

Semi-Automatic Alignment of Multilingual Parts of Speech Tagsets

S.Yashothara, R.T.Uthayasanker, W.S.N.Dilshani, G.V. Dias, S. Jayasena, S. Ranathunga,

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

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

Abstract. We cast the problem of mapping a pair of Parts of Speech (POS) tagsets as a labeled 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 of same language. We evaluate its usefulness using POS tagsets of 2 languages, Tamil and Sinhala and provide the alignment between these languages. The proposed approach shows that manual effort in prior approaches is drastically reduced due to the proposed algorithm and also eliminates the need of creating 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 to which a word is assigned in conformity with its morphosyntactic 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 to choose various POS tags which are word classes such as noun, verb, adjective, etc in a language.

The importance of POS tagging inspired various researchers to work independently in developing POS tagsets for a language. This limited the reusability of tagged corpus among NLP researchers of 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 that it is important to standardize POS tagsets for similar languages [4]. Having a POS agreement between multiple languages help cross-linguistic compatibility between different language corpora and guarantee that common categories of different languages are tagged in the same way [5]. Yet, most of the tagsets capture features of a particular language and hard for tagging data in other languages. This imbalance in tagsets obstructs interoperability and reusability of tagged corpora. This furthermore limited the reusability of tagged corpus among NLP researchers in low resource languages where scarcity of data,

particularly tagged data. POS agreement between multiple languages succors, 1. Re-usability of annotated corpora 2. Interoperability across different languages 3. Capture more detailed morphologic syntactic features of these languages 4. Achieve cross-linguistic compatibility between different language corpora 5. Guarantee that common category of different languages is tagged in the same way 6. Useful for building and evaluating unsupervised and cross-lingual taggers 7. Development of multilingual corpora [4]. POS agreement for multi languages can be used in applications like machine translation, building parse trees, Named Entity Resolution, Coreference Resolution, Sentiment Analysis, Question Answering in multiple languages and Code-mixing where it is dealt with two or more languages [4]. But, crucial challenges for POS agreement are the cost of multi-language experts, time consuming and need more manual effort.

Prior efforts on POS agreement primarily focused on developing a framework for standardizing POS tagsets of a given family of languages and mapping from different tagsets to universal set. Despite the standardization of POS tagsets, researchers kept developing new POS tagsets and evolving POS tagsets by considering morphosyntactic features deeply [6]. Therefore, a necessity of aligning the already generated POS tagsets for a language and in the midst of languages has been created. There are some approaches to map existing tagsets to a universal tagset [1]. However, there has not been any effort to establish an alignment within a language or between languages' tagsets. 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 tag sets between two languages P1 and P2 without creating new tagset. A set of correspondences is also called an alignment. POS alignment can be done in three ways 1. Equal alignment 2. Subset alignment 3. Complex alignment. It can be useful to integrate multiple POS tagsets. POS alignment is better than POS standardization because of better granularity and creation of new tagset is not essential.

In this study, we choose Tamil and Sinhala languages which gain importance since both of them are acknowledged as official languages of Sri Lanka. Further, since these two languages are considered low resource languages, these efforts gain more importance. The Sinhala language belongs to the Indo Aryan language family and the Tamil language belongs to the Dravidian family. As two languages that have been in contact for a long period of time, they share notable resemblances in morphology and syntax. This makes it sensible to align tagset that can exploit this similarity to facilitate the mapping of different tagsets to each other. So in this research, BIS tagset has been selected for the Tamil language as it is the standardized tagset for Indian languages and University of Moratuwa (UOM) tagset has been selected for the Sinhala language as it covers more morphosyntactic features. We derived a POS alignment between those tagsets using a semi-automatic approach. The semi-automatic approach was a better approach which addresses the issues such as the high cost of manual process and difficulty of finding multilingual experts. Automatic POS tagger has been used in both languages to tag the words. Then using word alignment tool, alignment between those words has fetched. After that manual evaluation has happened.

