ISSN 2007-9737

Semi-Automatic Alignment of Multilingual Parts of Speech Tagsets

S. Yashothara, R. T. Uthayasanker, G. V. Dias, S. Jayasena

University of Moratuwa,
Department of Computer Science and Engineering,
Sri Lanka

{yashoshan,rtuthaya,gihan,sanath}@cse.mrt.ac.lk

Abstract. We cast the problem of mapping a pair of Parts of Speech (POS) tagsets as a labelled tree mapping problem and present a general-purpose semi-automatic POS tree alignment algorithm to solve the alignment. This algorithm can be used to align two POS tagsets of different languages or the same language. We evaluate its usefulness using POS tagsets of two languages: Tamil and Sinhala. The proposed approach shows that manual effort in prior approaches is drastically reduced due to the proposed algorithm and eliminates the need to create new POS tagsets.

Keywords. Parts of speech, POS tagset mapping, POS tagset alignment, semi-automatic approach, BIS tagset, UOM tagset, Tamil NLP, Sinhala NLP.

1 Introduction

Parts of Speech (POS) is a category in which a word is assigned conforming to its morpho-syntactic functions [1]. The process of assigning the POS label to words in a given text is an important aspect of natural language processing. The initial task of any POS tagging process is choosing various POS tags that are word classes such as nouns, verbs, adjectives, etc., in a language.

The importance of POS tagging has led various researchers to work independently in developing POS tags for a language. It limited the ability to reuse tagged corpus among NLP researchers in the same language. Subsequently, there have been efforts to standardize POS tagset for a language [3]. While standardizing POS tagset for a given language, researchers also found the importance of standardizing POS tagsets for similar languages [4]. A multilingual POS agreement facilitates cross-language compatibility between different languages and ensures that common parts of different languages are tagged alike [5]. Yet, most of the tagsets capture features of a particular language, and it is not easy to tag data in other languages. The imbalance in tagsets obstructs the interoperability and reusability of tagged corpora. Furthermore, it limited the ability to reuse tagged corpus among NLP researchers in low resource languages with data shortages, especially tagged data.

POS agreement between multiple languages is useful because: (1) reusability of annotated corpora, (2) interoperability across different languages, (3) capture more detailed morphological and syntactic features of these languages, (4) achieve cross-linguistic compatibility among different languages corpora, (5) make sure the common category in different languages is tagged the same way, (6) useful for building and evaluating unsupervised cross-lingual taggers, and (7) development of multilingual corpora [4]. The POS Agreement for Multilingualism can be used for machine translation, parsing, named entity recognition, coreference resolution, sentimental analysis, question answering, and code-mixing [4]. However, alignment is still challenging due to the cost of multi-language experts, time-consuming and manual effort.

Prior efforts at the POS Agreement focused on developing a framework for standardizing POS tagsets for a given language family and mapping from different tagsets to universal sets. Despite the standardization of POS tagsets, researchers developed new and evolving tagsets by in-depth

consideration of morpho-phrasal features [6]. Therefore, aligning the already generated POS tagsets is necessary. There are some approaches to map existing tagsets to a universal tagset [1]. However, no attempt has been made to align within a language or between language tags. This paper focuses on a novel approach called 'POS tagset alignment of different languages'.

Further, it is the ever semi-automatic alignment of POS tagsets. POS alignment is the process of determining correspondences between tagsets between two languages P_1 and P_2, without creating a new tagset. POS alignment can be done in three ways: (1) equal alignment, (2) subset alignment, and (3) complex alignment. It can be useful to integrate multiple POS tagsets. POS alignment is better than POS standardization as it covers better granularity and no new tagset.

In this study, we chose Tamil and Sinhala languages, which gain importance since both languages are acknowledged as official languages in Sri Lanka. Furthermore, these efforts are gaining more reputation as these two languages are considered low resource languages. Sinhala language belongs to the Indo Aryan language family, and Tamil language belongs to the Dravidian family.

As two languages that have been associated for a long time, they share striking similarities in morphology and syntax. It makes sense for the alignment of tagsets that can utilize this similarity to facilitate mapping tagsets to each other.

Therefore, in this research, BIS tagset was selected for the Tamil language as it is the standard tagset for the Indian language. University of Moratuwa (UOM) tagset was chosen for the Sinhala language as it covers the most morpho-syntactic features. We derived a POS alignment between those tagsets using a semi-automatic approach. Semi-automated alignment was a better approach that simplified the challenges of alignment.

