Using Stanford Part-of-Speech Tagger for the Morphologically-rich Filipino Language

Matthew Phillip V. Go
De la Salle University
Manila, Philippines
matthew.go123@gmail.com

Nicco Nocon
De la Salle University
Manila, Philippines
noconoccin@gmail.com

Abstract

This research focuses on the implementation of a Maximum Entropy-based Part-of-Speech (POS) tagger for Filipino. It uses the Stanford POS tagger – a trainable POS tagger that has been trained on English, Chinese, Arabic, and other languages and producing one of the highest results in each language. The tagger was trained for Filipino using a 406k token corpus and considering unique Filipino linguistic phenomena such as high morphology and intra-sentential code-switches. The Filipino POS tagger resulted to 96.15% tagging accuracy which currently presents the highest accuracy and with a large lead among existing POS taggers for Filipino.

1 Introduction

A Part-of-Speech (POS) tagger is a software that classifies words into its word classes or lexical categories (Bird et al., 2009). POS tags and taggers have proven its importance in Natural Language Processing (NLP) when used in advanced NLP researches such as grammar checkers (Go and Borra, 2016), information extraction (Surdeanu et al., 2011), and word-sense disambiguation (Chen et al., 2009). In a pipeline architecture of an advanced research such as an information extraction system, POS taggers are usually found in the first section producing POS tags or tag sequences. These POS tags may be used as basic features or to produce more advanced features such as syntactic structures using a constituency parser and dependencies between words using a dependency parser (Surdeanu et al., 2011; Chen and Manning, 2014).

Despite being a fundamental NLP tool towards advanced NLP researches, there seems to be few researches made towards the development of a high-performing POS tagger for Filipino, the national language of the Philippines – a Southeast Asian country with a population of 101 million people[1].

The following are the list of POS taggers developed for Tagalog, the dialect from where Filipino was based on: TPOST (Rabo and Cheng, 2006), MBPOST (Raga and Trogo, 2006), PTPOST4.1 (Go, 2006), Tag-Alog (Fontanilla and Wu, 2006), and SVPOST (Reyes et al., 2011); Adding to the list is the recently published POS tagger designed for Filipino named SMTPOST (Nocon and Borra, 2016). The key difference between Tagalog and Filipino is the presence of accepted English words such as 'cellphone', 'laptop', 'professor', 'polo shirt' as part of the Filipino language leading to nonce borrowings (single word code switching) and even intra-word code switching such as *nag-conduct* 'conducted' (added prefix *nag-*) and *tinetext* 'texting (someone)' (added infix *-in-*).

Looking into the design of the taggers, TPOST and MBPOST are closely similar because both systems utilize a lexicon list, surrounding words, capitalization, and affix features using a stemmer; where tagging rules are extracted from the training to be used during testing. PTPOST4.1 uses Hidden Markov Model (HMM), Viterbi algorithm, lexicon list, stemmer, and the previous (left) tag before the word. SVPOST makes use of Support Vector Machines (SVM) with predefined features for its training and tagging. SMTPOST presents a novel

[1] Based on Philippine Census of Population 2015

31st Pacific Asia Conference on Language, Information and Computation (PACLIC 31), pages 81–88
Cebu City, Philippines, November 16-18, 2017

approach of using Statistical Machine Translation (SMT) in tagging by 'translating' feature representation of words to POS tags. For example, the verb *kumakain* 'eating' will be represented as *@um$ka* highlighting the infix *-um-* and the partial reduplication *ka* which is then paralleled to its respective POS tag VBTR_VBAF (imperfective actor-focus verb) during training.