2 Related Work

Prior efforts on POS agreement predominantly focused on either developing framework about how to standardize POS tagsets of a set of languages and using the guidelines of POS standardization to create a new standardized tagset or mapping from different tree-bank tagsets to universal set. Below, we present the literature review of both approaches.

2.1 Existing Approaches on POS Standardization

There are several POS standardization efforts carried by NLP researchers around the world. EAGLES guidelines [5] were an outcome of such an initial blow to create standards that are common across languages. The EAGLES Guidelines yield governance for analytic information about the language of a text, particularly for identifying morphosyntactic and syntactic features relevant in computational linguistics. In this approach, they did not create newly standardized tagset using the guidelines they gave. This became the foundation for several other researches ([4], [7], [8], [9]) in leveraging morphosyntactic and syntactic features to develop common standards across multiple languages.

LE-PAROLE project [7] formed a multilingual corpus for fourteen European languages; morphosyntactically annotated according to a common core PAROLE tagset, extended with a set of language-specific features. MULTTEXT [8] focused on tools, corpora and linguistic features for multi-languages, with the extension of other languages. But this project also mostly focuses on European languages to make the standardization among them. However, a spin-off MULTTEXT-EAST [9] gradually added morphosyntactic descriptions of sixteen languages, including Persian or Uralic languages. The MULTTEXT-EAST dataset embodies the EAGLES-based morphosyntactic specifications, morphosyntactic lexicons, and annotated multilingual corpora.

Early works on POS standardization were predominantly on European languages. One of the early works on standardizing Indian languages was by, Baskaran et al. who have focused on designing a common POS tagset framework for eight Indian languages by considering equivalent morphosyntactic phenomena consistently across all 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 Unified Parts of Speech (POS) Standard in Indian Languages with the consideration of morphologic syntactic features of Indian languages. According to the morphological features, the top level is subdivided into next two levels [3].

Nitish Chandra et al. claimed that the tagset for which taggers perform best should be the standard tagset to be followed, and sought for the POS tagset which yields the highest accuracy during the automatic POS tagging for a set of Indian languages [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 existing tagset to the standardized tagset is not considered in the above approaches. But there are some on mapping from different tree-bank tagsets to the universal tagset.

2.2 Existing Approaches On Mapping From Different Tree-Bank Tagsets To Universal Set

Instead of standardizing morphosyntactic tagging, there are some efforts of mapping existing tagsets to universal tagset which they created. A Universal Part-of-Speech Tagset was proposed by McDonald et al. The tagset consists of twelve universal part-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 consisted of common parts-of-speech for 22 different languages. When corpora with 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 used in cross-language parser adaptation [11]. In this approach, a tagset of a language is converted into the universal tagset using 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 have covered 20 tagsets in 10 languages. Zeman and Resnik claimed that their approach differs from Google universal tagset approach as McDonald et al. didn't want to learn the details of existing tagsets more deeply because they eliminate most of the language-specific information, except for the core parts of speech that they find universally. In contrary, Interset eliminates as little as possible because they kept what they find anywhere. Direct conversion from one language to another language didn't focus on this approach.

An international collaborative project called "Universal Dependencies project" proposes a scheme for the treebank annotation, which is suitable for a wide variety of languages and assists cross-linguistic study [12]. The universal annotation guidelines which are built on Google Universal Part of Speech tagset for POS, the Interset framework for morphosyntactic features and Stanford Dependencies were created by them ([13]-[14]) for dependency relations [12]. 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.

Majority of researchers have 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 for aligning tagset between multiple languages and need of aligning multiple tagsets to one tagset [15].