2 Related Work

Previous efforts at the POS Agreement focused primarily on developing a framework for POS tagset standardization of a language group and using the POS standardization guidelines to create a new standardized tagset or map from various tree-bank tagsets to a universal set.

2.1 Existing Approaches on POS Standardization

NLP researchers around the world focus on several POS standardization efforts. EAGLES guidelines [5] resulted from such an initial blow to create common standards across languages. EAGLES Guidelines provide analytical information about text language, especially for identifying morpho-phrases and syntax related to computer linguistics. In this approach, they did not create a new standardized tagset using their guidelines. It became the foundation for several other kinds of research [4, 7, 8, 9] in leveraging morpho-syntactic and syntactic features to develop common standards across multiple languages.

The LE-PAROLE project [7] established a multilingual corpus for fourteen European languages, an Extended morpho-syntactically annotated, language-specific set of features according to a common basic PAROLE tagset. MULTEXT [8] focused on multilingual tools, integration, and linguistic features, with extensions in other languages.

Still, this project also mostly focuses on European languages to make the standardization among them. However, a spin-off MULTEXT-EAST [9] gradually added morpho-syntactic descriptions of sixteen languages, including Persian or Uralic languages. The MULTEXT-EAST dataset embodies the EAGLES-based morpho-syntactic specifications, morpho-syntactic lexicons, and annotated multilingual corpora.

One of the earliest works on Indian language standardization was by Baskerville et al. in designing a common POS tagset for eight languages. Hierarchical and decomposable tagsets were used in the framework as it is a recognized method for creating a common tagset framework for multiple languages [4].

The BIS has released the Unified Parts of Speech (POS) Standard in Indian languages considering the morphologic, syntactic features of Indian languages. The top-level is subdivided into the next two levels [3]. Nitish Chandra et al. claimed that the tagset for which taggers perform best should be the standard tagset [10]. Unlike

prior efforts, designing a new common framework was not the focus of Nitish Chandra et al. [10].

POS standardization focuses on designing a common tagset framework that can exploit similarity. Mapping from the existing tagset to the standardized tagset is not considered in the above approaches. Nevertheless, there are some on mapping from different tree-bank tagsets to the universal tagset.

2.2 Existing Approaches to Mapping from Different Tree-Bank Tagsets to Universal Set

Instead of standardizing morpho-syntactic tagging, there are some efforts of mapping existing tagsets to universal tagset, which they created. A Universal Parts-of-Speech Tagset was proposed by McDonald et al. The tagset consists of twelve universal parts-of-speech categories. In addition to the tagset, they evolved a mapping from 25 different tree-bank tagsets to this universal set. As a result, this universal tagset and mapping generate a dataset consisting of common parts-of-speech for 22 different languages. When corpora with a common tagset are inaccessible, they manually define a mapping from the language or the tree bank-specific fine-grained tagset to the universal tagset [1].

Zeman and Resnik worked on Interset Project, which was used in cross-language parser adaptation [11]. In this approach, a tagset of a language is converted into the universal tagset using an encoding algorithm implemented in the support library. The above project serves as an intermediate step on the way from tagset A to tagset B. They covered twenty tagsets in ten languages.

Zeman and Resnik claim that their approach is different from McDonald et al. McDonald et al. did not need to be studied in-depth, as they removed much of the language-specific information, except for the basic parts of speech that are universally found. On the contrary, Interset eliminates as little as possible because they keep what they find anywhere. Direct conversion from one language to another language did not focus on this approach.

An international collaborative project called the "Universal Dependencies project" proposes a scheme for the treebank annotation, suitable for a wide variety of languages and assists cross-linguistic study [12]. The universal annotation guidelines were built on Google Universal Part of Speech tagset. Forty languages are covered in the current version 1.3. But in this approach also, they didn't focus on the direct conversion from one language to another language.

The majority of researchers focused on mapping several tagsets to a universal tagset using the guidelines developed. Despite the standards, researchers kept introducing tagsets, which posed key challenges for standardization using universal tagset. As POS tagsets become widely used, there is a growing need to align tagset between multiple languages and the need to align multiple tagsets to one tagset [15].

3 Background

We briefly introduce the Parts of speech tagset alignment problem in this section by adapting the knowledge from ontology and schema alignment. In the ontology alignment also, researchers matched entities to determine an alignment between different ontologies.