In terms of evaluation, an independent experiment was conducted to test the performance of the early POS taggers: TPOST, MBPOST, PTPOST4.1 and Tag-Alog using 120,000 words as data, 4% of which were used as testing data (Miguel and Roxas, 2007). The taggers scored 70%, 77%, 78.3%, and 72.5%, respectively, with PTPOST4.1 as the highest among the four. SVPOST on the other hand, conducted its own experiment on 122,318 words producing an 81% accuracy score. SMTPOST, being the most recent development among all Filipino POS taggers, produced 84.75% accuracy in its own 70,312 word dataset. These results however are relatively low compared to the state-of-the-art POS taggers for English (97.64%), French (97.8%), German (96.9%), Arabic (96.26%), and Chinese (93.46%) (Choi, 2016; Denis and Sagot, 2009; Toutanova et al. , 2003).

These low POS tagging results also hinders the progress of advanced NLP researches in the Filipino language. For instance, named entity recognition for Filipino is considered to be still in its infancy stage due to the limitation of researchers to either manually or semi-automatically tag their Filipino datasets which still requires a very tedious and time-consuming tagging or cleaning process (Lim et al., 2007).

Analysis show that works for Filipino and the other languages differ in two major factors: features and algorithms used. All of the POS taggers for Filipino uses few features: capitalization, presence of affixes, and partial/full reduplication which is produced during a pre-processing stage by handcrafted rules and a stemmer (Rabo and Cheng, 2006; Nocon and Borra, 2016). Incorrect stemming by the stemmer also cascaded as tagging errors as seen in TPOST which accounted for 25% of the tagging errors in the mentioned work. Algorithms used for Filipino which mostly relied on sentence template rules, affix features, feature-value(tag) pairs vary significantly than what algorithms the state-of-the-art POS taggers for the other languages are using: Conditional Random Fields, Maximum Entropy Cyclic Dependency Network, Maximum Entropy Markov Model, and others.

Due to significant developments in POS tagging, researches show that existing algorithms applied for these high-performing POS taggers are also usable for other languages, up to a certain extent. The Stanford POS Tagger[2], which uses maximum entropy cyclic dependency network as its core algorithm, has been applied in several languages and achieved decent tagging accuracy results: English (97.28%), Chinese (93.99%), Arabic (96.26%), French (not specified), and German (96.9%) with minimal tweaks such as character length of prefix and suffix to consider, unicode shapes for non-alphabetic languages, distributional similarity, and context window. The Stanford Part-of-Speech (POS) tagger has also been packaged in such a way that it is easy to use for training and testing custom models of different languages.

This research explores the usage of the Stanford POS tagger for the Filipino language taking into consideration the unique Filipino linguistic phenomena such as free word order structure, and a large vocabulary of root, derived, and borrowed words. This paper is organized as follows: in the next section, we discuss the Stanford POS Tagger, followed by the Filipino linguistic phenomena in Section 3; in Section 4, we describe the experiments conducted in creating a Filipino model for the Stanford POS Tagger; analysis of results are then shown in Section 5, ending the paper with the conclusion and future works in Section 6.

2 Stanford POS Tagger

The Stanford POS Tagger (SPOST), originally written by Kristina Toutanova in 2003 and maintained by the Stanford NLP Group since then, is one of the highest-performing POS tagger usable for multiple languages. It has been applied in at least four languages: English (97.28%), Chinese (93.99%), Arabic (96.26%), and German (96.9%) achieving top results for each language. The group has also released

[2] http://nlp.stanford.edu/software/tagger.shtml

the software publicly with extensive documentation written in Java discussing how fellow researchers can use the POS tagger for advanced researches or create tagging models for their target languages. The Stanford NLP community has also released packages of the POS tagger and making them usable in other programming languages (i.e. Python, PHP, Javascript, and others)[2].

The POS tagger uses maximum entropy cyclic dependency network as its core algorithm. It has been designed such that researchers can train models using different features called ExtractorFrames. Among these ExtractorFrames are tags, word shapes, unicode shapes, prefix, suffix, distributional similarity, which have shown impressive improvements when used/combined properly (Charniak et al., 1993).

3 Filipino Linguistic Phenomena

Understanding the linguistic phenomena of the Filipino language is important in determining the necessary features to be included when training a tagger model for Filipino. This section discusses the following linguistic phenomena: free-word order structure, high degree of morphology, code switches, and ambiguity of some Filipino words.