3 Background

We briefly introduce the Parts of speech tagset alignment problem in this section by adapting of knowledge from the ontology alignment and schema alignment. In the

ontology alignment also, researchers matched entities to determine an alignment between different ontologies. But, since direct mapping of same labeled tagsets is not possible in all cases of POS tagset alignment, this is more challenging problem compare to ontology alignment. Most of ontology alignment approaches are semi-automatic as they couldn't receive the best output by using automatic process. So in this paper also, the focus is based on 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 labeled 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 $M_{\alpha\alpha}$). Given two tagsets P_1 and P_2 , an alignment between P_1 and P_2 is a set of correspondences: (x_a, y_a, r) with $x_a \in P_1$ and $y_a \in P_2$ being the two matched entities, r being a relationship holding between x_a and y_a , in this correspondence.

$$\begin{aligned} M_{\alpha\alpha}: & \{ x_a, y_a, r \} \\ x_a : & \{ x_a^1, x_a^2, \dots, x_a^s \} \\ y_a : & \{ y_a^1, y_a^2, \dots, y_a^t \} \\ r : & \{ =, \subseteq, \supseteq, \dots \} \end{aligned}$$

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

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 categories in verbs but Tamil does have it. It is also possible that a POS tag in one language does not occur in another language at all. In this case, we won't 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 are not able to find an exact match for a tag, abstract level tagsets can be aligned through the adaptation knowledge of EAGLES guidelines.

4 Approach

In order to arrive at an agreement between multiple language POS tagset, researchers have adopted various strategies as we discussed. Some derived a new tagset capturing the morphosyntactic features of some specific set of the languages (Bureau of Indian Standard) and some mapped existing POS tagsets to a universal POS tagset. However, both approaches introduce new POS tagset. Unlike these prior approaches we took a completely new angle. We casted the problem of heterogeneity in POS tagsets as an alignment of two labeled 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 since, 1. we have accessed to necessary data and expertise 2. these languages are low resourced 3. they gain more importance as official languages of Sri Lanka. Below the rationales behind choosing the representative tagset from each language are described. Then, semi-automatic POS alignment algorithm is presented.

4.1 Tagset Selection

As there are several tagsets available in each language, selection of a POS tagset is essential for this study. While choosing a tagset of a language, the usability and standardization are considered. Next 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 University of Colombo School of Computing (UCSC) tagset which was developed by University of Colombo [16] and UOM tagset by University of Moratuwa [17]. The details of the tagsets are described in the next subsection. UCSC tag set contains 29 tags which include foreign word and Symbol. There are three versions in UCSC tag set. The University of Moratuwa has built an improved version of UCSC tagset by overcoming the issues, 1. Some word classes in Sinhala are not covered by UCSC tagset. 2. Out of the 100,000 words in the manually POS tagged corpus, 3989 words do not fall into any category 3. Some words being tagged with multiple tags in different places with a similar meaning. 4. Not comprehensive enough to cover the inflection based grammatical variations [17].

There are three levels in this tagset by 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 30 tags [17]. UOM tagset is selected for this study because of the above mentioned major limitations of 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 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 choose BIS Tamil Tagset since it is the officially accepted standard tag set 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 below reasons. 1) It has no apparent impact in most of the applications it used. 2) The deeper levels are at times inflectional forms than being truly POS classes 3) Tagging time increases as we need to split the word into morphemes 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. Then the parallel corpus of Languages P1 & P2 was annotated using respective automatic POS taggers. Then the tagged parallel corpus was word aligned using a word alignment tool. After that, best three mappings for each POS tag were selected based on the amount of word alignment and presented to human evaluators. The experts pruned the provided mappings and arrived at a final quality and complete alignment. Below we present the each and every workflow steps and tools used for this approach in a descriptive manner.

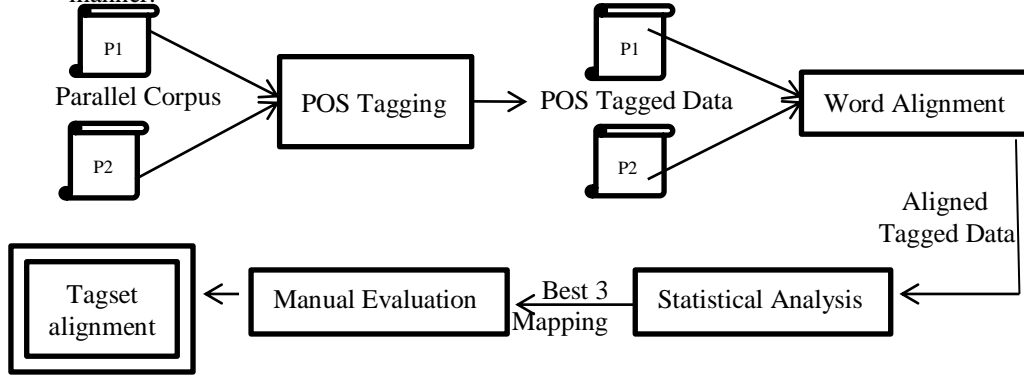


Fig. 1. Work flow of the semi-automatic POS tagsets alignment of P1 and P2 languages