Nevertheless, since the direct mapping of the same labelled tagsets is impossible in all POS tagset alignment cases, this is a more challenging problem than ontology alignment. Most ontology alignment approaches are semiautomatic as they couldn't receive the best output using an automatic process. So in this paper also, the focus is based on a semi-automatic process.

The POS tagset alignment problem is to find a set of correspondences between two languages' tagsets P_1 and P_2. Because tagsets can be modeled as trees, the problem is often cast as a matching problem between such trees. A tagset tree, P, is defined as P= (V, E), where V is the set of labelled vertices representing the tags and E is the set of edges representing the relations, which is a set of ordered 2-subsets of V.

Definition 1 (Alignment, correspondence Maα). Given two tagsets P_1 and P_2, an alignment between P_1 and P_2 is a set of correspondences: $(x_a, , y_\alpha, , r)$ with $x_a \in P_1$ and $y_\alpha \in P_2$ being the two matched entities, r being a relationship holding between x_a and y_α, in this correspondence:

$M_{a\alpha}$: { $x_{a,}$, $y_{\alpha,}$, r},
x_a : { x^1_a , x^2_a , ….., x^s_a },
y_α : { y^1_α , y^2_α , ……., y^t_α },
r {=, ⊆ , ⊇, … }.

Each assignment variable $M_{a\alpha}$, M is the confidence between the alignment of two languages, and x_a is the tag from one language, and y_α is the tag from another language. Here P_1 language has 's' no of tags and P_2 language has 't' no of tags. Many possible relationships are held between x_a and $y_{\alpha,}$ but they mostly fall into equal and subsumption relationships.

An equal relationship means one language tagset can equally align with another language tagset. Sometimes a POS tag in one language may not be mapped directly to another language POS tag. This mostly occurs when a number of aspects used in the specialization of a POS tag differ between languages.

For example, the Sinhala language does not have animate/ inanimate verb categories, but Tamil does. It is also possible that a POS tag in one language does not occur in another language. In this case, we will not be able to map the POS tag at all. Every language has some specific features. But we need to map these kinds of tags as well. If we cannot find an exact match for a tag, abstract level tagsets can be aligned through the adaptation knowledge of EAGLES guidelines.

4 Approach

To agree on multiple language POS tagset, researchers adopted various strategies as discussed above. Some derived a new tagset capturing the morpho-syntactic features of some specific languages (Bureau of Indian Standard), and some mapped existing POS tagsets to a universal POS tagset. However, both approaches introduce a new POS tagset.

Unlike these prior approaches, we took an entirely new angle. We cast the problem of heterogeneity in POS tagsets as an alignment of two labelled trees and proposed a novel semi-automatic approach algorithm to solve. We evaluated our algorithm using a representative POS tagset chosen from Sinhala and Tamil languages. We chose these language pairs because (1) accessibility of the data and expertise, (2) they are low resourced languages, and (3) they are official languages in Sri Lanka, where we conduct the research.

Below, the rationales behind choosing the representative tagset from each language are described. Then, a semi-automatic POS alignment algorithm is presented.

4.1 Tagset Selection

As there are several tagsets available in each language, selections of a proper POS tagset is essential for this study. While choosing a tagset of a language, usability and standardization are considered. The following subsections describe the identified POS tagsets of Sinhala and Tamil and how the proper tagset is selected to align.

Sinhala Tagsets. There are two tagsets available for the Sinhala language, such as the University of Colombo School of Computing (UCSC) tagset, developed by University of Colombo [16] and the UOM tagset by University of Moratuwa [17]. UCSC tagset contains 29 tags. There are three versions in the UCSC tagset. UOM tagset is an extended version of the UCSC tagset by overcoming the following issues: (1) uncovered word classes in the UCSC tagset (2) multiple words the 'any' category (out of the 100,000 words in the manually POS tagged corpus, 3989 words do not fall into any category), (3) inconsistent tagging, and (4) unfocus on inflection based grammatical variations [17].

There are three levels in this tagset, following a hierarchical structure. Altogether, they came up with 148 tags. Level I contains the primary top-level part of speech. Level II tagset is generated by adding inflected forms to Level I. Level II tagset is consisted of thirty tags [17]. UOM tagset is selected for this study because of the above mentioned significant limitations in the UCSC tagset. Table 1 shows the selected UOM tagset at the second level.