A sentence in Filipino can be written in multiple ways. For instance, the English sentence 'Juan went to the market.' can be translated as *Si Juan ay nagpunta sa palengke.* word-per-word translated as 'Juan [ay] went to market.' which follows the subject (focus)-predicate format. It can also be written in predicate-subject format *Nagpunta si Juan sa palengke.* In many cases, phrases can be re-ordered such as *Nagpunta sa palengke si Juan.* without any confusion / loss of information (Ramos, 1971).

The Filipino language has a high degree of morphology having at least 50 affix combinations, partial and full reduplication, and compound words. These morphologies are categorized into three: *inflectional*, a change in word form to represent case, gender, number, tense, person, mood, or voice such as the word *nagsisitakbuhan* 'running (present, actor focus, plural)' from the root word *takbo* 'run'; *derivational*, a change in word form that changes a word's part-of-speech (e.g. *nagsuot* 'wore (past, object focus, singular' from the root word *suot* 'clothes worn by a person'; and *compounding*, where independent words are concatenated together to form a new word (e.g. *anak* 'child' + *pawis* 'persipiration' = *anak-pawis* 'poor (noun)' (Bonus, 2003). Verb morphologies in Filipino are also complex with the different affix combinations that changes a verb's meaning, aspect (perfective, imperfective, contemplative), and focus (actor, object/goal, benefactive, locative, instrumental, and referential)[3].

Caused by past colonizations or settlements by countries such as Spain and America, the Filipino language has been greatly influenced by their languages having loanwords or Filipinized words (i.e. *bintana* 'window' from Spanish word *ventana*, *Keyk* from English word 'cake'), and having Filipinos naturally speaking English words (Americans were the last colonizers) as part of their Filipino sentences (e.g. *Computer Science ang course niya* 'His course is in Computer Science'). Additionally, rapid technology also led to more borrowed words such as 'cellphone', 'print', 'picture'. It is also common in the Filipino language to affixate English words to change its part-of-speech, for example *Phinophotoshop niya yung picture sa kanyang laptop.* 'He is editing his pictures on his laptop using Photoshop.' wherein the word 'photoshop' is affixated with a reduplication of the first syllable *Pho* and the infix *in* to denote an imperfective actor-focus verb.

Similar with English and other languages, Filipino also has its own sets of ambiguous words. Some words are ambiguous that they can be used as adjectives [JJD] or as common nouns [NNC] (i.e *balanse* 'balance' as [JJD] *balanse na buhay* 'balanced life' and as [NNC] *balanse sa buhay* 'balance in life'). Other examples include *indibidwal* 'individual' as single [JJD] or a particular person [NNC].

4 Filipino Model for Stanford POS Tagger

In creating a maximum entropy POS tagger for Filipino using the Stanford POS tagger (SPOST), features, or ExtractorFrames as Stanford calls it, that will capture unique Filipino linguistic phenomena were carefully considered and included along with the features that were commonly used in creating the other languages' tagger models: *left3words* (word and tag contexts), *naacl2003unknowns* (suffix and

[3]MGNN Tagset: `http://goo.gl/dY0qFe`

word shape feature extractors), and word shapes. Post-tagging processes were also included to augment and improve the tags provided by the Filipino tagger model and the SPOST.

For this tagger, the MGNN tagset originally presented in SMTPOST is used (Nocon and Borra, 2016)[3]. The tagset provides 230 tags consisting of 161 compound tags and 69 basic tags, revised and updated based from its predecessor, the Rabo tagset (Rabo and Cheng, 2006). The compound tags clearly present the features of the Filipino word such as: an adjective *magandang* 'beautiful' which has the ligature *-ng* with POS tag [CCP] attached to it, is denoted by [JJD_CCP] 'adjective with ligature' as compared to Rabo's tag [JJD]; and verbs' multidimensional features, where the word *kumakain* being an imperfective [VBTR] and actor-focused verb [VBAF] is denoted as [VBTR_VBAF] than Rabo's tagset that is only capable of tagging one or the other, that is either as [VBTR] or [VBAF].

