PANG-KAT: A DEDICATED TOKENIZER FOR THE TAGALOG LANGUAGE

Justin Louis L. Saavedra and Jaderick P. Pabico
Institute of Computer Science, College of Arts and Sciences, University of the Philippines - Los Baños

INTRODUCTION

Tokenization is a critical preprocessing step in many natural language processing (NLP) tasks, yet most general tokenization methods do not effectively recognize the unique grammatical features, named entities (NEs), and multi-word expressions (MWEs) of a language. Low-resource languages are more prone to these issues, including the Tagalog language, which is the most spoken language in the Philippines, but lacks a suitable tokenizer for accurately processing its language resources to promote its further NLP advancements.

- OBJECTIVE: Develop a dedicated tokenizer for the Tagalog language to address its lack of a suitable tokenizer, which would also incorporate the recognition of Tagalog NEs and MWEs, and Tagalog-English (Taglish) tokenization in an attempt to make a general tokenizer for real-world Tagalog language processing applications.

METHODOLOGY

To achieve its objective, the study introduces PANG-KAT, a Python module developed as a hybrid rule and dictionary-based tokenizer due to the lack of pre-annotated Tagalog resources for a machine-learning approach. To facilitate the recognition of Tagalog NEs and MWEs, a dictionary for PANG-KAT was manually created by extracting its vocabulary from various publicly available datasets and online resources. For PANG-KAT rules, these were manually extracted from two available Universal Dependencies (UD) treebanks in the Tagalog Language: Tagalog Reference Grammar (TRG) and Ugnayan, cross-referenced with KWF's Manwal sa Masinop na Pagsulat, and on the TweetTaglish corpus that is manually annotated to observe patterns in Taglish code-switching.

PANG-KAT employs initial short-unit tokenization using Inside–Outside–Beginning (IOB) tagging and longer-unit tokenization to identify Tagalog NEs and MWEs based on its dictionary categories and ruleset. PANG-KAT's performance was evaluated through unit testing and external validation on the manually annotated NewsPH-NLI dataset and May 2025 articles from Pilipino Star Ngayon using the performance evaluation metrics of accuracy, precision, recall, and F1 Score.

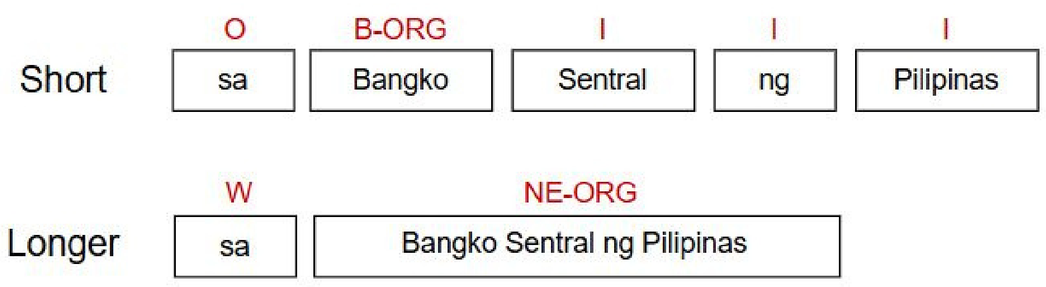

Figure 1. Example of short and longer unit tokenization with their corresponding labels. PANGKAT's dictionary categories include MWEs and NEs such as organization (NE-ORG), location (NE-LOC), and person (NE-PER) entities. Non-NEs and MWEs are labelled as "W", which corresponds to a word.

RESULTS & DISCUSSION

Table 1. The dictionary size and composition of PANG-KAT. Expanding PANG-KAT's dictionary is crucial for improving the overall performance of PANG-KAT in detecting and classifying Tagalog NEs and MWEs.

CATEGORY	TOKEN COUNT
Location entities (NE-LOC)	48,383
Organization entities (NE-ORG)	3,564
Person Entities (NE-PER)	7,321
Multi-word Expressions (MWE)	3,294
PANG-KAT's Total Dictionary Size	**62,562**

Table 2. The composition of the manually annotated TweetTaglish Corpus using IOB tagging, which is used for observing patterns in Tagalog and English (Taglish) code-switching to enable PANG-KAT to perform Taglish tokenization.

LEGEND	TOKEN COUNT
Sentences	1,858
Labeled Tokens	34,424

Table 3. The manually extracted rules of the Tagalog language, which were converted into code, integrated into the ruleset of PANG-KAT, and combined with the dictionary lookup for detecting Tagalog NEs and MWEs. Aside from these Tagalog-specific rules, PANG-KAT's ruleset also includes general rules on the proper usage of punctuation and other symbols, such as basic mathematical symbols.

