. Hoffman, *An Introduction to Bilingualism*, New York: Longman Inc., Ch. 5.
[6] J. Gumperz and D. Hymes, *Directions in Sociolinguistics: The Ethnography in Speaking*, New York: Holt, Rinehart, and Winston, 1962, pp. 407-434.
[7] Myers-Scotton, *Social Motivations for Codeswitching: Evidence from Africa*, Oxford: Clarendon Press, 1993.
[8] L. C. Yu, W. C. He, W. N. Chien, and Y. H. Tseng. (2013). Identification of code-switched sentences and words using language modeling approaches. *Mathematical Problems in Engineering*. [Online]. Available: http://www.hindawi.com/journals/mpe/2013/898714/cta/
[9] R. Metila, "Decoding the switch: the functions of codeswitching in the classroom," *Education Quarterly*, vol. 67, no. 1, pp. 44-61, 2009.
[10] N. Oco and R. Roxas, "Pattern matching refinements to dictionary-based code-switching point detection", in *Proc. 26th Pacific Asia Conference on Language Information and Computation*, 2012, pp. 229-236.
[11] J. Ilao, N. Oco, R. Roxas, and J. Wong, "Detectiong code-switches using word bigram frequency count," in *Proc. the 9th National Natural Language Processing Research Symposium*, Ateneo de Manila University, Quezon City, 2013, pp. 30-34.

Arianna Clarisse R. Bacatan was born in Marikina on June 12, 1993. She received secondary education at St. James College of Quezon City from 2006-2010. She is now a senior student at the University of Santo Tomas in Manila, Philippines. She is taking up bachelor of science in computer science and is expecting to graduate in April of 2014.

Bryan Loren D. Castillo was born in Mandaluyong on November 24, 1991. He received secondary education at Colegio San Agustin of Makati City from 2006-2010. He is now a senior student at the University of Santo Tomas in Manila, Philippines. He is taking up bachelor of science in computer science and is expecting to graduate in April of 2014.

Marjorie Janelle T. Majan was born in Manila on April 25, 1993. She received secondary education at St. Stephen's High School, Manila from 2006-2010. She is now a senior student at the University of Santo Tomas in Manila, Philippines. She is taking up bachelor of science in computer science and is expecting to graduate in April of 2014.

Verlia F. Palermo was born in Manila on February 28, 1994. She received secondary education at St. Mary's College, Quezon City from 2006-2010. She is now a senior student at the University of Santo Tomas in Manila, Philippines. She is taking up bachelor of science in computer science and is expecting to graduate in April of 2014.

Ria A. 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 currently teaching at the Department of Computer Science, College of Computer and Information Sciences, Polytechnic University of the Philippines in Sta. Mesa, Manila and a lecturer at the information and computer studies, Faculty of Engineering, University of Santo Tomas in Manila. Ms. Sagum has been a presenter at different conferences, including the 6th international conference on humanoid, nanotechnology, information technology, communication and control, environment, and management 2013 (HNICEM), 2012 international conference on e- commerce, e-administration, e-society, e-education, and e-technology and national natural language processing research symposium. She is a member of different professional associations including ACMCSTA and an active member of the Computing Society of the Philippines- Natural Language Processing Special Interest Group.