Comprehensive Part-Of-Speech Tag Set and SVM Based POS Tagger for Sinhala

Sandareka Fernando, Surangika Ranathunga, Sanath Jayasena, Gihan Dias
Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka
{sandarekaf,surangika,sanath,gihan}@cse.mrt.ac.lk

Abstract

This paper presents a new comprehensive multi-level Part-Of-Speech tag set and a Support Vector Machine based Part-Of-Speech tagger for the Sinhala language. The currently available tag set for Sinhala has two limitations: the unavailability of tags to represent some word classes and the lack of tags to capture inflection based grammatical variations of words. The new tag set, presented in this paper overcomes both of these limitations. The accuracy of available Sinhala Part-Of-Speech taggers, which are based on Hidden Markov Models, still falls far behind state of the art. Our Support Vector Machine based tagger achieved an overall accuracy of 84.68% with 59.86% accuracy for unknown words and 87.12% for known words, when the test set contains 10% of unknown words.

1 Introduction

Sinhala, the official language of Sri Lanka, which is used by a 16 million odd population, is a morphologically rich and highly inflected language. Sinhala belongs to Indo-Aryan family of languages and has its own alphabet. Compared to the advancement in the area of computational linguistics, Sinhala language lacks many linguistic resources, holding back natural language processing research for the same (Manamini et al., 2016; Palihakkara et al.,2015). A standard and accurate Part-Of-Speech (POS) tagger is one such basic resource.

Automatic POS tagging requires two main resources: a comprehensive tag set and a tagger. Further, a manually annotated corpus is required to train the tagger, when using supervised learning techniques. Comprehensiveness of the tag set can be defined as the ability of the tag set to represent all word classes of the language. As such any grammatically correct sentence of the language can be tagged using the tag set. Quality of the manually tagged corpus is a measurement of how accurately the words are tagged manually. Comprehensiveness of the tag set and quality of the corpus directly affect the performance of tagger.

Some research has been carried out in Sinhala POS tagging. They use the UCSC Tag Set, which has three versions. The latest version consists of 29 tags (Gunasekara & Weerasinghe, 2016). However, a closer inspection reveals that this tag set is not comprehensive. There are some word classes in Sinhala that are not covered by this tag set. As reported by Gunasekara & Weerasinghe (2016), out of the 100,000 words in the manually POS tagged corpus, 3989 words do not fall into any category of the UCSC Tag Set, which means that even manual POS tagging cannot achieve 100% accuracy. This limitation has created unnecessary ambiguities in the tagged corpus, resulting some words being tagged as unknown and some words being tagged with multiple tags in different places even when they appear in the same context with a similar meaning. In addition, this tag set is not comprehensive enough to cover the inflection based grammatical variations of Sinhala language. Sinhala noun base forms are inflected by suffixing a morpheme to indicate number, definiteness and case. Finite Verbs are inflected based on person, tense, number and gender. UCSC Tag Set does not capture such grammatical features in inflected nouns and verbs. In this research, we designed a complete, multi-level tag set[1] for Sinhala that covers all word classes and grammatical variations of Sinhala words, with the help of Sinhala language experts. The new tag set resolves the identified limitations of the previous tag set.

[1] https://github.com/sandarekaf/SinhalaPOS/blob/master/Sinhala%20POS%20tags.pdf

Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing,
pages 173–182, Osaka, Japan, December 11-17 2016.

Previous research on Sinhala POS taggers used Hidden Markov Model (HMM) based tagging models and a hybrid tagging model using HMM and morphological rules (Jayasuriya & Weerasinghe, 2013; Jayaweera & Dias, 2014; Gunasekara & Weerasinghe, 2016). The highest reported overall accuracy is 72% with 20% of average unknown words by the hybrid tagger. One reason for the low accuracy can be the ambiguities and limitations of the used tag set.

There are many other stochastic, rule based, and hybrid POS tagging techniques such as Support Vector Machines (SVM), Maximum Entropy, Brill Tagger, TnT and Neural Network based taggers that achieved higher accuracies (Ojha et al,2015; Antony & Soman, 2011; Kumar & Josan, 2010) when employed for other morphologically rich languages. Comparatively, SVM based taggers have provided promising results for POS tagging on other Indo-Aryan languages. SVM is a suitable method to use with a high dimensional feature space and it has proven results with a small training set. Given that Sinhala is an under-resourced language with a severe limitation in finding linguistic experts to prepare the annotated corpus, we opted to choose the SVM based approach.

This research is carried out as part of a larger initiative to build a computer assisted Sinhala to Tamil (another official language of Sri Lanka) translator for official documents. Based on the parent project, the corpus used for this research consists of official Sinhala documents such as official letters, annual reports and circulars. Creating a manually tagged corpus is a challenging task in Sinhala due to the lack of annotators with good knowledge in Sinhala linguistics.

