International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

Derivational and Inflectional Stemming Approach for the Filipino Part-of-Speech Tagging

Great Allan M. Ong
Technological Institute of the Philippines

Abstract: ***This paper aims to present an increase in the accuracy of the existing Tagalog part-of-speech tagging (TPOST) by introducing derivational and inflectional stemming approach through hybrid and modified stemming algorithm. A collected 89 derivational and inflectional assimilatory common words and 22 basic Filipino region names and languages for the KSTEM algorithm are utilized for the initial stemming process. The assimilatory and partial duplication of assimilatory word rules were modified using the same approach. Since TPOST showed a high percentage on stemming errors and wrong feature extraction in the part-of-speech tagging, the evaluation was focused on strengthening this methods. For the entire stemming testing sets, a 17.73% decrease in under stemming index and a 0.00967 over stemming index was produced. A total of 2.32 % and 3.42% assimilation word feature success rate was produced in the combined test sets for the part-of-speech tagging. The success of hybrid stemming relies on the assimilatory word search, therefore a trained data was listed and evaluated and produced a 9.39% assimilatory word features success rate. Despite of these variations, an innate morphological study by stemming modification and KSTEM strengthening focusing on the old Filipino assimilation of word and diversified samples are recommended.***
Keywords: Stemming, Tagalog, Natural Language Processing, Part of Speech, Morphology

I. INTRODUCTION

In recent years, computers have played an important role in most aspects of human lives, especially in the field of computational linguist. This field includes knowledge from linguistic, computer science and logic. Furthermore, its main goal is to make computer understand and speak natural language [1]. There are many interesting tasks in computer linguistics that many researchers may develop, such as machine translation, information extraction, questioning and answering, and parsing. These are considered high-level natural processing in which it requires fundamental process like stemming and part-of-speech tagging.

Filipino is the national language in the Philippines which commonly suggest the official name of the Tagalog [2]. The language is actually based on the mixture of native and contemporary Philippine languages rather than in Tagalog alone [3]. This means that the Filipino language is diverse and compliant. Because of its diversity and complexity, a constant and continuous update in the field of Filipino Natural Language Processing is necessary. Filipino machine translation also plays an important role in domains such as education, linguistics, theoretical and formal computations. However, the natural language processing of the Filipino language is still a young field of research with relatively fewer studies compared to other languages.

NLP plays a vital role in the research field; the vital role that NLP plays in the current business arena and the huge potential of the technology as an instrument for addressing the language problems in the country. However, there is low level of researches and awareness among students and support from educational institutions with regards to NLP in the Philippines [4].

II. PART OF SPEECH TAGGING

Part-of-speech (POS) tagging is one important phase in natural language processing. It is the process of labelling words in sentences with different part of speech. POS tagging is necessary in other NLP applications such as named entity recognition and syntactic analysis [5]. Part-of-speech tagging is an essential tool in many natural language processing applications such as word sense disambiguation, parsing, question answering, and machine translation[6].

Despite the developments of POS taggers in the country, the Filipino language's evolution requires constant updates on the tools and resources. Without these updates, the products become outdated in factors such as data contents, software usability, performance and availability [7]. Because of the complexity and the contemporary change that affects the language, and the limited yet diverge resources, a morphological and part of speech tagging analysis will be utterly challenging.

III.THE TPOST SYSTEM

TPOST is a template-based n-gram Part-Of-Speech (POS) tagger for Tagalog and is designed to utilize some lexical resources. The key to the algorithm is in the use of word features, which consists of (1) predefined words, (2) affixes, and (3) other word

characteristics and symbols such as capitalization and hyphens. Different variations on the algorithm were performed to reduce the errors, and to make TPOST algorithm a good foundation for further research in the field of POS Tagging [8]. However, TPOST shows 70% accuracy due to stemming errors, cascading errors, lack of resources, simple scoring, lack of training, and ambiguity. There are a lot of words that cannot be stemmed using simple string matching. A comprehensive lexicon and a morphological analyzer is needed to get the correct features. TPOST uses an existing Tagalog Stemming algorithm (TagSA).

