Factors Affecting Part-of-Speech Tagging for Tagalog*

Erlyn Manguilimotan and Yuji Matsumoto

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-Cho, Ikoma, Nara, 630-0192 Japan
{erlyn-m, matsu}@is.naist.jp

Abstract. This paper investigates factors contributing to the performance of the POS Tagger for Tagalog language. Tagalog, a morphologically rich language, exhibits complex morphological structure, makes use of morphological information in determining parts of speech of the word, aspect and voice. As word feature information plays important role in efficient tagging, tag set definition capturing word information also contributes to the success or failure of the tagger. A refinement of tag set is defined to possibly improve tagging performance.

Keywords: POS Tagging, Tagalog language, Morphology , Tag set

1 Introduction

Part-of-speech (POS) tagging is one important phase in natural language processing (NLP). It is the process of labeling words in sentences with the part of speech. POS tagging is necessary in other NLP applications such as named entity recognition and syntactic analysis. Automated POS tagging approaches include rule-based approach (Brill, 1995) and probabilistic approaches such as Hidden Markov Models (HMM), Decision Trees (Schmid, 1994), Maximum Entropy Model (Ratnaparkhi, 1996), and Conditional Random Fields (Lafferty et al., 2001). Many POS tagging applications have been developed for different languages such as English, Chinese, and Japanese. A few work on POS tagging for Tagalog has been done, however, results were not sufficiently high for these systems. In the comparative study of POS taggers in Tagalog (Miguel and Roxas, 2007), the tagger using Hidden Markov Model (HMM) has an accuracy rate of 78.3%, while a template-based, N-gram POS tagger reported only 70.0% accuracy. With the aim to improve POS tagging, this paper investigates the factors affecting POS tagging in Tagalog. Considering Tagalog word's morphological structure, morphological information is used as information in training the POS tagger.

2 Tagalog Language and Morphology

This section briefly introduces Tagalog language and describes morphological structure of Tagalog words.

Tagalog is a member of the Austronesian[1] languages, and is spoken by people from the central area of Luzon in Philippines. It is the national language of the Philippines (Bonus, 2003; Mistica, M. and T. Baldwin, 2009), used as principal language in national televisions and radios.

Tagalog is more morphosyntactically complex than English language (Nelson, 2004). Its morphological nature includes affixation, stress shifting, consonant alternation, and reduplication. These morphological characteristics are used in identifying part of speech, aspect

* The authors would like to thank the Center for Language Technologies of De La Salle University, Manila, Philippines for allowing us to use the Tagalog annotated data and the Rabo tag set.

[1] Austronesian languages are language family spoken in the islands of Southeast Asia and the Pacific

and voice (Nelson, 2004). Its complex system for affixation includes prefixation, infixation, suffixation and circumfixation, where words can have two or three affixes present. Table 1 shows examples of affixation types in Tagalog. A prefixed word can be hyphenated depending on the initial letter of the root word and the final letter of the prefix (Aquino, et al. 2008). Root words beginning with vowel are hyphenated if the prefix ends with a consonant. Suffixes *~an* and *~in* are attached to words ending in consonants, and *~han* and *~hin* are used for vowel ending words. Infixation in Tagalog does not occur only inside root words but can also appear inside a prefix as demonstrated in the words ***pin****agtanim (was asked to plant)* and ***pin****apuntahan* (*was asked go and see*) in Table 1. The last row in the table shows an example of circumfixation, where a word has prefix, infix and suffix.

Another morphological characteristic of Tagalog is the use of reduplication. Reduplication can be partial, where the first syllable in the stem is repeated, such as in the word ***da****dalaw (to visit)* , or full reduplication, where the entire stem is repeated, such as in the word *araw-araw (everyday)* (Bonus, 2003). Reduplication expresses pluralization for nouns, imperfectivity for verbs, or expresses diminutiveness for adjectives.