RULE	EXAMPLE
Intensification of word meaning	malungkot na malungkot, magandang maganda, darating at darating
Continuity of action	tumakbo nang tumakbo, palakas nang palakas
Repeating words	gaya-gaya, kani kanila, tumalon-talon, nagtuloy tuloy, ang lungkot-lungkot
Contractions	ako'y, dalawampu't apat, s'ya, '97, 'wag
Tagalog date and time	ika-15 ng Abril, a-kinse ng Abril, ika-tatlo ng hapon, alas tres ng hapon
Borrowed Spanish conjunction "y"	alas tres y medya, trenta y dos, Jose Protasio Rizal y Alonso Realonda
Daglat	Kgg. Pangilinan, G. Pangilinan, Pangilinan, PhD, DepEd
Spelled-out large Tagalog numbers	limampu't isang libo, walong daang libo, sampung milyon, 11 bilyon
Beginning Markers for Taglish code-switching (Tagalog prefix + English word)	nagma-marites, ipa-consult, pagkabitter, napaka close

PANG-KAT operates on a linear, first-match-wins architecture, wherein when a token is labeled, it won't be reconsidered by later rules. Thus, rule sequencing is crucial for ensuring accurate token labeling, which starts with more structured, Tagalog-specific rules, followed by general rules on the proper usage of punctuation, and the broadest dictionary lookups. When a punctuation mark has a language-specific use, this specific function must take precedence within its corresponding rule block.

Token labels are only updated based on specific triggers that are mainly applicable for Tagalog NEs. For instance, in the phrase "Pasig Mayor Vico Sotto," the token Pasig would be initially labelled as a location (B-LOC), but would be updated as part of a person entity (B-PER) after matching the full entity using the B-PER dictionary.

In performance evaluation, PANG-KAT achieved F1 scores exceeding 0.9 on both unit testing and external validation for both short and longer unit tokenization. Given that the F1-score ranges from 0 to 1, with 1 indicating perfect classification performance for the model, PANG-KAT F1 scores indicate good performance in accurately tokenizing and classifying Tagalog NEs and MWEs.

Table 4. Performance evaluation results of PANG-KAT's short-unit tokenization for both unit testing and external validation.

DATASETS	TOKEN COUNT	ACCURACY	PRECISION	RECALL	F1-SCORE
TRG	734	0.9959	0.9649	0.9821	0.9735
Ugnayan	1,025	1.0	1.0	1.0	1.0
TweetTaglish	34,424	0.9943	0.9753	0.9981	0.9865
NewsPH-NLI	28,922	0.9730	0.9632	0.9443	0.9537
Pilipino Star Ngayon Articles	5,787	0.9758	0.9852	0.9407	0.9624

Table 5. Performance evaluation results of PANG-KAT's longer-unit tokenization for both unit testing and external validation.

DATASETS	TOKEN COUNT	ACCURACY	PRECISION	RECALL	F1-SCORE
TRG	734	0.9958	0.9473	0.9730	0.9600
Ugnayan	1,025	1.0	1.0	1.0	1.0
TweetTaglish	34,424	0.9931	0.9443	0.9968	0.9698
NewsPH-NLI	28,922	0.9764	0.9034	0.9202	0.9117
Pilipino Star Ngayon Articles	5,787	0.9801	0.9612	0.9036	0.9315

The deductions on PANG-KAT's F1 score is mainly impacted by its strong reliance on the word patterns in its ruleset; it lacks the ability to recognize deeper contextual meanings and word relationships. Strict word patterns also lead to misclassifications when handling misspellings or incorrect usage of punctuation. Additionally, its dictionary has limited vocabulary and is specifically tailored to the Philippine context. Integrating PANG-KAT with additional Tagalog pre-processing modules and expanding its dictionary to broader contexts would further improve its tokenization and classification capabilities.

CONCLUSION

The main objective of this study is to develop PANG-KAT to address the lack of language-specific NLP tools for the Tagalog language, particularly a suitable tokenizer. Through rigorous development and testing, the performance evaluation results of PANG-KAT indicated good performance in accurately tokenizing and classifying Tagalog NEs and MWEs. These results affirm the effectiveness of its ruleset and dictionaries in capturing the patterns in Tagalog and Taglish texts. Further improvements in PANG-KAT's performance can still be achieved by integrating it with additional Tagalog pre-processing modules and expanding its dictionary, all to develop a dedicated Tagalog tokenizer that could serve as a foundation for the development of more advanced Tagalog NLP tools and bridge the gap that impedes its potential NLP advancements.