IV. THE TAGSA SYSTEM

TagSA is a Tagalog Stemming Algorithm that can be used specifically for morphological analysis focuses on deriving root words. It uses the principle of iterative affix removal and is context sensitive. The initial method refers to the non-stemming stage that handles the hyphen-search and dictionary-search routines. Second stemming method refers to the stemming stage. In this stage, every removal of an affix requires a dictionary look-up to avoid overdoing a stemming process [9].

TAGSA is intended to be used as a pre-process to a morphological analyser; it is also extended in the handling assimilatory changes in words, except that it relies solely on dictionary look-up and causes a number of over stemming errors.

V. STEMMING MODIFICATION

KSTEM was employed as a pre-stemmer, consisting of 89 derivational and inflectional collected assimilatory common words and 22 common Filipino clusters of natives and languages. The collected lexicons are used in the training and test data evaluation of stemming and word feature extraction for the part of speech tagging.

KSTEM is a morphological analyzer that reduces morphological variants to a root form. Unlike previous stemmers, KSTEM tries to avoid conflating variants that have different meanings. The process starts by evaluating sample word. The word will then pass through dictionary look-up. Initial process takes place in finding words in the lexicon then checks its derivational and inflectional structure from the rules generated. The procedure will then return words instead of truncated word forms as seen if figure 1.

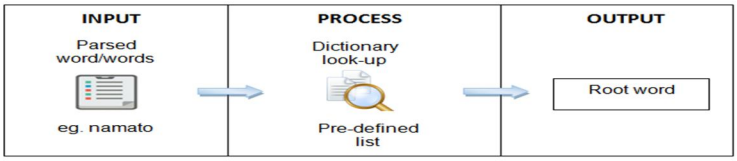

Figure 1. KSTEM Process

In general, KSTEM requires a word and root word to be in the lexicon (the basic list of words that the system knows about) before it will reduce one word form to another. If the input word was not found in the lexicon, an affix look-up will be instigated. The next process follows rule checking of the fragmented affix and derived inflected word in the rule table. Since KSTEM is a morphological analyzer that focuses on reducing morphological variants into a root form, the basic and most common Filipino verbs, nouns and pronouns that can be formed by applying the principle of inflectional and derivational based approach were listed in the predefined dictionary. Table 1 shows samples of the generated predefined dictionary for the pre-stemmer. 89 basic common words and 22 common Filipino clusters of natives and languages, including the inflected and derived words were listed.

Table 1. KSTEM Dictionary

Root Word (Derived)	Basic words (Inflected)
agad	agaran, agarin
bahay	namahay, namamahay, mamahay, mamamahay
bangka	namangka, namamangka
baril	namaril, namamaril
bayad	kabayaran,bayaran,binabayaran,babayaran,binayaran
bato	namato, namamato
bili	namili,namimili
bingwit	namingwit, namimingwit, mamimingwit, mamingwit
bukid	bukirin
bulaklak	namumulaklak, namulakalak, mamulaklak, mamumulaklak

VI. ASSIMILATORY AND PARTIAL DUPLICATION MODIFICATION

Tagalog words undergo assimilatory changes when a prefix pam, pan, man, mang, nam, nan, nang were found.
The conditions used are:

1) If the form starts with a consonant, then go to TagSa Stemming algorithm.
2) If the form (i.e., after removal of a prefix) starts with a vowel, then attach and substitute the conceived letter/s to the form and to the consonant/s following the first vowel of the form that is common to the prefix ending/s, respectively.

a. pam -- b/p
e.g. pamato→pam+ato /pamasok→pam +asok
b. pan -- s/t
e.g. panangga→pan+sangga / panahi→pan+ahi
c. man -- t / s
e.g.manukso→man+ukso/manigaw→man+igaw
d. mang -- k
e.g. mangantiyaw→mang+antiyaw
e.nam --- b /p
e.g. namigay→nam+igay / namili→nam+ili
f. nan --- s/t
e.g. nanulat→nan+ulat / nanahi→nan+ahi
g. nang--- k
e.g. nangulekta→nang+ulekta