Currently this study is carried out using a corpus of 70,000 words. We are continuously increasing the size of the corpus. The corpus is manually tagged using the second level of the new tag set that consists of 30 tags. The second level of the new tag set resolves the ambiguities of the previous tag set. From the third level onwards, separate tags are provided for each inflected form of a word. In Sinhala, inflecting factors are Number, Gender, Person, Animacy, Definiteness, Case and Tense. After tagging a word from a second level tag, expanding it to the third level based on above inflecting factors is straightforward and unambiguous. Thus, manually tagging at level 3 is straightforward and does not need high level of language skills. Automatic tagging at third level can also be done using a simple classification process followed by morphological analysis. Therefore, although we did not use the full tag set to annotate the corpus (due to time constraints and resource limitations), getting this done in a later stage is much straightforward.

An SVM based tagger was created using the SVM based sequential POS tagger generator provided by Giménez and Màrquez (2004a). Our tagger was successful in achieving an overall accuracy of 84.68%, with 87.12% and 59.86% accuracy for known words and unknown words, respectively. This accuracy was achieved when 10% of words are unknown words in the test set.

The rest of the paper is arranged as follows. Section two explains the current status of POS tagging research in Sinhala language and other south Asian languages. The new tag set is presented in detail in section three, explaining how it solves the problems identified in the previous tag set. Section four explains the corpus creation process. Section five explains the SVM based tagger and the selected feature set. Section six presents the experiment results and a discussion on the results. Finally section seven concludes the paper.

2 Literature Review

POS tagging related research done up to now for Sinhala language has used the UCSC Tag Set, which now has three versions. The latest version of UCSC Tag Set consists of 29 tags where 27 of them are language related tags. The remaining two tags are for Foreign Words and Symbols. Changes from the UCSC Tag Set Version 1 to Version 3 include addition of Common Noun Root tag and splitting the Verb Participle tag to four sub categories. All three versions of this tag set are hereafter collectively referred to as 'UCSC Tag Set', unless otherwise specifically referred to by the version number. UCSC Tag Set has two limitations. The main limitation is that some Sinhala words do not fall under any POS tag in the tag set, thus happens to be tagged as Unknown. Some examples are හැකි - hæki "can" ,යුතු - yuthu "should/must", නොහැකි - nohæki "cannot", කුමන - kumana "which", ඉටු - itu[2] ,සිදු - sidu[2] ,පත් - path[2] and බව - bava[2]. Such unknown words fall in to a small set of distinct categories that have special language characteristics, thus can be grouped under new tag categories. The second limitation is

[2] No comparable word in English

inflection based grammatical variations of words have not been captured in the tag set. For example, common nouns in Sinhala that get inflected based on cases (Nominative: ගස - gasa "the tree", Accusative :ගසක් - gasak "a tree", Dative :ගසට - gasata "to the tree", Genitive :ගසේ/ගසෙහි - gase/gasehi "in the tree", Instrumental :ගසෙන් - gasen "from the tree") are tagged under a single tag.

Previous research reported on creating Sinhala POS taggers (Jayasuriya & Weerasinghe,2013; Jayaweera & Dias, 2014; Gunasekara & Weerasinghe, 2016) has used the UCSC Tagged Corpus built from Sinhala newspaper articles which was tagged using the UCSC Tag Set. Jayasuriya and Weerasinghe (2013) used an HMM based statistical method to train a tagger and achieved 62% overall accuracy using an 80,000 word tagged corpus for training. Jayaweera & Dias (2014) reported an accuracy of 90% for known words. The accuracy for unknown words is not reported. Improving the previous work, Gunasekara and Weerasinghe (2016) have built a hybrid POS tagger using HMM based statistical tagging followed by a morphological rule based tag prediction technique for unknown words, and reported an overall tagging accuracy of 72% with 20% of average unknown words.