To cover the high degree of morphology in the Filipino language in which prefixes, infixes, suffixes, and combination of them are evident, features extracting prefixes of length one to six for prefixes ranging from *i-* (*i-* + *tayo* 'stand up' = *itayo* 'put up') to *pinaka-* (*pinaka-* + *matalino* 'smart' = *pinakamatalino* 'smartest') and infixes with length of two for the infixes *-in-* (*-in-* + *bati* 'greet' = *binati* 'greeted') and *-um-* (*-um-* + *takbo* 'run' = *tumakbo* 'ran') were included in some tests.

A post-tagging process of overwriting POS tags of English common nouns[4] such as 'ability', 'locker', 'structure' from [NNC] 'common noun' to [FW] 'foreign word' were also included in some tests. This is done after consulting with two Filipino linguists that such words should be tagged as [FW] and not left as [NNC].

In the SPOST training tagger properties files, a number of tags from the MGNN tagset were also defined as closed class, or tag groups that have a limited number of words / symbols as its members namely: [PRS], [PRP], [PRSP], [PRO], [PRQ], [PRL], [DTC], [DTCP], [DTP], [DTPP], [LM], [CCA], [PMP], and [PMC]. Other configuration in the tagger properties file were kept similar to the configurations used in most tagger properties files of other languages trained using SPOST.

5 Results & Analysis

For this research, we used a Filipino corpus containing 15,166 sentences with a total of 406,509 tokens (54,583 of which are unique). The corpus consists of English Wikipedia sentences that were manually translated to Filipino and tagged with part-of-speech tags by Filipino linguists. The corpus has been divided into two parts: training and testing data following the 80/20 split.

For comparison of results, the Filipino tagger model in SPOST was compared with SMTPOST (Nocon and Borra, 2016) and HPOST, which is an upgraded version of SMTPOST with additional post-tagging rule-based processes. As recalled, SMTPOST is the highest-performing POS tagger for Filipino at this time of writing. All three taggers: SMTPOST, HPOST, and SPOST were trained and tested on the same corpus. Table 1 clearly shows the significant lead of the maximum entropy-based SPOST achieving 96.15% accuracy compared to the statistical machine translation-based SMTPOST's 89.11% and SMT with rule-based post-tagging HPOST's 91.63%.

POS Tagger	Accuracy
SMTPOST	89.11%
HPOST	91.63%
SPOST (best model)	96.15%

Table 1: Comparative Results of Existing POS Taggers

5.1 Finding the Best Feature Set

The best model[5] mentioned in Table 1 uses the *left5words* macro extractor frame which uses two words before, two words after, and two tags before the word to be tagged; *naacl2003unknowns* extractor frame which extracts word shape features and suffixes of the word; *word shapes(-1,1)* or the word shapes of the word before, word to be tagged, and the word after; and distributional similarity of words, which are the same set of extractor frames found in most tagger models of other languages created on SPOST. The distributional similarity was

[4] http://www.desiquintans.com/nounlist

[5] Filipino model for SPOST: https://github.com/matthewgo/FilipinoStanfordPOSTagger

trained on a Filipino Wikipedia corpus containing 17.18 million tokens. Extractor frames extracting *prefixes* of lengths one to six and *infixes* of lengths two are also included in this tagger model. These set of features allow SPOST to understand Filipino morphology and use it for tagging. Furthermore, this model uses a post-tagging process of overwriting English common nouns that were tagged as [NNC] to [FW] instead. For example, the word 'forum' if tagged as [NNC] but is in the English dictionary, the tag will be replaced into [FW].