Tamil Tagsets. For the Tamil language, there are plenty of tagsets. We considered nine tagsets [3, 6, 10, 18, 19, 20, 21, 22, 23, 24] before choosing an appropriate one for this study. Bureau of Indian Standards (BIS) is recommended as a common tagset for POS annotation of Indian languages. Many tags in BIS are same as LDC-IL tagset. It

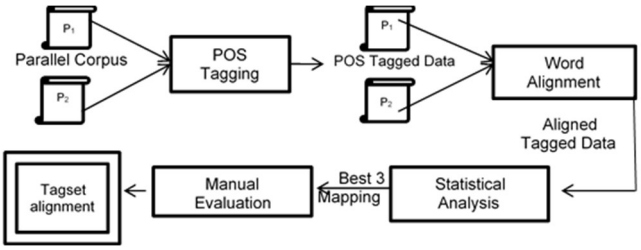

Fig. 1. Workflow of the semi-automatic POS tagsets alignment of P_1 and P_2 languages

groups unknown, Punctuation and residual into one tag. It has 11 tags in level I and 32 tags in Level II tags.

Level II is made by further subdividing the level I tags [3].

We chose BIS Tamil Tagset since it is the officially accepted standard tagset for Tamil language.

In our approach, the third level of both language tagsets is not considered. The third level captures inflection based grammatical variations of the language. We choose to omit Level III for the following reasons: (1) no apparent impact in most applications, (2) the deeper levels are inflectional forms than being POS classes, (3) more time for tagging, and (4) a large number of tags will lead to more complexity, which reduces the tagging accuracy [19].

4.2 Semi-Automatic Algorithm for POS Tagset Alignment

We proposed a semi-automatic approach for the tagsets alignments. Figure 1 describes the workflow of the semi-automatic POS tagsets alignment. The proposed semi-automatic approach requires parallel corpus. Therefore, the parallel corpus of languages P_1 & P_2 were annotated using respective automatic POS taggers.

Then the tagged parallel corpora were word-aligned using a word alignment tool. Then, the top three maps for each POS tag were selected based on the word order and presented to the human evaluators. The experts pruned the provided mappings and arrived at a final quality and complete alignment. Below we present every workflow step and tool used for this approach in a descriptive manner.

We have access to the Sinhala-Tamil parallel corpus of government official documents, which contains approximately 40,000 words. The parallel corpus was manually cleaned and aligned by three professional translators. Then, the parallel corpus was annotated using the automatic POS tagger of both languages. We used an automatic POS tagger developed by Dhanalakshmi et al. for the Tamil language as it gave higher accuracy among all available taggers. Likewise, we used an automatic POS tagger [17] based on SVM from the University of Moratuwa to annotate the Sinhala corpus.

Once the annotation was done for both the sides of the parallel corpus, parallel text was word-aligned using a word alignment tool. This study used GIZA++ [25] as a word alignment tool, giving our dataset higher accuracy. GIZA++ can perform word alignments in two directions for each pair of languages by considering one language as the source and another as the target. The intersection of both directions is taken as the resulting alignment [25].

Based on the word alignment, we retrieved the best-aligned words for the given words. It resulted in any tag of one language can be mapped to any tag of the other. There are 35 tags from the BIS tagset and 30 tags from the UOM tagset in our study. Therefore, there can be 30*35 (1050) possible alignments of tags. Further, to refine this alignment, statistical values of this mapping was considered. The highest three mappings were considered as the possible aligned tags.

The highest three mappings were derived using an automatic program by counting words belonging to each mapping. The general idea is to consider all the tag alignments of both languages generated from the GIZA++ algorithm and choose the most frequent of them as the correct alignment.