As seen above, very limited amount of research has been done on Sinhala POS tagging. Other south Asian languages (such as Hindi, Urdu, Bengali, Nepali, Bhojpuri, etc.) belonging to the same language family as Sinhala have comparably higher amount of research in the field of POS tagging (Modi & Nain, 2016; Joshi et al.,2013; Dandapat et al.,2007; Chakrabarti & CDAC, 2011; Gupta et al.,2016; Ekbal & Bandyopadhyay,2008; Shahi et al.,2013; Singh & Jha, 2015). These works have used different POS tagging techniques such as HMM, Maximum Entropy, Conditional Random Field(CRF) and SVM. Comparisons between the experimented POS tagging methodologies for south Asian languages have shown that, SVM based POS tagging method has shown promising results (Antony & Soman, 2011). For example, an SVM based tagger for Bengali has obtained an accuracy of 86.84% and found to be outperforming other tagging systems based on HMM, Maximum Entropy and CRF (Ekbal & Bandyopadhyay,2008). A POS tagger for Nepali, based on SVM has shown an accuracy of 93.27% and reported to be accurate than the TnT tagger (Shahi et al.,2013). POS tagger for Bhojpuri used an SVM based tagger with 87.67% accuracy when 3.7% words are unknown (Singh & Jha, 2015), where training is done using 10,440 tokens. Ojha et. al (2015) have done a comparison of two POS tagging methods: CRF and SVM for three Indo-Aryan languages : Hindi, Odia and Bhojpuri. Error rate in POS tagging is found to be lower in SVM for two languages except Bhojpuri. Similarly, English POS taggers based on SVM has achieved comparable results with the state of art (Giménez & Marquez, 2004b). Based on the above observations, SVM appears to be a promising option to create an accurate POS tagger.

3 Tag Set

This section describes the new comprehensive multi-level Sinhala tag set. This tag set was created based on the available tag set for Sinhala UCSC Tag Set. The UCSC Tag Set was improved based on the consultation with Sinhala language experts and some comparable tags were borrowed from the Penn Treebank tag set. The UCSC Tagged Corpus was taken as a reference to analyse language usage in creating the new tag set. The new tag set is defined in multiple levels, where in each new level, tags are divided in to sub tags based on inflecting factors or contextual definitions. The complete Sinhala POS tag set contains 148 tags. The hierarchical nature of the new tag set allows users to select the appropriate level of tagging for their application or purpose. Because of this multi-level nature, it was straightforward for us to tag the new corpus using only a 30 sub-set of this tag set.

3.1 Tags in Level One

Sinhala language has five primary top level parts of speech: Nouns (නාම - nāma), Adjectives (නාම විශේෂණ - nāma viśēṣaṇa), Verbs (ක්‍රියා - kriya), Adverbs (ක්‍රියා විශේෂණ - kriya viśēṣaṇa), and Nipāta (නිපාත).

3.2 Tags in Level Two

Each primary tag at level 1 is divided in to sub categories at level 2 based on context definitions.

Noun Categorization at Level Two

Nouns are divided in to 7 categories at secondary level based on the definition. Those are *Common Noun*, *Proper Noun*, *Pronoun*, *Noun in Compound Verb*, *Questioning Pronoun*, *Deterministic Pronoun*,

and *Question Based Pronoun*. From these, the first four tags are the most obvious and can be found in the UCSC Tag Set as well. The remaining three are newly introduced.

Common nouns in Sinhala are similar to common nouns in any other language and denote a class of objects or a concept. Similarly, *proper nouns* identify an exact entity (person, place or thing) and cannot have an indefinite form. *Pronouns*, similar to any other language, are words that can be substituted for a noun or a noun phrase. *Questioning pronouns*, a special category of pronouns, are words used to ask a question. This category is comparable to WH-Pronouns in Penn Treebank tag set. *Questioning pronoun* is a new tag introduced in the tag set. Sinhala examples for *questioning pronouns* are කුමක්ද - kumakda "what" ,කෙසේද - kesēda "how" ,කවදාද - kavadāda "when" and කොහේදීද - kohēdīda "where" . An example usage of a questioning pronoun කුමක්ද - kumakda "what" would be ඔබට අවශ්‍ය කුමක්ද? – obata avashaya kumakda? "What do you want?". As seen from the examples, all *questioning pronouns* in Sinhala end in letter 'ද'-'da'. In the UCSC annotated corpus that uses the UCSC Tag Set, these words have been broken down to two parts where the last letter 'ද'-'da' is separated. In the UCSC tagged corpus, 'ද'-'da' is tagged as particle. Tagging of the first part was also ambiguous. For example කුමක්ද - kumakda "what", is first broken up in to 'කුමක්' + 'ද', former part කුමක් - kumak "which" is tagged as Pronoun. At the same time, කුමක් - kumak "which" is tagged as unknown in some other places. We refer to this first part as *question base pronouns*. *Question base pronouns* are used to show the uncertainty of a noun/noun phrase of interest. As discussed above, *questioning pronouns* are created by adding the suffix 'ද'-'da' to *question base pronouns*. An example usage of *question base pronoun* කුමක් - kumak "what" would be ඔබ කුමක් කළේද? – oba kumak kaleda? "What did you do?". *Deterministic pronouns* are words built up from a combination of a determiner (discussed below) and a pronoun. For example සමහරෙක් - samaharek "some of them" is a word in Sinhala derived from සමහර - samahara "some", which is a determiner and දෙනෙක් - denek "them", which is a pronoun. Finally, *Noun in compound verb* is a common noun followed by a verb to build up a compound verb.