Before achieving the best model, Table 2 shows the different models created and their corresponding accuracies sorted by the sequence of updates performed on the tagger model. Note that all the models discussed in Table 2 uses *naacl2003unknowns* and *wordshapes(-1,1)* configurations.

To begin with, it is noteworthy to discuss that the initial model using the default features: *left3words*, *naacl2003unknowns*, *wordshapes(-1,1)* and the conjugate gradient search method (*cg*) alone already scored 95.67% which is 4.04% higher than the state-of-the-art for Filipino on the same train and test data. All succeeding models however were trained using the quasi-newton search method (*owlqn2*) because of its faster training time, relatively higher accuracy, and that it is the search method used in training models of other languages for SPOST.

As seen in Table 2, series of experiments on tagger models and the improvements after inclusion or change of features are shown. The first experiment started with comparing two search methods: *cg* and *owlqn2*. After discovering that *owlqn2* performs explicitly better in terms of training speed and accuracy, the next comparison was to choose which context features to use: either *left3words* which looks at word features of the words x_{-1}, x_0, and x_1 – wherein x_0 is the word to be tagged, and x_{-1} and x_1 are its left and right adjacent words, respectively, and tags t_{-2} and t_1 as features; or extending it to *left5words* which uses the features of x_{-2}, x_{-1}, x_0, x_1 and x_2, and the same tags. The experimentation was followed by using *pref(6)* as feature and distributional similarity (*distsim*) learned from a 17.18 million word corpus. Another testing captured both prefixes *pref6* and infixes *inf2* such as *-um-* and *-in-*. Next, combined distributional similarity, prefixes, and infixes which showed higher results than the previous experimentations. Lastly, to account for the intra-sentential code switches in Filipino, an overwrite process was performed for which English common nouns[4] tagged as [NNC] into [FW]. The tagger model using *left5words*, *distsim*, *pref6*, *inf2*, and with the post-tagging process of overwriting English common nouns [NNC] as [FW] showed the highest performance among all models, achieving 96.15% accuracy – that is 78,469 out of 81,610 words were correctly tagged. With this in mind, its high accuracy shows that the tagger is significantly closer to the human's tagging reliability whose estimated error rate is at 3% (Manning, 2011).

Feature Set	Accuracy
cg-left3words	95.67%
left3words	95.80%
left5words	95.81%
left3words-pref6	95.80%
left5words-pref6	95.83%
left3words-distsim	95.89%
left5words-distsim	95.89%
left3words-pref6-inf2	95.84%
left5words-pref6-inf2	95.84%
left3words-distsim-pref6 -inf2	95.90%
left5words-distsim-pref6 -inf2	95.92%
left3words-pref6-inf2 -engNNCasFW	96.08%
left5words-pref6-inf2 -engNNCasFW	96.12%
left3words-distsim-pref6 -inf2-engNNCasFW	96.13%
left5words-distsim-pref6 -inf2-engNNCasFW	96.15%

Table 2: Results of Tagger Models

5.2 Breakdown of Errors

Adding to the results, top 10 POS tags with the highest frequency and distribution in the gold test data is shown at Table 3. The Common Noun [NNC] tag is the highest in terms of frequency and distribution with 11,015 and 13.5%, respectively; while Determiner for Common Noun (Plural) [DTCP] tag is the lowest, with 2,546 counts and 3.12% distribution.

POS Tag	Frequency	Dist. %
NNC	11,015	13.5%
NNP	7,834	9.6%
CCB	5,104	6.25%
CCT	4,952	6.07%
CCP	4,075	4.99%
DTC	3,959	4.85%
PMC	3,921	4.8%
FW	3,188	3.91%
PMP	3,039	3.72%
DTCP	2,546	3.12%