The generated forms will be considered as candidates, a dictionary-search will decide on the winning candidate. However, the very first word that matches a candidate will be considered as the accepted form.
From the previous algorithm (TAGSA), partial duplication method will be strengthened, if the generated form was not found in the dictionary, the following conditions are implemented:
ex. Nangungulekta (input) → nang+ungulekta (assimilation) →kungulekta (output).
The following will be added to the algorithm for the assimilation method for the partial duplication process.
a. If the assimilation method uses a prefix look –up of “nang”, and “mang” the next three letters after substituted letter will be removed.
e.g. Nangungulekta → kungulekta→ kulekta
b. If the assimilation method uses a prefix look –up of “nam”, “nan”, ”pam” ,”man” and “pan”, the next two letters after substituted letter will be removed.
e.g. namimigay→bimigay→bigay pananahi→tanahi→tahi

VII. TEST DATA

TAGSA was evaluated using samples which were derived from three different sources, namely, the Philippine Constitution, The -09 Babilonia Wilner Foundation-Balikas (September issues) website (http://bwf.org/balikas/dati_09_03.shtml), and from a dissertation entitled “Isang Feministang Pagbasa kay B.S Edina, JR.” by Alice Gregorion-Nicolas (September, 1997).For comparative data analysis, the modified hybrid stemming algorithm approach was evaluated using samples taken from Philippine Constitution, The Bible - Book of Philippians Chapters 1-3, and El Filibusterismo.

VIII. WORD GROUPINGS

The hybrid approach manually list concept groups as patterned to the TAGSA’s method of word groupings. The generated words in the group are not necessarily semantically related.

1) aaga, maaga, pinakamaagang, umaga
2) alamin,ipagbigay-alam, kaalaman, kaalamang, kinalaman, makialam
3) naglalaman, nilalaman
4) .kapulungan,kapulungang,magpulong,nagpulong,pagpupulong,pagpupulungan, pangkapulungan, pulungan, pulungin

Sample processed words in the word groups by TAGSA yields:

5) aga, aga, aga, aga
6) alam, alam, alam, alam, alam, laman, alam

7) laman, laman
8) pulung, pulung, pulong, pulong, pulong, pulung, pulung, pulung, pulung

IX. COMPUTATION OF ERROR RATES

The hybrid stemming was evaluated based on TAGSA's computation of error rates. This focus on the error counting of word samples derived from the actual text. Under stemming Errors - words referring to the same concept are not reduced to the same stem. Over stemming Errors - words referring to a distinct concept are reduced to the same stem.

X. HYBRID STEMMING RESULTS

The modified stemming (under and over stemming) indexes result serves as an indicator of its effectiveness by having an increase in its accuracy. Table 2 shows different data result, and how the stemmer performs based on the number of word samples derived from different sources.

Table 2.
Error Stemming Result

Sources		Total Word Samples	UI	OIx10^{-5}
TAGSA	Philippine Constitution(PC)	1909	0.14471 14.47%	4.29292 0.00429%
	Social Science Dissertation(SSD)	2691	0.13465 13.47%	4.86886 (0.00487%)
	BWF-Balikas Website	1782	0.1479 14.79%	4.5476 0.00455%
Hybrid Stemming	Philippine Constitution (PC2)	2450	0.12062 12%	0.00 0%
	Bible Verse (BV)	1850	0.05434 5%	0.00 0%
	El Filibusterismo (EF)	1132	0.08196 8%	4.0384 0.00404%

For the entire six corpuses, to be compared with the TAGSA's under stemming and over stemming, the Hybrid Stemming approach indices indicate lower error rates. This shows a 17.73% decrease in under stemming index and 0.00967 over stemming index for all the combined group corpuses.

XI. POS TRAINING SETS

Previous TPOST was trained using 1,983 words with 450 distinct features, from the first three chapters of the Book of Philippians. The tagset includes 59 tags that are classified under 10 major POS tags. The tagger was tested using another text under the same domain with 539 words with 221 distinct word features, and has achieved less than 8% and 11% errors for general and specific POS tag errors, respectively.