Adjective Categorization at Level Two

At the second level, adjectives are divided in to 3 categories: *adjective*, *adjectival noun* and *adjective in compound verb*. The whole purpose of an *adjective* is to describe a noun, and cannot be used as any other word type. This same tag is present in the UCSC Tag Set as well. *Adjectival noun* is a noun that acts as an adjective to describe another noun based on the context. Therefore the same word form can act as a *common noun* and *adjectival noun* based on the context. For example, පාසල් - pāsal "schools" is a *common noun* but in the phrase පාසල් වත්ත - pāsal vatta "school ground", පාසල් - pāsal "schools" is used to describe the වත්ත - vatta "ground", thus tagged as an *adjectival noun*. Another observation here is, adjectival nouns in Sinhala take the plural, base form of its related common noun. In contrast, English language uses the singular form of the common noun, even when it is used as a modifier of another noun. In the UCSC Tag Set Version 3, the base form of a common noun is identified as c*ommon noun root* that is always plural. But *Common Noun Root* can either be used as a *common noun* alone, or as an *adjectival noun*. So in our new tag set, we use two tags, a*djectival noun* or c*ommon noun* to tag common noun roots, based on the context. *Adjectival noun* is a new tag introduced in our tag set. In contrast, in the Penn Treebank tagged corpus, as well as the UCSC tagged corpus, all adjectival nouns are tagged as some variation of *common noun*. Advantage of having an *adjectival noun* is it helps to identify noun phrases that need to be treated as a single entity. Finally, *adjective in compound verb* is an adjective followed by a verb to create a compound verb.

Verb Categorization at Level Two

Verbs are divided into five sub categories, based on the definition: *verb finite*, *verb participle*, *verbal noun*, *verb non-finite* and *modal auxiliary*. *Verb finite* refers to verbs used in a sentence ending. All other verb types are used in the middle of sentences. *Verbal noun* is an inflected form of a verb that acts as a noun. *Verb participle* is another inflected form of a verb that is used in a sentence to modify a noun, noun phrase, verb or a verb phrase. Thus it plays the same role as adjective or adverb. *Verb non-finite* contains all other inflected forms of the verb that do not belong to *verb finite*, *verb participle* and *verbal noun*. Sinhala has a set of words similar to English modal verbs: හැකි - hæki "Can', යුතු - yuthu "Should/ Must", නොහැකි - nohæki "Cannot". To cover this word group, *modal auxiliary* tag is borrowed from

the Penn Treebank tag set. UCSC Tag Set has not defined this tag, and consequently, the corresponding words in the corpus have been marked as unknown.

Adverbs, similar to other languages, are words used to describe a verb, and are not divided in to sub categories at level two.

Nipātha Categorization at Level Two

Nipātha are further divided in to 8 categories: *Postposition*, *conjunction*, *particle*, *interjection*, *determiner*, *nipathana*, *case marker* and *preposition in compound verb*. *Postpositions* in Sinhala are words used after nouns, verbs and sometimes even after adjectives and adverbs to show their relationship to other words in order to build up a meaningful sentence. *Conjunctions* are words used to connect words, phrases or sentences. *Interjections* in Sinhala, similar to other languages, are words used to show the emotion or feeling. *Determiners* are words that are used before a noun to show which particular example of the noun is referred to. *Postposition*, *conjunction*, *particle*, *interjection* and *determiner* are present in the UCSC Tag Set as well. *Nipathana* is a special subset of *Nipātha*. Usually, *nipātha* words cannot be used alone. In contrast, *nipathana* can be used alone in some contexts, and can be used as a *postposition* as well if needed. Examples are ඇති - æti "enough/have/in[3]"and පුළුවන් - puluwan "able".

Sinhala nouns are morphologically inflected based on the case. A suffix is added to the noun to show the case. For animate nouns and inanimate singular nouns, suffix ට - ta is added for dative case, suffix ගේ - ge is added for genitive case and suffix ගෙන් - gen is added for instrumental case. For inanimate plural nouns, suffix වලට - valata is added for dative case, suffix වල - vala is added for genitive case and suffix වලින් - valin is added for instrumental case. According to Sinhala language rules, it is wrong to separate these case marking suffixes from the main noun (Dissanayaka, 2008; Dissanayaka, 2014). However, some Sinhala writers tend to separate this case marking suffix from the main noun. To cope with such cases, a POS tag called *case marker* is introduced. In the corpus tagged using the UCSC Tag Set, case markers have been handled in an ambiguous manner. For example, dative 'ට' case marking suffix is tagged as particle whereas dative 'වලට' case marking suffix is tagged as a noun. Finally, *preposition in compound verbs* are words that do not have a meaning by themselves but, when combined with another verb, make up a compound verb.