Table 3: Tags Distribution

Gold Tag	Mistagged	Recall %
NNC	366 / 11,015	96.68%
JJD	265 / 2,037	86.99%
VBW	219 / 810	72.96%
FW	167 / 3,188	94.76%
RBD	160 / 282	39.72%
JJD_CCP	155 / 1,430	89.38%
VBOF	128 / 795	71.32%
VBTS_VBOF	108 / 123	12.2%
RBW	106 / 723	85.34%
VBTS	103 / 1,730	94.05%
VBTR	103 / 1301	92.08%
RBD_CCP	95 / 230	58.7%
VBTS_VBAF	55 /59	6.78%

Table 4: Tagging Errors Breakdown

Table 4 shows the tagging errors from the test using the best tagger model, namely those POS tags that have been mistagged, the number of mistagged words and their respective recall rate – analyzed to understand the current limitations of the Filipino SPOST, and the linguistic phenomena or other reasons causing the errors. Among the list, [NNC] or common nouns have the highest number of mistagged words in terms of frequency, accounting 366 out of the total 3,141 tagging errors in the test data (11.65%). Words that should have [NNC] were mistagged as [VBW] (141), [JJD] (110), [FW] (57), and others. Interestingly, 137 out of 141 [NNC]s that were incorrectly tagged as infinitive verbs [VBW] had the prefix *pag* or *pag-* such as *pag-angkat* 'import', *pagbabago* 'change', and *pagbaha* 'flooding'. This is mainly because there are some [VBW]s that actually uses the same prefixes such as *pagbigay ng ligtas..* 'to give a safe..', and *pagkatapos* 'after finishing..' leading to confusion once detecting the prefix feature *pag* or *pag-*. On the other hand, 140 out of the 219 mistaggings of [VBW] were tagged as [NNC] and 105 of these also uses the prefix *pag* and *pag-* such as *paglalahad* 'access/approach' and *pagsabi* 'telling'.

Mistaggings [NNC] into [JJD] and vice-versa are seen in the results having 110 and 125 instances, respectively. This shows that there are Filipino words that are ambiguous and can possibly be tagged with either of the POS tags. For instance, the word *opisyal* 'official' can act as a noun such as *ang opisyal ng bayan* 'the town official' or as an adjective such as *ang opisyal na bilang* 'the official count'. The same applies to the word *bilog* 'circle' which can act as a noun or an adjective 'circular' with the same Filipino spelling.

Common nouns or [NNC]s were also mistagged as [FW]s 57 times according to the gold standard. However, results show that 27 of these 'errors' are actually English words that should be tagged with [FW], showing tagging inconsistencies by linguists whom have created the gold standard. With this in mind, tagging errors by SPOST exhibits the amount of difficulty to distinguish between [NNC] and [FW] as they are used in the same context, just that [FW] tags are borrowed English common nouns used in Filipino sentences. On the other hand, 125 out of the 167 misclassified [FW] were tagged as [NNC]; this is mainly attributed to the fact that the English nouns list used in this research only used 4,401 English common nouns, which has missed out many other English [NNC]s that can be overwritten as [FW]s. In this case, increasing the English nouns list may reduce these tagging errors.

'How' adverbs [RBD] (160) and 'how' adverbs with ligature *-ng* as suffix [RBD_CCP] (106) accounts for 8.47% (266 / 3,141) of the mistaggings in the test data. 89 of [RBD]s were mistagged as adjectives [JJD]s and 51 of [RBD_CCP] were mistagged as adjectives with ligature *-ng* [JJD_CCP]. In the English language, majority of the adverbs can be distinguished apart from adjectives by having the suffix '-ly' such as 'happily', and 'safely'. Whereas for the Filipino language, the distinction does not apply as

instances such as *galit na lumabas* 'angrily exited' and *galit na lalaki* 'angry man' uses the same word *galit* 'angry' as an adverb and adjective. But it must be noted that one can distinguish the proper POS tag by looking at t_2 – that if it is a verb or an adjective, t_0 should be [RBD], and if it is a noun, then [JJD] should be the tag. However, this research did not use the feature *bidirectional5words* which uses t_1 and t_2 instead of *left5words* due to its high memory usage requirement.