Table 1: Types of Affixation in Tagalog

Type	Root	Affix	Affixed word
Prefixation	sama away	mag~	*mag*sama (*to live together*) *mag*-away (*to fight*)
Suffixation	kain puna	~in/hin	kain*in* (*to eat*) puna*hin* (*to mind*)
Infixation	bili sayaw sama	~in~ ~um~ ~in~	*bin*ili (*bought*) s*um*ayaw (*danced*) p*in*agsama (*joined*)
Circumfixation	punta	pa~ ~in~ ~han	*pina*punta*han* (was *asked to go*)

2.1 Verbal Focus and Aspect

Verbal focus in Tagalog defines the semantic relationship between the verb and the grammatical subject, or topic. This relationship is being shown in verbal affixes. The inflected verb indicates whether the topic is the agent or actor, object or goal, location or referent, or instrument of the action (Kess, 1979). For example, affixes like *mag~*, *~um~*, *mang~* indicate that the subject is performing the action, while affixes like *~in* (suffix), *~in~* (infix), indicate that the subject is the receiver of the action or the object. This kind of information is important in tagging verbs to be able to clearly distinguish a verb that is actor focus from an object focus verb. Kroeger (1993) referred to this verbal affix as *voice marker* and added benefactive voice to the four roles mentioned in Kess (1979).

Verbs are also inflected for three aspects, according to Schachter and Otanes (1972) as quoted by Kroeger (1993): *"Tagalog verbs are inflectable for three aspects: perfective, imperfective, and contemplated"*. Perfective aspect indicates that the action has been completed, while the imperfective aspect implies that the action has begun but not completed. The imperfective form of the verb can be used as a past progressive verb, present progressive as well as habitual actions (Kroeger, 1993). Aside from affixation, verbs also take reduplication to indicate progressive action. This characteristic defines the difference between perfective and imperfective aspects in terms of morphological structure, as in the word *laba* (*wash clothes*) and the prefix *nag* to form perfective "*nag*laba" (*washed clothes*) and imperfective "*nag***la**laba" (*washing clothes*). On the other hand, contemplated aspect refers to actions not started yet. Just

like the imperfective, this verbal aspect has partial reduplication by repeating the first syllable of the stem. However, instead of prefix *nag*, which indicates past action, imperfective aspect takes the neutral or infinitive affix *mag*. Lastly, Tagalog verbs also include aspect for recent past, or actions just *recently completed*. Verbs in recent past take the affix *ka~* and repeating the first consonant-vowel pair of the verb stem, or the first vowel, if the verb starts with a vowel , such as in t*um*ula *(recited)* → *ka*tutula *(just finished reciting)* (Santiago and Tiangco, 2003).

3 POS Tagging in Tagalog

POS tagging work for Tagalog has been done in the past. However, these taggers have low accuracy rates. As mentioned in Miguel and Roxas (2007), a template-based, N-gram POS tagger has 70% accuracy, while a rule-based tagger is 72.5% accurate, and a memory-based POS tagger is about 77% accurate. A supervised probabilistic POS tagger which made use of HMM performed better than the other taggers with 78.3% accuracy. This tagger made use of predefined words and affixes as information features during the training. The POS taggers made use of manually tagged corpora with 120,000 word tokens and punctuation marks and symbols, and made use of 65 specific tags (Miguel and Roxas, 2007). Errors were attributed to small sized corpora, stemming algorithms, and algorithms for identifying unknown words.

This paper further investigates other factors that affect POS tagging for Tagalog. It makes use of affixes and repeated syllable, if available, as feature information. To be able to use more features, conditional random fields (Lafferty et al, 2001) model is used. This model allows the use of features to condition the input values. Conditional random field (CRF) is a probabilistic framework used for labeling and segmenting sequential data (Wallach, 2004; Lafferty et al., 2001). As a conditional model, CRF specifies the probabilities of label sequences given a range of observation sequence. This conditional probability of label sequence can depend on non-independent features of the observation sequence (Lafferty et al., 2001).