Compound Verbs

Three tags discussed above under Nouns, Verbs and Nipāta, deserve further discussion: *noun in compound verb*, *adjective in compound verb*, and *preposition in compound verb*. There are verbs in Sinhala that cannot be written using a single word, thus they need two words. These verb types are referred to as 'compound verbs' hereafter. Second word of a compound verb is always a verb type. Examples for such compound verbs are පාඩම් කරනවා - pāḍam karanavā "study" ,අඩු කරනවා - adu karanavā "reduce", අඩු වෙනවා - adu venavā "reducing" සිදු කරනවා - sidu karanavā "make something happen". The latter words of above examples, කරනවා - karanavā "doing" and වෙනවා venavā "happening" are verbs. When analyzing the former words, it can be a *common noun* as පාඩම් - pāḍam "lesson", an *adjective* as අඩු - adu "less/lesser" or a word that does not have a meaning on its own such as සිදු - sidu. Respectively, these three former words are tagged using *noun in compound verb*, *adjective in compound verb* and *preposition in compound verb*. *Noun in compound verb* and *adjective in compound verb* tags are present in UCSC Tag Set under the names *noun in kriya mula*, and *adjective in kriya mula*, respectively. Words that we categorize under *preposition in compound verb* are taken as unknown words in the UCSC Corpus. '*kriya mula*' in Sinhala means 'base verb' and it could be misleading. Thus we substituted the term *kriya mula* with the term c*ompound verb*.

Additional Tags at Level Two

Apart from the above discussed sub categorization of 5 primary tags, there are 5 other POS tags that are added to the tag set. These are *number*, *abbreviation*, *full stop*, *punctuation*, and *foreign word*, which are self-explanatory. Another special tag called *sentence ending* is introduced to mark all the words that end a sentence but do not belong to the category *verb finite*. In Sinhala, sentences can end in an inflected form of a *noun*, *adjective* or a *nipātha*. Examples are ගසකි - gasaki "a tree", which is a *noun*, සඳහායි - saňdahāyi "for" which is a *postposition* ,and විශේෂිතයි - viśēṣitayi "special", which is an *adjective*. In

[3] English meaning can vary based on the context

the UCSC Tag Set, these words are tagged using their original tag: *common noun*, *postposition*, and *adjective*, respectively.

These categorizations mark the second level of the tag set consisting of 30 tags.

3.3 Tags at Level Three

Nouns and verbs in Sinhala can be inflected based on number, gender, person, animacy, definiteness, case, and tense. From third level onwards, each tag at second level is further categorized based on inflection factors.

For example, *Common noun* can be inflected based on animacy (animate/ inanimate), gender (masculine/ feminine), number (singular/plural), definiteness (definite/indefinite) and case (nominative/accusative/dative/genitive /instrumental). *Finite verb* is further inflected based on person (first/second/third), tense (past/non past), number (singular/plural) and gender (masculine/ feminine).

As per the requirement, tag set can be extended up to more levels by taking the selected set of inflecting factors at each level. For example, one can do the third level classification only using animacy and gender for *common nouns*, and then further categorize each third level noun based on definiteness at fourth level. Granularity of categorization can be decided based on the requirements of the specific application. At the most fine grained level, our tag set contains a total of 148 tags.

Sentence ending tag holds the possibility of further categorization depending on its original word class such as Noun Finite, Adjective Finite and Nipātha Finite, which is not included in the current tag set and thus not used to tag the corpus.

There are some *postpositions* in Sinhala that can be inflected by suffixing a *particle*. For example, සඳහා - saňdahā "for" can be inflected as සඳහාම - saňdahāma "especially for" and සඳහාද් - saňdahāda "even for". Such inflections are not captured in this tag set.

4 Corpus Creation

This research is initiated as part of a larger project for creating a 'Sinhala to Tamil Machine assisted translation system for official documents in Sri Lanka'. Therefore, the corpus used for the research is built up using official documents used in various government organizations, such as official letters, circulars and annual reports.

A corpus of 70,000 words was created using official letters, circulars and annual reports. Corpus was manually annotated using the second level of the tag set, consisting of 30 tags. Training was given for each annotator before commencing annotation. Continuous feedback was provided for annotators to reduce errors.

Annotation was done by 7 annotators who are native Sinhala speakers. However, their Sinhala linguistic knowledge is naive. Due to the nature of their knowledge on the Sinhala language and human errors, annotators tend to make mistakes in tagging. Therefore a verification process was carried out on the initial phases of manually tagging during the training period of the annotators to overcome this limitation.