In the experiments conducted, the verb group, denoted by the POS tag prefix [VB-], accounts for 36.29% (1,140 / 3,141) of all the mistaggings in the test dataset and 832 of these are *specific* errors wherein verbs are incorrectly assigned with other verb POS tags. An important reason why such mistaggings happen is the push for the use of two dimensions in verb POS tags. In Filipino, affixes changes a verb's aspect (perfective, imperfective, and contemplative) and its focus (actor, object, locative). Combinations of certain affixes will then immediately provide a verb's aspect and focus.

One type of verb tagging error concerns with distinguishing between perfective [VBTS] and imperfective [VBTR] verbs. SPOST incorrectly tagged [VBTS] as [VBTR] 41 times and the opposite 59 times. The two tags uses almost the same set of affix sets except that [VBTR] uses a partial reduplication to denote that the action is still ongoing, as seen in the example *nagbabahagi* 'sharing' [VBTR] vs *nagbahagi* 'shared' [VBTS] from the root *bahagi* 'share', and *kinukulang* 'lacking' vs *kinulang* 'lacked' from the root *kulang* 'lack'. As part of the analysis, the partial reduplication phenomena was not included as a feature for the Filipino SPOST, which may have caused the tagging errors between the two.

Another type of verb tagging errors is about tagging verbs with two dimensions such as [VBTS_VBOF] and [VBTS_VBAF] with low recall rates at 12.2% (15 / 123) and 6.78% (4 / 59), respectively. For the case of [VBTS_VBOF], it was mistagged as [VBTS] 67 times and as [VBOF] 29 times. As for [VBTS_VBAF], it was mistagged as [VBAF] 54 times and as [VBTS] 1 time.

According to a linguist who has participated in creating the gold standard tagged data, the Filipino verb phenomena is challenging to tag. Significantly in the case of the sentence *Tumunog ang orasan ng cellphone.* 'The cellphone's alarm clock has rang.', wherein the word *tumunog* 'rang' has the infix *-um-* is usually used in a perfective actor-focus verb such as *kumain ang bata ng avocado* 'The kid ate an avocado'. As the first verb's / sentence's focus is clearly the 'alarm clock' (an object), should the verb be tagged as object-focus or as actor-focus because of the infix *-um-*? Another example are the words *tinanggihan* 'rejected' and *pinuntahan* 'went to', where both words have the same set of affixes – the infix *-in-* and suffix *-han* but they have different focuses in the examples *tinanggihan niya ang offer* 'He rejected the offer' and *pinuntahan niya ang bayan* 'He went to the city', having object and locative focus, respectively because the focus word 'offer' is an object and the word *bayan* 'town' is a location, respectively. Note that the other words in the sample sentences have the same respective set of POS tags. These examples show the difficulties in Filipino verbs which requires understanding of the semantic meaning of the nouns, root verb in tagging the focus of a verb.

Lastly, for the case of [VBTS_VBAF], it has been mistagged as VBAF 54 times out of all its 59 instances. Upon observation, the training data shows that it contains 2,352 [VBAF] tagged words and only 137 [VBTS_VBAF] words affecting the tagging results on the test data. This simply shows that the gold standard data still has room for improvement as verbs' aspect should be easy to identify.

6 Conclusion & Future Works

This research presents an implementation of a trained Filipino POS tagger using the Stanford POS tagger, which uses Maximum Entropy Cyclic Dependency Network as its core algorithm. By using built-in features such as *left5words*, *naacl2003unknown* which also captures suffixes, *wordshapes*, distributional similarity (*distsim*), adding *prefix(6)* and *prefix(2,1)* features to capture Filipino prefixes and infixes, and adding a simple POS tag overwrite function to replace English common nouns' POS tags [NNC] with foreign word [FW], the Filipino Stanford POS tagger scored 96.15% accuracy – beating other existing POS taggers even on the same train and test dataset.