We have access to the Sinhala-Tamil parallel corpus of government official documents. This parallel corpus is manually cleaned & aligned by three professional translators. This corpus contains more than 40,000 words. This parallel corpus was annotated using the automatic POS tagger of both languages. For the Tamil language, we have used an automatic POS tagger developed by Dhanalakshmi et al of AMRITA University, Coimbatore. The system was trained with a corpus of twenty-five thousand sentences and they claimed accuracy of 95.63% [19]. We have used an automatic POS tagger based on SVM which was developed by the University of Moratuwa, Sri Lanka to annotate the Sinhala corpus. Researchers reported an overall accuracy of 84.68% [17].

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

In order to proceed with tagset alignment, initially, a number of words belong to each tag was calculated in either language which resulted in most of the words into “common noun” category. Based on the word alignment, a tag alignment was retrieved. This resulted any tag of one language can be mapped to any tag of the other. In our study, there are 35 tags from BIS tagset and 30 tags from UOM tagset. So 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 mapping were derived using an automatic program by counting words belongs to each mapping. The general idea is to take into consideration all the tag alignments of both languages that were generated from the GIZA++ algorithm and chose the most frequent of them as the correct alignment. But, in our approach, we chose top three frequent aligned tags and cross-checked it with bilingual experts to finalize the alignments. For example “Nipathana” in UOM tag aligned with “Verb Finite” and “Common noun” mostly in BIS tagset. But through the linguistic point of view, it does have to align with “Verb finite”.

5 Results and Discussion

Through the experiment, there are some possible relationships holding between BIS tagset and UOM tagset. In this section, the details of four types of relationship and the examples are focused. The results of POS tagset alignment of Tamil and Sinhala languages after manually proven are tabulated in Table 1. Results are based on word alignments and two linguists’ opinion. There are 8 equal relationships, 22 sub-sumption relationships, 1 complex relationship and no non mapped relationships.

TABLE I. 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	Com-mon/proper	க்கு, உடைய	ට, ගේ	to, 's
Proper noun	Proper noun	ஜான்	ජොන්	John
Pro-noun/Deterministic Pronoun	Personal Pro-noun	நான், நீ	මම, ඔබ	I, you
Pronoun	Reflexive Pro-noun	தான்	-	Myself
	Reciprocal Pronoun	ஒருவருக்கொருவர், அவனவன்	එක එකකෙනාට, ඔවුනොවුන්	each other
Questioning Pro-nouns	Question words	என்ன, எப்படி	කුමක්ද, කෙසේද	what, how