4 Experiments and Discussion

4.1 Experiment Set-up

The experiments made use of manually annotated data from De La Salle University (DLSU), Philippines. The annotated data consisted of novels, news paper articles, short stories, and Bible chapter and were stored as separate files. For the training, three (3) files were put together, summing up to 114,096 tokens. Testing data consisted of small files with a total of 3,555 tokens. The tagset used was devised by DLSU with 9 coarse-grained tags, 60 specific tags, and 5 tags for punctuations and other symbols (Roxas et al. 2008). This tag set, which was based on the Penn Treebank tagset, originated from the Rabo tagset with 59 tags and was later revised to include more tags (Miguel and Roxas, 2007). CRF training and testing were done using the CRF++ Toolkit[2].

4.2 Feature Templates

Feature template is the set of information used to condition the model to predict the correct label or tag of the word. The experiments were done using different feature templates in training the CRF model. This was to show how morphological information in Tagalog may affect the POS tagging process. The trainings were done using unigram templates and bigram templates. Here, a bigram template generates a combination of features from the previous output token and the current token. Table 2 summarizes the feature information used in the training. Feature template F1 includes information of the words in its normalized form. It makes use of current word, w_c,

[2] CRF++ ToolKit by Taku Kudo, http://crfpp.sourceforge.net/

the previous two words (w_{c-1},w_{c-2}), and the next two words (w_{c+1},w_{c+2}) as features. This was done to set a reference for the next feature templates where models were trained using root word and affixes information, such as template F2, where using information from F1 template, root word and the affixes information, if available, were added. Affixes in this template were divided as prefix (pre_c), infix(inf_c), and suffix(suf_c). The third template made use of the same information in F2 and adding root and affixes of neighboring words.

With the assumption that reduplication information may contribute correct tagging, reduplication has been added for the last feature template. As mentioned in Section 2.1, Tagalog exhibits reduplication to express progressive action of a verb. Reduplication may occur in the first syllable (consonant-vowel pair) of the stem, or of the prefix (e.g. nagpa→ nagpa*pa*).

Table 2: Summary of Feature information

F1: w_c, w_{c-1}, w_{c-2}, w_{c+1}, w_{c+2}
F2: w_c, w_{c-1}, w_{c-2}, w_{c+1}, w_{c+2} , pre_c, inf_c, suf_c, r_c
F3: w_c, w_{c-1}, w_{c-2}, w_{c+1}, w_{c+2} , pre_c, pre_{c-1}, pre_{c-2}, pre_{c+1}, pre_{c+2}, inf_c, inf_{c-1}, inf_{c-2}, inf_{c+1}, inf_{c+2} , suf_c, suf_{c-1}, suf_{c-2}, suf_{c+1}, suf_{c+2}, $r_{c,}$, r_{c-1}, r_{c-2}, r_{c+1}, r_{c+2},
F4: w_c, w_{c-1}, w_{c-2}, w_{c+1}, w_{c+2} , pre_c, pre_{c-1}, pre_{c-2}, pre_{c+1}, pre_{c+2}, inf_c, inf_{c-1}, inf_{c-2}, inf_{c+1}, inf_{c+2} , suf_c, suf_{c-1}, suf_{c-2}, suf_{c+1}, suf_{c+2}, $r_{c,}$, r_{c-1}, r_{c-2}, r_{c+1}, r_{c+2}, $rpt_{c,}$, rpt_{c-1}, rpt_{c-2}, rpt_{c+1}, rpt_{c+2},

4.3 Experiment Results

Results from these experiments showed that models with unigram templates perform better than models with bigram feature templates. Table 3 shows the results of the experiments. These results are higher compared to the performances of the taggers mentioned in Section 3, where the highest accuracy rate is only around 78.3% for supervised HMM POS tagger. Adding the root and affix information of the word (feature template F2) into the training increased the performance of the POS tagger by 3.33%. However, adding affixes and roots of neighboring words did not improve tagging using unigram templates and degraded by -0.09% using bigram templates. Although reduplication information corrected some tags, this did not significantly affect the results as there are few verbs with reduplications in the test data.