Table 1 shows the composition of the corpus in terms of frequency of frequencies of unique words. Our corpus of 70,000 words contains 12% of unique words.

No. of occurrences	1	2 - 5	6 -10	11 -50	51 - 100	100 – 200	200 – 300	300 – 400	400 - 500	> 500
% of unique words	49%	31%	7%	9%	1%	0.6%	0.2%	< 0.1%	< 0.1%	< 0.1%

Table 1: Word frequency of frequencies

Finally, it should be noted that we were able to assign a tag for each word in our corpus, unlike the corpus tagged with the UCSC Tag Set.

5 Tagger

SVM is a supervised machine learning algorithm for binary classification (Cortes & Vapnik, 1995). Given a set of training examples, where each instance is a vector in multidimensional space, SVM learns

a Maximum Margin Hyperplane that separates positive examples from negative examples. Margin is the distance from the hyperplane to the nearest positive and negative examples in the vector space.

POS tagging is a multi-class classification problem. SVM, which by default is a binary classifier, is used to solve the multi-class classification problem by taking one POS tag at a time as positive class and rest of the tags as negative. Following this technique, the sequential POS tagger generator based on SVM (Giménez and Màrquez 2004a) was used to train a POS tagging model for Sinhala.

Three feature types are considered in tagger generation: word features, POS features, and lexicalized features. Word features are word unigrams, bigrams and trigrams. POS features are POS unigrams, bigrams and trigrams. Lexicalized features used for this experiment are prefixes, suffixes and word length. Lexicalized features related to English language that are based on character capitalization are irrelevant to Sinhala language.

A centred window of size N around the word to be disambiguated is considered in feature generation. N value 7 is used for feature generation in English (Giménez & Marquez, 2004b). But it may not be optimal for a language like Sinhala, which is highly inflected. To find out the best N value for Sinhala language, an experiment was done for N= 7, 5 and 3. From the 70,000 word corpus 55,000 words were used as training data and remaining 15,000 were used for testing. The tagger uses simple left-to-right tagging, so POS tags of following words are not decided at run time. To cope with this problem, ambiguity class tags are defined for proceeding context words. Ambiguity class for a word is a concatenation of all possible POS tags for that word. Each individual tag of ambiguity classes is taken as a 'May Be' binary feature. For example, if a word has an ambiguity class NN_VV (that word can be a Noun or a Verb), then May Be features are defined as "Following class May Be NN" and "Following class May Be VV". Feature set used for window size 3 is shown in Table 2.

To check the effect of ambiguity class related features, a tagger model is recreated with optimal window size. Features related to ambiguity classes (May Be's, POS unigrams for current and next (right) tag, and POS bigrams) were removed from the feature set. This is because those features use the ambiguity class of a specific word as its POS tag if its POS tag is not yet decided at run time. Finally, to analyze the effect of lexical features (prefixes, suffixes and word length), the experiment was carried out by removing them from the feature set.

Word Unigrams	w_{-1}, w_0, w_{+1}	**Ambiguity Classes**	a_0, a_1
Word Bigrams	$(w_{-1}, w_0)(w_{-1}, w_{+1})(w_0, w_{+1})$	**May Be's**	m_0, m_1
Word Trigrams	(w_{-1}, w_0, w_{+1})	**Prefixes**	a(2), a(3), a(4)
POS Unigrams	p_{-1}	**Suffixes**	z(2), z(3), z(4)
POS Bigrams	(p_{-1}, a_{+1})	**Word Length**	L

Table 2: Feature set used in window size 3

6 Evaluation

Table 3 shows the performance of the tagger when features are generated using a centred window of size 7, 5 and 3. As observed, overall accuracy was increased when window size is reduced from 7 to 3. Sinhala words are inflected based on morphology. When compared with languages such as English, same information can be given using a lesser number of words. This may be the reason for increased accuracy when window size is decreasing. Further, the training time has drastically reduced when window size is reduced from 7 to 3. Based on these observations, window size 3 is selected for further experiments.

	N=7	**N = 5**	**N= 3**
Overall Accuracy	84.24%	84.43%	84.53%
Known Word Accuracy	86.50%	86.61%	86.78%
Unknown Word Accuracy	61.23%	62.26%	61.57%
Training Time (Sec)[4]	170	125	88
Tagging Time (Sec)	11	8	7

Table 3: Performance of SVM POS tagger for window size 7, 5 and 3

[4] Intel Core i3 CPU – 1.7GHz, RAM – 8GB

	Overall	Known word	Known unambiguous word	Known ambiguous word	Unknown word
Ambiguity class related features removed	84.68%	87.12%	91.98%	83.37%	59.86%
Lexical Features Removed	84.08%	87.13%	91.98%	83.39%	53.01%