TPOST is also tested in other corpus, the wordings from Psalm 23 (Bible), business news, entertainment news and in an essay. These test are used in the training test of the hybrid approach, since it shows significant evidence on the errors pertaining to the stemming.

XII. FEATURE EXTRACTION

Since TPOST shows a high percentage of stemming errors in part-of-speech tagging, the test data evaluation will focus on the feature extraction of derivational and inflected words specifically the assimilation conditions. Table 8 shows that an added feature code will be used to label word under assimilation conditions. This will enable assimilatory words to correctly extract its feature for the post tagging processes.

Table 3.
Modified Feature Extraction Table

Feature Code	Description
#	Predefined Word
:F	1st letter Capitalized
:FS	1st word of Sentence
*	No Features, Whole word
~	Prefix
@	Infix
+	Suffix
<>	Assimilatory condition
$	Duplicated Characters
-	Hyphen

The corpuses used in the stemming process were utilized in the POST tagging feature extraction. Test data are manually extracted and get its features. From the extracted features, words that are incorrectly extracted due to the assimilatory condition were listed. From the TPOST Training data, Old Testament, Business News, Entertainment News and in an Essay, which shows a large number of tagging error due to stemming, the number of words and assimilatory words are manually collected for the features extraction processes as seen in table 4.

Table 4.
Stemming Error and Assimilation Success Percentage

Corpus	Number of words	TPOST Stemming Error	No. of Assimilated Words	Assimilation Percentage
Old Testament (OT)	128	16	2	1.56%
Business (BN)	131	9	0	0%
Entertainment (EN)	207	16	3	1.44%
Essay (ES)	240	23	1	0.42%

The data also shows the number of percentage of correctly extracted feature from the test data, Old Testament, Business News, Entertainment News and in an Essay with the corresponding 128, 131, 207, and 240 collected words respectively. An evaluation of 1.56% of OT, 0% of BN, 1.44 of EN and 0.42% of ES are correctly extracted. A decrease in the stemming error and an increase in POS tagging accuracy are projected with the total of 3.42% assimilation success percentage.

XIII. WORD FEATURE EVALUATION

Table 5 shows the number of percentage of the correctly extracted feature from the test data, Philippine Constitution, Bible Verse and El Filibusterismo with the corresponding 2450, 1850, and 1132 collected words respectively. An evaluation of 1.81% of PC, .70% of Bible verse, and .44% of El Filibusterismo are correctly extracted its word features for the POST tagging processes.

Table 5.
Assimilation Success Percentage

Corpus	Number of words	No. of Assimilated Words	Assimilation Success Percentage
Phil. Constitution(PC)	2450	29	1.18%
Bible Verse(BV)	1850	13	.70
El Filibusterismo(EF)	1132	5	.44

From the collected corpus, which composes of 5,432 words, and 47 assimilatory words, a total of 2.32% assimilation word feature success rate was manually generated. This shows that the hybrid stemming algorithm contributes in decreasing the stemming error in TAGSA and in TPOST processes.

XIV. TRAINED DATA

Since the success of the hybrid stemming algorithm relies on the assimilatory word search, a trained data was listed. The trained data includes 45 sentences with different assimilatory words. Table 12 shows the collected assimilatory and the success rate of the feature extracted words.

Table 6.
Collected Assimilatory Words and Feature Success Rate

Corpus	No. of Sentences	No. of Words	Assimilatory Words	Assimilation Word Feature Success Rate
Trained Corpus	45	490	46	9.39%

From the collected 46 assimilatory word on the training data, which compose of trained assimilatory words, a 9.39% assimilation word feature success rate was manually generated.

XV. SUMMARY AND CONCLUSIONS

A derivational and inflectional KSTEM algorithm was employed to distress TPOST stemmer, mainly the assimilatory and partial duplication stemming inaccuracies. KSTEM, which is composed of pre-collected assimilatory words, serves as pre-stemmer that deals with the lexicon look-up process. Upon look-up, the established lexicon will then pass through feature extraction for the part of speech tagging.

A modified rule was established in assimilatory words focusing on the word prefixes. The partial duplication rule was also crafted since TAGSA shows specified handling error in this class. A modification to the previous TPOST word feature rule was added to solve limitation of previous stemmer. This modification rule was added to ascertain assimilatory word for POS tagging.