Table 3: Experiment Results using Unigram and Bigram Feature Template

Feature Template	Accuracy (%)	
	Unigram	Bigram
F1	83.69	82.31
F2	87.02	86.54
F3	87.02	86.45
F4	87.21	86.89

Affix information increased the tagger's performance, however, the same information has been observed to have caused the errors. Among word categories in Tagalog, verbs, nouns and adjectives are the most inflected words. Verbs are inflected according to focus and time, while

nouns and adjectives are inflected to express their semantic information. However, Tagalog nouns, verbs and adjectives share some common affixes. Using this information resulted to some conflicts in conditioning the model. This conflict has been seen in experiments F2, F3, as well as F4.

5 Analysis and Discussions

5.1 Affixes and Ambiguity

The problem of affix information as mentioned in Section 4.3 is observed not only between word categories but also within the verbs. Aside from the morphological nature of the verb, the tag set used also contributed to the errors. The Rabo tag set has separate tags for verbal focus (actor focus, object focus, etc.) and verbal aspects (perfective, imperfective, contemplated). For example, the verb *d**um**ating (arrived)*, expresses that the subject is the *doer* of the action (actor focus), and the action has been *completed* (perfective aspect). Verbs with affix *~um~* in the train data have VBAF or actor focus tags. However, during the testing, some verbs with *~um~*, were tagged as VBTS or perfective aspects, such as in the words *p**um**ulot (picked up),* and *d**um**ilim (became dark),* or VBTR or time imperfective, such as *l**um**alabas (going out)* and *pagpup**um**ilit (forcing).* This conflict affected results in experiments F2, F3, and F4.

Another observation concerning verbs is between contemplated verbs and neutral/infinitive verbs. One possible reason for this error in tagging is that contemplated and neutral/infinitive verbs share the same set of prefixes (*mag~, ma~, mang~*), with the difference that, in contemplated aspect, verbs repeat the first consonant-vowel pair, or first vowel in the stem. We recall that in the previous section, results of models which made use of the repeated syllable information as features have made very little improvement in the results. The same problem has been observed for perfective and imperfective verbs, which share the same characteristics as contemplated and neutral/infinitive verbs for prefixes *nag~,na~, nang~.*

5.2 Other Factors

Other than affixed words, words in its basic form also caused some errors. These errors were found to be in Tagalog's nature where words have different semantic use depending on its context (Roxas et al, 2008). One of the common errors due to these ambiguities is the tag for the word *na*, which can either be an adverb enclitic or as ligatures. Enclitics *na c*onvey different nuances in meaning, which could mean *now*, *already* or *yet* as expressed in the sentence *kain **na*** (*eat now*). On the other hand, ligature *na* are used to connect a modifier and a modified word, such as *mahirap **na** buhay* (*difficult life*). Ligature *"na"* is used to connect two words if the first word ends with consonant, except *"n"*.

Modifier and modified word relationship is not only expressed using ligature "na". This relationship is also found with words (verbs, nouns, adjectives) with suffix "ng*" (magandang buhay, → beautiful life)* for words ending with a vowel, and *"g"* for words ending in *"n"*. Since the experiments only used affixes and reduplication, this feature of a word was not included, thus contributed to errors in the tagger. Lastly, aside from verbs and adjectives, errors were also attributed to nouns, specifically, proper nouns. Most proper nouns were tagged as common noun since the training did not include information, such as capitalization, to distinguish proper nouns from common nouns.

6 Features and Tag Sets

6.1 Tag set Refinement

The experiments show that morphological information can help improve tagging, especially for Tagalog, a morphologically rich language. The use of affixes improved tagging accuracy but may not be the only factor contributing to a successful POS tagging. As explained in Section 5.1, the tag set used also affected the result of tagging.