Question-Based Pronouns	Relative Pronoun	எங்கே, எது	கொனே, கவர	where, which
Determiners	Deictic	இவன், இவள்	மே, கிடெ	this, all
	Relative	அவ்வீடு, இவ்வீடு	ஃ றெடர், மே றெடர்	That home, this home
Verbal Participle	Verbal participle	பார்த்து	பெ	Looked
Verb finite	Verb finite	செய்தான்	கலேய	Did (he)
Preposition in compound verb		-	ஓபு, கிடெ	-
Nouns in Compound Verb		படிக்கின்றான்	பாமி கரனலா	Study
Adjective in Compound Verbs		கூட்டப்படுகின்றது	புமி கரனலா	Increasing
Nipathana		போதும், காணாது	ஈதி, துதி	Enough/ not having
Modal auxiliary	Verb auxiliary	முடியும், வேண்டும்	காதி, டுது	Can, should
Verb Non-Finite	Infinite Verb	விழ	புபிமெ பகே	like to fall
	Conditional Verb	நடந்தால்	ஈபிடேஊன்	If walk
Verbal Noun	Verbal Gerund	படித்தல்	ஓகெதிம	Studying
	Verbal noun	படிப்பு	-	Study
Adverb	Adverb	விரைவாக	பெகெயன்	Fast
Adjective	Adjective	மிருதுவாக	கூபுடி	Smooth
	Relative Participle	நடந்த	ஈபிட	Walked (kid)
Conjunction	Coordinator	உம், மற்றும்	கை, கை	Or, and
	Subordinator	என்று, என	யது, யுதி	That
Particle	Default Particles	மட்டும், கூட	ய, டு, ம	Only, also
	Classifier	அட்டும்	-	-
	Intensifier	அதி, வேக, மிக	ஓதா	Most, speed
	Negation	இல்லை	தா, துத	No
Interjection	Interjection	ஐயோ	ஈகெயன்	Oh
Postposition	Postposition	பற்றிகுறித்து	காது	Related
Number	Cardinal	ஒன்று, 1	பக, 1	One, 1
	Ordinal	முதல், இரண்டாம்	பகபிபன, டேபன	First, second
Punctuation/Full	Punctuation	/?,:''	/?,:''	/?,:''

stop	Symbol	\$. &*,(\$. &*,(\$. &*,(
Foreign word	Foreign Residuals	கார்	කාරය	Car
Abbreviation	Unknown	மு.ப	මෙ.ව	a.m

5.1 Equal relationship

There are some POS alignments which hold an equal relationship. Equal relationship implies one language tagset can equally align with another language tagset. As mentioned in Table 1, some POS alignments fall under the equal relationship. The adverb in the Tamil language can be directly mapped to 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 the equal relationship as it has same features. Questioning pronouns words are used to ask a question. So that is equivalently aligned with question words in BIS tag set.

5.2 Subsumption relationship

In most of the cases, a POS tag in the Sinhala language is not mapped directly to Tamil language POS tag. Most of those tags fall under subsumption relationship. Nipathana is a category in the Sinhala language, but which does not have direct mapping tag in the Tamil language. So Nipathana does have to map with the finite verb category in the Tamil language (subsumption \subseteq, \supseteq). 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 \subseteq Relationship). This mostly occurs when a number of aspects used in the specialization of a POS tag differ between languages. BIS tagset does have five categories of pronouns while there are only four categories in UOM tag set. As a result, we are not able to equal align those tags. The Personal, Reflexive and Reciprocal pronouns from BIS tagset are subsumptionally aligned with Pronoun tag in UOM tag set. Deterministic pronouns in UOM tagset are aligned to personal pronouns in BIS tag set. 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 this tag aligns with the Relative pronoun in BIS tag set. But Relative pronoun 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 BIS tag set while UOM tag sets have only one category. The subcategories Deictic and Relative are aligned to Determiners tag. Particles are further divided into five sub-categories in BIS tag set while there is only a

parent node Particles in UOM tag set. Hence, the subcategories are mapped to Particles in UOM tagset using subsumption relationship. General, ordinal and cardinal are the three categories of Quantifiers in BIS tag set. Yet, UOM tag set only have Number category. Thus, three subcategories are aligned with Number category. Full stop in UOM tagset does have subsumption relationship with punctuation in BIS tag set. Like that, Symbol in BIS tag is aligned with punctuation category of UOM tag set. 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 BIS tag set are aligned to the Common noun in UOM tag set.

A noun in Compound Verb is another category of the noun in the Sinhala language. It is a combination of noun and verb. The noun which makes compound verb is called as nouns in the compound verb. There is no matching translation in English and Tamil since all compound verbs in the Sinhala language is a normal verb in English and Tamil. In this example, First part of the verb is identified as ‘Noun in the compound verb’. So this ‘Noun in Compound verb’ tag is subsumptionly mapped with 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 which describes the main common noun ‘වත්ත (garden)’. But according to the Tamil grammar rule, if a noun expresses another noun it cannot be categorized under adjective category. So those ‘Adjectival noun’ is mapped with common noun in BIS tagset.