Future works for improving the tagging accuracy of the developed POS tagger include experimentations on using *bidirectional5words* as feature, further cleaning of the train and test dataset, and experiments on solving the tagging difficulties on adjectives vs adverbs, verb groups, English words as [FW], and others.

Acknowledgments

This work was supported by the Department of Science and Technology (DOST) – ERDT, DOST-SEI, and Philippine Commission on Higher Education (CHED). Thanks to Nathaniel Oco and Joey Chua who have assisted us in this research.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Don Erick Bonus. 2003. The Tagalog Stemming Algorithm. Master's Thesis. De la Salle University.

Eugene Charniak, Curtis Hendrickson, Neil Jacobson, and Mike Perkowitz. 1993. Equations for Part-of-Speech Tagging. *AAAI 11*, pages 784-789.

Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser Using Neural Networks. *Proceedings of EMNLP 2014*.

Ping Chen, Wei Ding, Chris Bowes, and David Brown. 2009. A Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge. *Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 28-36.

Jinho Choi. 2016. Dynamic Feature Induction: The Last Gist to the State-of-the-Art. *Proceedings of the NAACL-HLT 2016*, pages 271-281.

Shirley Chu. 2009. Language Resource Development at DLSU-NLP Lab. The School of Asian Applied Natural Language Processing for Linguistics Diversity and Language Resource Development ADD-4: Language Resource Technology.

A. Cortez, D.J. Navarro, R. Tan, and A. Victor. 2005. PTPOST: Probabilistic Tagalog Part-of-Speech Tagger. De la Salle University.

Pascal Denis and Benot Sagot. 2009. Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. *Proceedings of the 23rd Pacific Asia Conference on Language, Information, and Computation*.

G.K. Fontanilla, and H.w. Wu. 2006. Tag-Alog: A Rule-based Part-of-Speech Tagger For Tagalog. De la Salle University-Manila.

K. Go 2006. PTPOST 4.1: Probabilistic Tagalog Part-of-Speech Tagger. De la Salle University-Manila.

Matthew Phillip Go and Allan Borra. 2016. Developing an Unsupervised Grammar Checker for Filipino Using Hybrid N-grams as Grammar Rules. *Proceedings of the 30th Pacific Asia Conference on Language, Information, and Computation (PACLIC30)*, 105-113.

L.E. Lim, J.C. New, M.A. Ngo, M.C. Sy, and N.R. Lim. 2007. A Named-Entity Recognizer for Filipino Texts. *Proceedings of the 4th National Natural Language Processing Research Symposium*.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics?. *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*.

Dalos D. Miguel and Rachel Edita O. Roxas. 2007. Comparative Evaluation of Tagalog Part-of-Speech Taggers. *Proceedings of the 4th National Natural Language Processing 2007*.

Nicco Nocon and Allan Borra. 2016. SMTPOST: Using Statistical Machine Translation Approach in Filipino Part-of-Speech Tagging. *Proceedings of the 30th Pacific Asia Conference on Language, Information, and Computation (PACLIC30)*, 391-396.

Vladimir Rabo and Charibeth Cheng. 2006. TPOST: A Template-based Part-of-Speech Tagger for Tagalog. *Journal of Research in Science, Computing and Engineering*, 3(1).

Rodolfo Raga Jr. and Rhia Trogo. 2006. Memory-based Part-of-Speech Tagger. De la Salle University-Manila

Teresita V. Ramos. 1971. *Makabagong Bararila ng Pilipino*. Rex Book Store.

C. D. E. Reyes, K. R. S. Suba, A. R. Razon, and P. C. Naval Jr.. 2011. SVPOST: A Part-of-Speech Tagger for Tagalog Using Support Vector Machines. *Proceedings of the 11th Philippine Computing Science Congress*

Mihai Surdeanu, David McClosky, Mason R. Smith, Andrey Gusev, and Christopher D.Manning. 2011. Customizing an Information Extraction System to a New Domain. *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics (RELMS 2011)*, pp. 2-10.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of HLT-NAACL 2003*, pp. 252-259.