In a language where inflectional complexity is an issue in tagging, a more refined set of tag that captures this complex characteristic may be considered. Although coarse-grained set of tags may have less margin of errors and may avoid the confusions on how a word should be tagged (Mistica and Baldwin, 2009), fine-grained information can improve a tagger's performance (Dredze and Wallenberg, 2008). Tagging words to its semantic details can be useful in future syntactic-semantic analysis for the language.

The tag set in this research defines specific tags according to the features or classification of its higher level category. For example, NNC and NNP are tags for common noun and proper noun, respectively, while VBTS and VBOF are specific tags for time perfective and object focus verbs, respectively. Having two separate tags for verb features (Section 2.1) had resulted to errors in tagging (Section 5.1). Although the specific tags were defined according to some features of the word category, the tags cannot capture more refined characteristics of a word.

Table 4: Refined feature information for word category

Word Category	Feature Information
Noun (N)	Category: common, proper, abbreviation
	Common: singular, plural
	Proper: name, date, location, organization
	Gender: feminine, masculine, neuter
Pronouns (P)	Category: personal, possessive, demonstrative, interrogative, locative, comparison, indefinite, found
	Number: singular, plural
Verbs (V)	Focus: actor, object, benefactive, locative, instrumental, referential
	Aspect: infinitive, perfective, imperfective, contemplated, recent past
Adjectives (A)	Function: describe, compare, number,
	compare: same level, comparative, superlative, negation
	Number: singular, plural

Table 4 presents the feature information that can be used for each category's specific tags. This information were defined based on the existing tags in Rabo set of tags, with further refinements include other features such as name, date, location for proper nouns, number for adjectives, i.e. N-PN for proper noun and P-PRS for personal pronoun, singular. Inclusion of this information can help resolve ambiguity, and consequently may increase POS tagging accuracy. It is noticed that nouns do not include pluralization as a feature. Pluralization in Tagalog is always marked with a determiner *mga,* preceding the pluralized word. Inclusion of this information would also imply the need of a lexicon defining this information. The present work did not use lexicon in its experiments.

The tags for verbs in this refinement should be able to solve the problems of conflicting tags as observed in the experiments. Both focus and aspect features of verbs will come in one tag, e.g. V-AFP for verb, actor focus, perfective aspect. Adjective features for comparison are branched out to different kinds of comparisons (same level, comparative, superlative, and negation). The suggested refinements are only for nouns, pronouns, adjectives and verbs as the other word categories in the present tag set already defined in such fine-grained manner.

6.2 Coarse-grained Tag as Features

Recognizing the possibility that coarse-grained tag would help in correct tagging, a separate training was done to include coarse-grained tag in the feature template as one of the information. As it is not possible in the CRF tool used to predict coarse-grained tag simultaneously from the specific tags, the coarse-grained tag was given during the training. Results show that an upper bound of accuracy can be achieved with the coarse-grained information.

Feature templates *F2-Unigram*, *F4-Bigram,* and *F4-Unigram* were the only feature sets included in this trial, since the templates achieved the first 3 high scores in the previous experiments. Results are presented in Table 5.

Table 5: Results of Experiments with Coarse-grained Tag Information

Feature Sets	Without Coarse-grained Tag (From Table 3)	With Coarse-grained Tag
F2-Unigram	87.02	93.34
F4-Bigram	86.89	94.36
F4-Unigram	87.21	94.27

The increase of accuracy is attributed to correcting the tags of same affix set conflict among verbs, nouns and adjectives. It also corrected the problems between the enclitic “na” and ligatures “na” as explained in Section 5.2. However, this experiment was only done to see how inclusion of the coarse-grained POS tag could affect the CRF training, but requires further studies to simultaneously predict the coarse-grained tag given the fine-grained POS tag.