Further, adjectives are categorized into three subcategories Adjective, Adjectival Noun, and Adjective in Compound Verbs. As we saw above, Adjectival Noun tag is aligned to 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 Adjective in compound verbs. In the example ‘වැඩි කරනවා (increase)’, වැඩි is an adjective and කරනවා is a verb. But Tamil we can write this as ‘கூட்டப்படுகிறது’. Hence, there is no matching translation in Tamil for the adjective in the compound verb, since all compound verbs in Sinhala is a normal verb in Tamil. Thus ‘Adjective in the Compound verb’ is mapped with Finite verb tag of the BIS tagset. Remaining subcategory ‘Adjective’ is aligned to Adjective in BIS tag set.

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 map with adjective in UOM tag set. Similarly, gerunds and verbal nouns BIS tagset is aligned to Verbal noun in UOM tagset. At the same time, however, they retain their verbal arguments. Usually, these words are tagged as forms of verbs. Likewise, infinite verb and conditional verb in BIS tag set are aligned to non-finite verb category in UOM tag set.

Some other categories in UOM tagset also fall under the Verb category of 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. But Tamil we can write this as ‘செய்கிறார்’. Hence, there is no matching translation in Tamil for the preposition in the compound verb, since all compound verbs in Sinhala is a normal verb in Tamil. Thus ‘Preposition in the Compound verb’ is mapped with Finite verb tag of the BIS tagset.

Nipathana is a tag in UOM tagset which is used alone in some contexts and as a postposition. But Tamil language does not have an exact match for this category. This category is mapped with 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 POS tagset are unique to the particular language. Those features may map to another category or categories when we come to alignment. There are some complex alignments when we try map POS tagsets of Sinhala and Tamil language. Hence, we went deep in the grammar of both languages to find out the relationship for those categories.

Sinhala and Tamil nouns are morphologically inflected based on the case. To indicate case, a suffix is attached. According to Sinhala language rules, it is incorrect to detach these case marking suffixes from the main noun. However, some Sinhala writers tend to separate this case marking suffix from the main noun. So unlike the Tamil language, the Sinhala language has space in between the noun and its case marker. Subsequently, there is a new POS tag added “Case marker” in Sinhala, but not in Tamil. Case marker does not have an English meaning on its own. This tag set has to align with a common noun or proper noun according to the previous tag set alignment in the Sinhala language. So this alignment falls into the composite relationship. For an 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 case marker. But in the Tamil language, it will be “மரத்துக்கு” and tagged under the common noun category. This correspondence is fallen into composite relationship.

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

6 Conclusion and Future Works

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

this algorithm. We have presented a quality alignment between Sinhala UOM tagset and Tamil BIS tagset. Even though, these two languages have been in contact for a long period of time, the grammars are not identical between these languages 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. To come up with alignment, the highest three mapping derived using an automatic program by counting words belongs to each mapping. But, in our approach, even though we choose top three frequent mappings, all alignments fall within top two mappings. The solutions we proposed follow the ultimate goal that information loss is minimized and no need of creating a new tagset. This approach is language independent and we could apply for the different tagsets which belong to a language. POS alignment used to study the similarity and dissimilarity of grammar quantitatively. Also it is the good starting point for the study of language divergence. In the future, we plan to extend this study for different tagsets which wither belong to different language or same language.