Table 4: Accuracies with ambiguous and lexical features removed

Table 4 presents the accuracies of the tagger when ambiguous class related features and lexical features are omitted from the feature set, respectively. After removing ambiguous class related features from the features provided in Table 2, results provided an overall accuracy of 84.68% which is a further improvement. However, accuracy of unknown word tagging was reduced to 59.86%. Comparing results provided in Table 3 and Table 4, best unknown word accuracy is obtained when window size is 5 whereas best overall accuracy is obtained when window size is 3 and ambiguous class related features are omitted. Improvement in results for known word accuracy has contributed to the increase in the overall accuracy at this case. This opens up an interesting experiment to find out the reason behind unknown word accuracy decrement and known word accuracy increment, when features are generated from window of size 5 and 3, respectively.

When lexicalized features (prefixes, suffixes and word length) are removed from the feature set of window size 3, overall accuracy was reduced to 84.08%. This was due to the reduction in accuracy for unknown words to 53.01%. Therefore we can conclude that lexical features have contributed directly on determining POS tags for unknown words.

SVM based POS tagger for Sinhala was successful in obtaining an highest overall accuracy of 84.68% with known word accuracy of 87.12% and unknown word accuracy of 59.86% when test set contains 10% unknown words. Here, lexical features have helped improving the unknown word accuracy. Table 5 summarizes the feature set used to obtain the highest overall accuracy.

Word Unigrams	w_{-1}, w_0, w_{+1}	**Prefixes**	a(2), a(3), a(4)
Word Bigrams	(w_{-1}, w_0)(w_{-1}, w_{+1}) (w_0, w_{+1})	**Suffixes**	z(2), z(3), z(4)
Word Trigrams	(w_{-1}, w_0, w_{+1})	**Word Length**	L
POS Unigrams	p_{-1}		

Table 5: Feature set used in obtaining the best results

Table 6 provides tagging accuracy per each language related POS tags. *Question pronoun*, *question base pronoun*, *modal auxiliary*, *pronoun*, *case marker*, *conjunction*, *postposition*, *particle*, *determiner* and *nipathana* have obtained tagging accuracy of 90% and above. Not surprisingly, these are closed class words. As discussed before, *common nouns* and *finite verbs* are two tags that will be further categorized based on inflection factors at third level tagging. These two tags have achieved 89% and 88% accuracy respectively at second level. Since third level tagging is straightforward and unambiguous, this will contribute to an increased accuracy of the tagger even when third level tagging is done. *Preposition in compound verb*, a new tag we introduced to tag set, has achieved 80% accuracy. Words belonging to this tag were tagged as Unknown in the UCSC tag set. Thus the new addition has contributed positively to POS tagging of Sinhala. *Adjectival noun*, again introduced in our tag set, has achieved a 67% of accuracy. This is due to ambiguity when the same word is used as an *adjectival noun* and *common noun* in two contexts. *Sentence ending*, another newly introduced tag has only achieved 48% accuracy. Accuracy of tagging *Adjectival Noun* and *sentence ending* can be improved by increasing the size of the corpus and avoiding the errors in manual tagging.

Tag	Accuracy	Tag	Accuracy	Tag	Accuracy
Question Pronoun	99%	Verbal Noun	91%	Preposition in Compound Verb	80%
Question Base Pronoun	99%	Determiner	90%	Adverb	72%
Conjunction	98%	Common Noun	89%	Proper Noun	72%
Modal Auxiliary	97%	Particle	89%	Adjective in Compound Verb	67%
Pronoun	95%	Finite Verbs	88%	Adjectival Noun	55%
Postposition	93%	Verb Non Finite	88%	Noun in Compound Verb	59%
Nipathana	95%	Verb participle	87%	Sentence Ending	48%
Case Marker	94%	Adjective	82%		

Table 6: Tagging accuracy per POS tag

7 Conclusion and Future Work

This study presented a comprehensive, multi-level Sinhala POS tag set. This tag set covers most of the word classes and inflection based grammatical variations of the language. The new tag set overcomes the identified ambiguities and limitations of the UCSC Tag Set. The new tag set was designed by analysing the UCSC Tag set and the UCSC tagged corpus, which was a corpus of news articles. The new tag set is then used to tag a corpus created from official documents, a different domain, and found to be successful. Further, an SVM based approach is followed in creating an automatic tagger for Sinhala, which is found to outperform existing taggers proposed for Sinhala language up to now.

The current accuracy of the tagger can be further improved by increasing the size of the corpus. Human errors in manual tagging has contributed to a certain percentage of errors in automatic tagging. The quality of the manually tagged corpus should be verified and improved further. Moreover, the tag set should be tested with other corpora of different domains to check the validity. Finally, the tagged corpus using the new tag set should be tested with advanced NLP tasks, such as machine translation, to evaluate the correctness and effect of the new tag set and the corpus tagged with it.