Table 1. Alignment of BIS tagset and UOM tagset

UOM Tags	BIS Tags	Example		
Common Noun	Common Noun/Echo words	மரம்	ගස	Tree
Adjectival Noun		பாடசாலை,	පාසල්	School
Case marker	Common/proper	க்கு ,உடைய	ට, ගේ	to, 's
Proper noun	Proper noun	ஜான்	ජොන්	John
Pronoun/Deterministic Pronoun	Personal Pronoun	நான், நீ	මම, ඔබ	I, you
Pronoun	Reflexive Pronoun	தான்	-	Myself
	Reciprocal Pronoun	ஒருவருக்கொருவர,அவனவன்	එක එක්කෙනාට,ඔවුනොවුන්	each other
Questioning Pronouns	Question words	என்ன, எப்படி	කුමක්ද, කෙසේද	what, how
Question-Based Pronouns	Relative Pronoun	எங்கே, எது	කොහේ, කවර	where, which
Determiners	Deictic	இவன், இவள்	මේ, සියලු	this, all
	Relative	அவ்வீடு, இவ்வீடு	ඒ ගෙදර, මේ ගෙදර	That home, this home
Verbal Participle	Verbal participle	பார்த்து	බලා	Looked
Verb finite	Verb finite	செய்தான்	කළේය	Did (he)
Preposition in compound verb		-	ඉටු, සිදු	-
Nouns in Compound Verb		படிக்கின்றான்	පාඩම් කරනවා	Study
Adjective in Compound Verbs		கூட்டபடுகின்றது	වැඩි කරනවා	Increasing
Nipathana		போதும், காணாது	ඇති, මදි	Enough/ not enough
Modal auxiliary	Verb auxiliary	முடியும், வேண்டும்	හැකි, යුතු	Can, should
Verb Non-Finite	Infinite Verb	விழ	වැටීමට වගේ	like to fall
	Conditional Verb	நடந்தால்	ඇවිද්දොත්	If walk
Verbal Noun	Verbal Gerund	படித்தல்	ඉගෙනීම	Studying
	Verbal noun	படிப்பு	-	Study
Adverb	Adverb	விரைவாக	වේගයෙන්	Fast
Adjective	Adjective	மிருதுவாக	සුමුදු	Smooth
	Relative Participle	நடந்த	ඇවිද	Walked (kid)
Conjunction	Coordinator	உம், மற்றும்	හෝ, සහ	Or, and
	Subordinator	என்று, என	යනු, යැයි	That
Particle	Default Particles	மட்டும், கூட	ය, ද, ම	Only, also
	Classifier	அட்டும்	-	-
	Intensifier	அதி, வேக, மிக	ඉතා	Most, speed
	Negation	இல்லை	නෑ, නැත	No
Interjection	Interjection	ஐயோ	අයියෝ	Oh
Postposition	Postposition	பற்றி,குறித்து	ගැන	Related
Number	Cardinal	ஒன்று, 1	එක, 1	One, 1
	Ordinal	முதல், இரண்டாம்	පළමුවන, දෙවන	First, second
Punctuation/Full stop	Punctuation	/?,:"	/?,:"	/?,:"
	Symbol	$, &,*,(	$, &,*,(	$, &,*,(
Foreign word	Foreign Residuals	கார்	කාරය	Car
Abbreviation	Unknown	மு.ப	පෙ.ව	a.m

Nevertheless, in our approach, we chose the top three frequent aligned tags and cross-checked them with bilingual experts to finalize the alignments. For example, "Nipathana" in UOM tags aligned with "Verb Finite" and "Common noun" mostly in BIS tagset. From the linguistic point of view, it does have to align with "Verb finite".

5 Results and Discussion

Through the experiment, some possible relationships are held between the BIS tagset and the UOM tagset. We reported identified four types of relationships with examples. After the manual inspection, table 1 shows the POS tagset alignment between the BIS tagset and the UOM tagset.

Two linguistics did a manual review to avoid bias. There are eight equal relationships, 22 subsumption relationships, one complex relationship and no non mapped relationships.

5.1 Equal Relationship

Some POS alignments hold an equal relationship. An equal relationship implies one language tagset can equally align with the tagset in another language. As mentioned in Table 1, some POS alignments fall under the equal relationship. The adverb in the Tamil language is directly mapped to the Sinhala language adverb node.

Modal auxiliary in UOM tagset and Verbal auxiliary in BIS tagset are equally aligned. Verbal participle, Common noun, Postpositions, Foreign words and Punctuation in both languages are fallen in an equal relationship as it has the same features. Questioning pronouns words are used to ask a question. Therefore, that is equivalently aligned with question words in the BIS tagset.

5.2 Subsumption Relationship

In most cases, a POS tag in the Sinhala language is not mapped directly to the Tamil language POS tag. Most of those tags fall under the subsumption relationship. Nipathana is a category in the Sinhala language but does not have a direct mapping tag in the Tamil language. Therefore, Nipathana has to map with the finite verb category in the Tamil language (subsumption ⊆ ⊇). A conjunction is specialized into subordinator and coordinator in the Tamil language. So these two subcategories are aligned to parent node conjunction in Sinhala language (subsumption ⊆ Relationship). It often happens when some of the features used to specialise a POS tag vary between languages.