7 Conclusion and Future Work

Existing POS taggers for Tagalog has lower accuracy rates, thus a better performing POS tagger is aimed by identifying the factors contributing to successful tagging. This paper showed that using morphological information as features Tagalog POS tagger using CRF achieved higher accuracy compared to previous POS taggers presented in Miguel and Roxas (2007) discussed in Section 3. It was also shown that the tag set affected the result due to conflicts of affix information and specific tag used. Further refinements in the tag set can be done to improve the performance of the POS tagger and resolve problems of conflicting tags. A hierarchical tag set, such as in Kudo, et al (2004) can be adopted to implement the suggested POS tags presented in Table 4. Using a hierarchical tag set and defining coarse-grained POS and fine-grained POS in the tags may resolve ambiguities in tags. Lastly, we plan to work on predicting coarse-grained and fine-grained tags for Tagalog simultaneously, or in a pipeline approach.

References

Adrafre, S.F. 2005. Part of Speech Tagging for Amharic using Conditional Random Fields. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Language, pages 47-54*

Bonus, D.E. 2003. The Tagalog Stemming Algorithm. In *Natural Language Processing Research Symposium, Manila Philippines.*

Brill, E. 1995. Unsupervised Learning of Disambiguation Rules for Part Speech Tagging.[3]

Cheng, C.K. and V.S. Rabo. 2005. A POS-Tagger for Tagalog using Minimal Lexical Resources. In *Proceedings of the 5th Philippine Computing Science Congress, pages 98-103.*

Dredze, M. and J. Wallenberg. 2008. Icelandic Data Driven Part of Speech Tagging. In *Proceedings of ACL-08: HLT, Short Paper, pages 33-36*

Kess, J. 1979. Focus, Topic, and Case in the Philippine Verbal Paradigm[4].

Kroeger, P. 1993. *Phrase Structure and Grammatical Relations in Tagalog.* CLSI Publications.

Kudo, T., K.Yamamoto, and Y. Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237.*

Lafferty, J., A. McCallum and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282 -289.

McCallum, A. 2003. Efficiently Inducing Features of Conditional Random Fields. *The Nineteenth Conference on Uncertainty in Artificial Intelligence.*

Miguel, D. and R.E.O. Roxas. 2007. Comparative Evaluation of Tagalog Part-of-Speech Taggers. *Proceedings of the 4th National Natural Language Processing, Manila, Philippines.*

Mistica, M. and T. Baldwin. 2009. Recognising the Predicate-Argument Structure of Tagalog. In *Proceedings of NAACL HLT 2009: Short Papers, pages 257-260.*

Nelson, H. 2004. A Two-Level Engine for Tagalog Morphology and a Structured XML output for PC-KIMMO. Brigham Young University.[5]

Peng, F. F. Peng, and A. McCallum. 2004. Chinese Segmentation and New Word Detection using Conditional Random Fields. In *COLING 2004, pages 562-568.*

Pinto, D., A. McCallum., X. Wei and W. Bruce Croft. 2004. Table Extraction using Conditional Random Fields. In *Proceedings of the 26th annual international ACM SIGIR Conference, pages 235 – 242.*

Ratnaparhi, A. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *Conference on Emperical Methods in Natural Language Processing, pages 133-142.*

Roxas, R.E.O, A. Borra, C.Ko Cheng., N.R. Lim, E.C. Ong, and M.W.Tan. 2008. Lang Resources and Evaluation, 42: 183-195.

Santiago, A. and N. Tiangco. 2003. *Makabagong Balarilang Filipino*. Rex Printing.

Schmid, H. 1995. Probabilistic Part-of-Speech Tagging using Decision Trees. [6]

Wallach, H. 2004. Conditional Random Fields: An Introduction. *University of Pennsylvania CIS Technical Report MS-CIS-04-21*[7].

[3] http://acl.ldc.upenn.edu/W/W95/W95-0101.pdf
[4] http://sealang.net/sala/archives/pdf4/kass1979focus.pdf
[5] http://contentdm.lib.byu.edu/ETD/image/etd465.pdf
[6] http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf
[7] http://www.cs.umass.edu/~wallach/technical_reports/wallach04conditional.pdf