Acknowledgment

This research is funded by a Short Term Research Grant from University of Moratuwa. We would like to extend our gratitude to the language experts who helped in designing the new tag set, Professor J B Dissanayaka and Professor Sandagomi Koparahewa. We also thank Mr. Pasan Dissanayaka for his help in developing the verification tools for manual POS tagging, and annotators who did the manual tagging.

References

Antony, P. J., & Soman, K. P. (2011). Parts of speech tagging for Indian languages: a literature survey. International Journal of Computer Applications (0975-8887), 34(8), 22-29.

Chakrabarti, D., & CDAC, P. (2011). Layered parts of speech tagging for Bangla. Language in India, www. languageinindia. com, Special Volume: Problems of Parsing in Indian Languages.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Dandapat, S., Sarkar, S., & Basu, A. (2007, June). Automatic part-of-speech tagging for Bengali: An approach for morphologically rich languages in a poor resource scenario. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (pp. 221-224). Association for Computational Linguistics.

Dissanayaka, J. B. (2014) Sinhala Reethiya 7 - Pada Nirmanaya. Sri Lanka: Sumitha Books.

Dissanayaka, J. B. (2008) Basaka Mahima 10; Nama Padaya. Sri Lanka: S. Godage & Brothers.

Ekbal, A., & Bandyopadhyay, S. (2008, December). Part of speech tagging in Bengali using support vector machines. In Proceedings of the International Conference on Information Technology, (pp. 106-111). IEEE.

Giménez, J., & Marquez, L. (2004a). SVMTool: A general POS tagger generator based on Support Vector Machines. In Proceedings of the 4th International Conference on Language Resources and Evaluation.

Giménez, J., & Marquez, L. (2004b). Fast and accurate part-of-speech tagging: The SVM approach revisited. Recent Advances in Natural Language Processing III, 153-162.

Gunasekara, N. A. K. B. D., & Weerasinghe, A. R. (2016). Hybrid Part of Speech Tagger for Sinhala Language. In Proceedings of the International Conference Advances in ICT for Emerging Regions (ICTer) (pp.41-48). IEEE.

Gupta, V., Joshi, N., & Mathur, I. (2016). POS tagger for Urdu using Stochastic approaches. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (p. 56). ACM.

Jayasuriya, M., & Weerasinghe, A. R. (2013). Learning a stochastic part of speech tagger for Sinhala. In Proceedings of the International Conference on Advances in ICT for Emerging Regions. (pp. 137-143). IEEE.

Jayaweera, A. J. P. M. P., & Dias, N. G. J. (2014). Hidden Markov Model Based Part of Speech Tagger for Sinhala Language. arXiv preprint arXiv:1407.2989.

Joshi, N., Darbari, H., & Mathur, I. (2013). HMM based POS tagger for Hindi. In Proceeding of the 2013 International Conference on Artificial Intelligence (pp. 341-349), Soft Computing.

Kumar, D., & Josan, G. S. (2010). Part of speech taggers for morphologically rich Indian languages: a survey. International Journal of Computer Applications, 6(5), 32-41.

Manamini, S. A. P. M., Ahamed, A. F., Rajapakshe, R. A. E. C., Reemal, G. H. A., Jayasena, S., Dias, G. V., & Ranathunga, S. (2016, April). Ananya-a Named-Entity-Recognition (NER) system for Sinhala language. In 2016 Moratuwa Engineering Research Conference (pp. 30-35). IEEE.

Modi, D., & Nain, N. (2016). Part-of-Speech Tagging of Hindi Corpus Using Rule-Based Method. In Proceedings of the International Conference on Recent Cognizance in Wireless Communication & Image Processing (pp. 241-247). Springer India.

Ojha, A. K., Behera, P., Singh, S., & Jha, G. N. (2015). Training & Evaluation of POS Taggers in Indo-Aryan Languages: A Case of Hindi, Odia and Bhojpuri. In the proceedings of 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (pp. 524-529).

Palihakkara, S., Sahabandu, D., Shamsudeen, A., Bandara, C., & Ranathunga, S. Dialogue Act Recognition for Text-based Sinhala. In Proceedings of the 12th International Conference on Natural Language Processing.

Shahi, T. B., Dhamala, T. N., & Balami, B. (2013). Support vector machines based part of speech tagging for Nepali text. International Journal of Computer Applications, 70(24), 38-42.

Singh, S., & Jha, G. N. (2015, August). Statistical Tagger for Bhojpuri (employing Support Vector Machine)s. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (pp. 1524-1529). IEEE.