BIS tagset does have five categories of pronouns, while there are only four categories in the UOM tagset. As a result, we are not able to equally align those tags.

The Personal, Reflexive and Reciprocal pronouns from the BIS tagset are subsumption aligned with the Pronoun tag in the UOM tagset.

Deterministic pronouns in the UOM tagset are aligned to personal pronouns in the BIS tagset. Furthermore, the category of personal pronouns can contain other words except for deterministic pronouns.

Question-based pronouns are used to show the uncertainty of a noun/noun phrase of interest. So it aligns with the Relative pronoun in the BIS tagset. But Relative pronouns can contain other words than question-based pronouns.

E.g.: *I don't know who did this.*

இதை யார் செய்தது என்று எனக்கு தெரியாது.

මෙය කළේ කවුදැයි මම නොදනිමි.

There are two types of demonstrative in the BIS tagset, while the UOM tagset has only one category. The subcategories Deictic and Relative are aligned to the Determiners tag. Particles are further divided into five subcategories in the BIS tagset, while only a parent node Particles are in the UOM tagset.

Hence, the subcategories are mapped to Particles in the UOM tagset using a subsumption relationship. General, ordinal and cardinal are the three categories of Quantifiers in the BIS tagset. Yet, the UOM tagset only have a Number category. Thus, three subcategories are aligned with the Number category.

Full stop in the UOM tagset does have a subsumption relationship with Punctuation in the BIS tagset. Like that, Symbol in the BIS tagset is aligned with the Punctuation category of the UOM tagset. As BIS tagset do not have a proper tag for Abbreviation in UOM tagset, it takes the subsumption relationship with Unknown tag. Echo

words in the BIS tagset are aligned to the Common noun in the UOM tagset.

A noun in Compound Verb is another category of noun in the Sinhala language. It is a combination of nouns and verbs. The noun, which makes a compound verb, is called as a noun in the compound verb. There is no matching translation in English and Tamil since all compound verbs in the Sinhala language is normal verb in English and Tamil. In this example, the first part of the verb is identified as 'Noun in the compound verb'. Therefore, this 'Noun in Compound verb' tag is subsumption mapped with the Finite verb tag of the BIS tagset.

E.g. එයා පාඩම් කරනවා.

He is studying.

அவன் படிக்கிறான்.

The adjectival noun is a common noun that acts as an adjective to describe another noun. When a common noun is used as an adjectival noun, it always takes the base, plural form of the common noun. For example, in a noun phrase like 'පාසල් වත්ත (school garden)', 'පාසල් (school)' is an adjectival noun that describes the main common noun 'වත්ත (garden)'. However, according to the Tamil grammar rule, if a noun expresses another noun, it cannot be categorized under the adjective category. So that 'Adjectival noun' is mapped with the common noun in the BIS tagset.

Further, adjectives are categorized into three subcategories Adjective, Adjectival Noun, and Adjective in Compound Verbs. As we saw above, the Adjectival Noun tag is aligned to the Common noun tag.

The adjective in Compound Verb is a combination of Adjective + Verb. The first word in such compound verbs will be tagged as an adjective in compound verbs. In the example' වැඩි කරනවා (increase)', වැඩි is an adjective and කරනවා is a verb. However, Tamil, we can write this as 'கூட்டபடுகின்றது'.

Hence, Tamil has no matching translation for the adjective in the compound verb since all compound verbs in Sinhala are normal in Tamil. Thus, 'Adjective in the Compound verb' is mapped with the Finite verb tag of the BIS tagset. The remaining subcategory, 'Adjective,' is aligned to the adjective in the BIS tagset.

Non-finite and finite verb forms often constitute mixed categories from the syntactic point of view. The syntactic properties of participles overlap with adjectives. Relative participle from verb category in BIS tagset also maps with the adjective in UOM tagset. Similarly, gerunds and verbal nouns BIS tagset are aligned to Verbal nouns in the UOM tagset. However, they retain their verbal arguments. Usually, these words are tagged as forms of verbs. Likewise, infinite verb and conditional verb in the BIS tagset align to the non-finite verb category in the UOM tagset.