The algorithm was manually tested using the previous test data and other trained corpus. The stemming evaluation, for the entire corpuses shows that the hybrid approach presents lower error rates. A 17.73% decrease in under stemming index and 0.00967 over stemming index for all the combined group corpuses were produced. Also, a decrease in the stemming error and an increase in POS tagging accuracy were projected with the total of 3.42% assimilation success percentage for the same TPOST test corpora. A trained data was also trained, since the success of the hybrid stemming algorithm relies on the assimilatory word search. A 9.39% assimilation word feature success rate was manually generated. Thus, this shows and generally contributes in decreasing the stemming and POS tagging inaccuracy of the previous algorithms.

XVI. RECOMMENDATIONS

A further research can be performed on the Filipino word stemming and the part-of-speech tagging from the results concluded from this study. With the implementation of derivational and inflectional approach of KSTEM algorithm, a more collection of lexicon is needed to strengthen the pre-stemming phase. An innate morphological study, stemming modification or KSTEM strengthening focusing on the old Filipino assimilation of word dealing with the linking verb “ay” like nuo’y, landasi’y, and duo’y are recommended. Automatic word grouping generator is recommended to expedite future test corpus. An improvement of the feature extractor or construction of a morphological analyser dealing with the apostrophe “ ‘ ” search in assimilatory word and feature look-up of foreign entry will enhance stemming and part of speech tagging results. The algorithm adaptation in diversified domain types is recommended to generate more results in stemming and features extraction. With the presentation of the derivational and inflectional approach through the use of KSTEM algorithm, and the assimilation and partial duplication stemming modification, the execution of study through the use of existing stemming and part of speech tagging is recommended. The evaluation of other post tagging errors, which will fortify the study, is limited without the application from the previous system. This includes parameter check, scoring errors, cascading errors, word ambiguity check, and the part-of-speech tagging results.

A. Acknowledgment

The researcher would like to express sincerest gratitude and deepest appreciation to all the help and support of the following persons: To Mr. Felizardo Reyes , a responsible and understanding adviser, for giving ideas in the improvement of the study; his patience, understanding and encouragement. To Ms. Nenette Rogel and Ms. Jelisa Victorino, for sharing their knowledge and talents about the subject matter. To all friends, families, and love ones, Thank you for understanding and encouragement in many moments of crisis. Because of their unconditional love and prayers, the researcher had the chance to fulfill this study.

REFERENCES

[1] ("Master ' s Thesis Word Segmentation and Part-of-Speech Tagging for Lao Language Insisiengmay Alivanh," 2017)
[2] Wolff, J.U. (2010). Concise Encyclopedia of Languages of the World.Elsevier. pp. 1035–1038. ISBN 978-0-08-087775-4.
[3] Hualde J.I., Olarrea A., O'Rourke E. (2012). The Handbook of Hispanic Linguistics.John Wiley &Sons.p. 49.ISBN 978-1-4051-9882-0.
[4] Raga, R. Jr., &Trogo, R. (2006).Memory-Based Part-Of-Speech Tagger. De La Salle University-Manila,.
[5] Manguilimotan, E., & Matsumoto, Y. (2009). Factors Affecting Part-of-Speech Tagging for Tagalog. 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC-23), 763–770.
[6] Pisceldo, F., Adriani, M., & Manurung, R. (2009). Probabilistic Part of Speech Tagging for Bahasa Indonesia. Proceedings of the 3rd International MALINDO Workshop, Colocated Event ACL-IJCNLP.
[7] Nocon, N., & Borra, A. (2016). SMTPOST : Using Statistical Machine Translation Approach in Filipino Part-of-Speech Tagging, (Paclic 30), 391–396.
[8] Cheng, C. K., & Rabo, V. S. (n.d.). TPOST : A Template-Based , n-gram Part-Of-Speech Tagger for Tagalog.
[9] Bonus, D. E. J. (1995). The Tagalog Stemming Algorithm (TagSA), 1552(63), 63–67.