7 Acknowledgement

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.: Universal Dependency Annotation for Multilingual Parsing. In Proceedings of ACL, Sofia, Bulgaria, (2013).
2. Hardie, A.: The Computational Analysis of Morpho-syntactic Categories in Urdu. PhD thesis submitted to Lancaster University. (2004).
3. Bureau Indian Standard. (n.d.). Unified Parts of Speech (POS) Standard in Indian Languages.
4. Sankaran Baskaran, Kalika Bali, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, Rajendran S, Saravanan K, Sobha L.: Designing a Common POS-Tagset Framework for Indian Languages. The 6th Workshop on Asian Language Resources, (2008).
5. Leech, G and Wilson, A.: Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Re-port EAG-TCWG-MAC/R, (1996).
6. Selvam, M., & Natarajan, A. M.: Improvement of rule based morphological analysis and pos tagging in tamil language via projection and induction techniques. International journal of computers, 3(4), 357-367, (2009).
7. Volz, Norbert, and Suzanne Lenz.: Multilingual Corpus Tagset Specifications. *MLAP PAROLE 63æ386 WP 4.4* (1996).
8. Nancy Ide and Jean Véronis: Multext (multilingual tools and corpora). In Proceedings of the 15th International Conference on Computational Linguistics (COLING-94), Kyoto, Japan, (1994).

9. Tomaž Erjavec.: Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisboa, Portugal, (2004).
 10. Nitish Chandra, Sudhakar Kumawat, Vinayak Srivastava.: Various Tagsets for Indian Languages and Their Performance In Part Of Speech Tagging. Proceedings of 5th IRF International Conference, ISBN: 978-93-82702-67-2, (2014).
 11. Zeman D.: Reusable Tagset Conversion Using Tagset Drivers. In Proceedings of LREC, Marrakech, Morocco, (2008).
 12. Francis M. Tyers, Joakim Nivre: Universal Dependencies for Turkish. Proceedings of COLING, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3444–3454, Osaka, Japan, (2016).
 13. Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D.: Universal Stanford Dependencies: a cross-linguistic typology. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), Reykjavík, Iceland, May. European Language Resources Association (ELRA), (2014).
 14. Reut Tsarfaty: A unified morpho-syntactic scheme of Stanford Dependencies. In Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL), pages 578–584, (2013).
 15. AnHai Doan, Alon Y., Halevy.: Semantic-Integration Research in the Database Community A Brief Survey. AI Magazine Volume 26 Number 1, (2005).
 16. Dilmi Gunasekara, W.V. Welgama, A.R. Weerasinghe.: Hybrid Part of Speech Tagger for Sinhala Language. International Conference on Advances in ICT for Emerging Regions (ICTer), 041 – 048. (2016).
 17. Fernando, S., Ranathunga, S., Jayasena, S., & Dias, G.: Comprehensive Part-Of-Speech Tag Set and SVM Based POS Tagger for Sinhala. WSSANLP 2016, 163, (2016).
 18. Central Institute of Indian Languages. (n.d.).
 19. Dhanalakshmi, V., Padmavathy, P., Soman, K. P., & Rajendran S.: Chunker for Tamil. In Advances in Recent Technologies in Communication and Computing. ARTCom'09. International Conference on (pp. 436-438). IEEE, 2009, October, (2009).
 20. Dandapat, S.: MSRI Part-of-Speech Annotation Interface.
 21. Lakshmana Pandian, S, & Geetha, T.: Morpheme based Language Model for Tamil Part-of-Speech Tagging. Polibits, (38), 19-25, (2008).
 22. Ramanathan, M., Chidambaram, V., & Patro, A.: An Attempt at Multilingual POS Tagging for Tamil.
 23. IIIT Homepage, http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines, last accessed 2018/01/03. IIIT-tagset. “A Parts-of-Speech tagset for Indian languages”.
 24. Ramasamy, Loganathan and Žabokrtský, Zdeněk.: Tamil Dependency Treebank v0.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, (2014).
 25. F. Och.: Minimum Error Rate Training in Statistical Machine Translation. in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, (2003).
 26. Daniel Zeman: Hard Problems of Tagset Conversion.
 27. Leech, G.: Grammatical Tagging. In Corpus Annotation: Linguistic Information from Computer Text Corpora. ed: Garside, Leech and McEnery, London: Longman, (1997).
 28. Slav Petrov, Dipanjan Das, Ryan McDonald: A Universal Part-of-Speech Tagset. arXiv:1104.2086v1 [cs.CL], (2011).
- Zeman D.: Parsing with a Statistical Dependency Model. PhD thesis, Univerzita Karlova v Praze, (2004).