Some other categories in UOM tagset also fall under the Verb category of the BIS tagset. Similar to 'Adjective in Compound Verb', 'Preposition in the compound verb' is one of the categories in the UOM tagset, which does not have a meaning by them but, when combined with another verb, make up a compound verb. In the example' ඉටු කරයි (does)', ඉටු is a preposition and කරයි is a verb. However, Tamil, we can write this as 'செய்கிறார்'. Hence, Tamil has no matching translation for the preposition in the compound verb, since all compound verbs in Sinhala are normal in Tamil. Thus, 'Preposition in the Compound verb' is mapped with the Finite verb tag of the BIS tagset.

Nipathana is a tag in the UOM tagset, which is used alone in some contexts and as a postposition. However, Tamil language does not have an exact match for this category. This category is mapped with the Finite verb tag by considering the usability of this category:

E.g., ඇති (*Enough*) - போதும்,

නැති (*not having*) – கிடையாது.

5.3 Complex Relationship

Some features in the POS tagset are unique to the particular language. Those features may map to another category or categories when it comes to alignment. There are some complex alignments when we try to map POS tagsets of the Sinhala and Tamil languages. Hence, we went deep into the grammar of both languages to find out the relationship for those categories.

Sinhala and Tamil nouns are morphologically inflected based on the case. The suffix is attached to the nouns to indicate the case. According to Sinhala language rules, detaching these case marking suffixes from the main noun is incorrect.

However, some Sinhala writers tend to separate this case marking suffix from the main noun. Therefore, unlike the Tamil language, the Sinhala language has space between the noun and its case marker. Subsequently, a new POS tag was added, "Case marker" in Sinhala but not Tamil. The case marker does not have an English meaning on its own. According to the previous tagset alignment in the Sinhala language, this tagset must align with a common noun or proper noun. Therefore, this alignment falls into the composite relationship.

For example, nominative form of ගස - gasa "the tree" can be inflected as ගසට - gasata "to the tree". ගසට - gasata can be written as ගසට - gasata or ගස ට - gasa ta. In the second case ට - ta has to be tagged as a case marker.

However, in the Tamil language, it will be "மரத்துக்கு" and tagged under the common noun category. This correspondence is fallen into the composite relationship.

POS alignment depicts the grammar of the language to a certain level. In addition, it is a good starting point for the study of language divergence.

6 Conclusion and Future Works

We showed that heterogeneity in POS tagsets can be cast into the labelled tree alignment problem. We presented a generic language-independent semi-automatic algorithm to align POS tagsets to provide high-quality alignment. Manual effort and time are reduced compared to previous approaches by this algorithm.

We have presented a quality alignment between the Sinhala UOM tagset and the Tamil BIS tagset. Even though these two languages have been in contact for an extended period, the grammars are not identical and have a significant difference. We listed numerous examples from real tagsets of Tamil and Sinhala languages to illustrate the most difficult parts of tagsets alignment. Each mapping includes the top three maps obtained using an automated word counting program to present the layout. However, in our approach, even though we choose the top three frequent mappings, all alignments fall within the top two mappings.

The solutions we propose follow the ultimate goal of minimizing information loss and creating a new tagset. This approach is language-independent, and we could apply for the different tagsets, which belong to a language. POS alignment is used to study the similarity and dissimilarity of grammar quantitatively. In addition, it is a good starting point for the study of language divergence.

In the future, we plan to extend this study for different tagsets, which either belong to a different language or the same language.

Acknowledgment

This project was partly supported by a Senate Research Committee (SRC) Grant funds awarded by the University of Moratuwa and funds from the Department of Official Languages of Sri Lanka.

References

1. **McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Tackstrom, O., et al. (2013).** Universal dependency annotation for multilingual parsing. 51st Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 2, pp. 92–97.
2. **Hardie, A. (2004).** The computational analysis of morphosyntactic categories in Urdu. PhD thesis Lancaster University.
3. **Bureau Indian Standard. (n.d.).** Unified parts of speech (POS) standard in Indian languages. Ministry of Communications & Information Technology Govt. of India.
4. **Baskaran, S., Bali, K., Bhattacharya, T., Bhattacharyya, P., Jha, G. N., Rajendran, S., Saravanan, K., Sobha, L. (2008).** Designing a common POS-Tagset framework for Indian languages. 6th Workshop on Asian Language Resources, pp. 89–92.

5. **Leech, G., Wilson, A. (1996).** Recommendations for the morphosyntactic annotation of corpora: EAGLES document EAG-TCWG-MAC/R. Expert Advisor Group on Language Engineering Standard.
6. **Selvam, M., Natarajan, A. M. (2009).** Improvement of rule based morphological analysis and POS tagging in Tamil language via projection and induction techniques. International journal of computers, Vol. 3, No. 4, pp. 357–367.
7. **Suzanne, N., Volz, L. (1996).** Multilingual corpus tagset specifications. MLAP PAROLE 63œ386 WP 4.4.
8. **Ide, N., Véronis, J. (1994).** Multext: Multilingual text tools and corpora. 15th International Conference on Computational Linguistics, Vol. 1, pp. 588–592. DOI: 10.3115/991886.991990.
9. **Erjavec, T. (2004).** Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. 4th International Conference on Language Resources and Evaluation, pp. 1535–1538.
10. **Chandra, N., Kumawat, S., Srivastava, V. (2014).** Various tagsets for Indian languages and their performance in part of speech tagging. 5th IRF International Conference.
11. **Zeman, D. (2008).** Reusable tagset conversion using tagset drivers. Sixth International Conference on Language Resources and Evaluation (LREC'08), pp. 28–30.
12. **Sulubacak, U., Gokirmak, M., Tyers, F., Coltekin, C., Nivre, J., Eryigit, G. (2016**). Universal dependencies for Turkish. 26th International Conference on Computational Linguistics: Technical Papers, pp. 3444–3454.
13. **De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. M., Manning, C. D. (2014).** Universal Stanford dependencies: a cross-linguistic typology. 9th International Conference on Language Resources and Evaluation (LREC), pp. 4585–4592.
14. **Tsarfaty, R. (2013).** A unified morpho-syntactic scheme of Stanford dependencies. 51st Annual Meeting of the Association for Computational Linguistics, Vol. 2, pp. 578–584.
15. **Doan, A., Halevy, A. Y. (2005).** Semantic-Integration Research in the Database Community A Brief Survey. AI Magazine, Vol. 26, No. 1. DOI: 10.1609/aimag. v26i1.1801.
16. **Gunasekara, D., Welgama, W. V., Weerasinghe, A. R. (2016**). Hybrid part of speech tagger for Sinhala language. 16th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 041–048. DOI: 10.1109/ICTER.2016.7829897.
17. **Fernando, S., Ranathunga, S., Jayasena, S., Dias, G. (2016).** Comprehensive part-of-speech tagset and SVM Based POS tagger for Sinhala. 6th Workshop on South and Southeast Asian Natural Language Processing, pp. 173–182.
18. **Central Institute of Indian Languages. (n.d.).** Ministry of Education Government of India. http:///www.ciil.org/.
19. **Dhanalakshmi, V., Padmavathy, P., Anan-Kumar, M., Soman, K. P., Rajendran, S. (2009).** Chunker for Tamil. International Conference on Advances in Recent Technologies in Communication and Computing, pp. 436–438. DOI: 10.1109/ ARTCom.2009.191.
20. **Dandapat, S. (2010).** MSRI Part-of-Speech Annotation Interface.
21. **Lakshmana, P. S., Geetha, T. (2008).** Morpheme based language model for Tamil part-of-speech tagging. Polibits, No. 38, pp. 19–25.
22. **Ramanathan, M., Chidambaram, V., Patro, A. (n.d).** An attempt at multilingual POS tagging for Tamil. Available from http://pages.cs. wisc.edu/~madhurm/CS769_final_report.pdf.
23. **Shiva IIIT (2018).** A Parts-of-Speech tagset for Indian languages. http://shiva.iiit.ac.in/ SPSAL2007/iiit_tagset_guidelines.
24. **Ramasamy, L., Žabokrtský, Z. (2014).** Tamil dependency treebank v0.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
25. **Och, F. J. (2003).** Minimum error rate training in statistical machine translation. 41st Annual Meeting on Association for Computational Linguistics, pp. 160–167. DOI: 10.3115/1075096.1075117.

26. Zeman, D. (2010). Hard problems of tagset conversion. 2nd International Conference on Global Interoperability for Language Resources, pp. 181–185.

27. Leech, G. (1997). Grammatical tagging. **Garside, Leech and McEnery, eds.**, Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman.

28. Petrov, S., Das, D., McDonald, R. (2011). A universal part-of-speech tagset. arXiv:1104.2086v1. DOI: 10.48550/arXiv.1104.2086.

29. Zeman D. (2004). Parsing with a statistical dependency model. PhD thesis, Univerzita Karlova v Praze.

Article received on 12/09/2018; accepted on 07/01/2021. Corresponding author is S. Yashothara.