Improvement of Rule Based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques

M. Selvam, A.M. Natarajan

Abstract— Morphological analysis and part of speech (POS) tagging are very essential for natural language processes like generation of Treebanks, training of parsing models and parsing. Rule based approach is applicable to the languages which have well defined set of rules to accommodate most of the words with inflectional and derivational morphology. Rule based morphological analysis and POS tagging are very difficult and cannot accommodate all combinations through the rules due to inflections and exceptions especially in languages like Tamil. Statistical methods are very important which in turn need large volume of electronic corpus and automated tools which are very rare in Tamil. Since English is very rich in all aspects, POS tags can be projected to Tamil through alignment and projection techniques. Rule based morphological analyzer and POS tagger can be built from well defined morphological rules of Tamil. They can be further improved by the root words induced from English to Tamil through the sequence of processes like alignment, lemmatization and induction with the help of sentence aligned corpora like Bible corpora, TV news, newspaper articles since finding the root in the inflected words is very difficult and leads to ambiguity. In our experiments, rule based morphological analyzer and POS tagger were built with 85.56% accuracy. POS tagged sentences in Tamil were obtained for the Bible corpus through alignment and projection techniques and categorical information had been obtained. Root words were induced from English to Tamil through alignment, lemmatization and induction processes. Further 7% improvement was made in rule based morphological analyzer and POS tagger using categorical information and root words obtained from POS projection and morphological induction respectively via sentence aligned corpora.

Keywords— Alignment, Induction, Lemmatization, Morphological Analysis, Parsing models, POS Tagging, Projection, Tamil Language, Treebank.

I. Introduction

Natural Language Processing (NLP) is a set of computational techniques for analyzing and representing text in Natural Language (NL) with linguistic analysis for achieving human-like language processing for a range of tasks or applications. It deals with interactions between computer and human (natural) languages. Computational linguistics (CL) is an interdisciplinary field dealing with statistical and/or rule based modeling of NL from a computational perspective. Major components driven by these broad areas are syntax, semantics and pragmatics. NLP applications like Grammar Checker and Language Analyzer need a parser with an optional parsing model. Parsing is the process of analyzing the text automatically by assigning syntactic structure according to the grammar of NL. A parser is a computational system which processes input sentences and builds one or more constituent structures called parses or parse trees. For a simple parsing task considering languages with a limited vocabulary, parsers using rule based techniques are usually sufficient. For applications requiring a large vocabulary parsers based on more sophisticated parsing models are needed, for example models which use probability distributions over parses to accomplish the disambiguation task in ambiguous sentences. Adequate parsing models can be created by adding structural components in statistical methods which satisfy the constraints needed for the parsing process [1], [2].

In order to build a parsing model, large volume of Treebank is needed. Treebank is a corpus with linguistic annotation beyond word level. Part of Speech (POS) tagging and phrasing are essential for the development of Treebank. POS tags are generated through morphological analysis. POS tagging is used to assign or select correct POS tag to a word before syntactic analysis. Phrasing is the process of applying morpho-syntactic relations among words in the formation of constituents which in turn build parse trees. Collection of phrase structured sentences together constitutes a Treebank. Morphological analysis is the process of segmenting a given word into a sequence of morphemes. It is closely related to POS tagging but word segmentation is required for natural languages because morpheme boundaries are not indicated. Inflectional morphology gives different forms added to a root word whereas derivational morphology derives new words by inclusion of affixes. Lexical and surface levels of words are studied through morphological analysis. Based on that, POS tags are suggested to words in a sentence.

Manuscript received September 2, 2009:

M.Selvam is an assistant professor in department of Information Technology, Kongu Engineering College, Perundurai, Erode, Tamilnadu, India – 638052 (phone : +91 9486655106; fax : +91-4294-220087; e-mail : amm_selvam@yahoo.co.in, am_selvam@kongu.ac.in)

A.M. Natarajan is chief executive and professor at Bannari Amman Institute of Technology, Sathyamangalam – 638 401, Erode, INDIA (e-mail: amnatarajan2006@yahoo.co.in).

Rule based morphological analysis is applicable to the languages which have well defined set of rules. For a new language, morphological analysis and POS tagging are very difficult, time consuming and laborious. Rule based morphological analysis and POS tagging are very difficult to achieve and all combinations of affix patterns cannot be accommodated through rules due to inflections and exceptions especially in a language like Tamil. Rule based morphological analysis depends upon pattern matching with affixes and processing with partial stems. Correct POS tag cannot be assigned to a Tamil word when there is an ambiguity. Due to that, categorical information cannot be determined. For these problems, rule based approaches will not help. Hence, moderate accuracy in annotation can be obtained in rule based approach. Rule based morphological analyzer and POS tagger have been developed with 85.56% accuracy. Shortfall in accuracy is due to partial stemming and misidentification of exact root words. It is difficult to further improve rule based approach. An alternate method is needed to augment it. From the literature, it is found that categorical information and root word are needed for assisting rule based approach for the improvement.

There is no standard method for the identification of root words in Tamil. Stemmers and lemmatizers are not available in Tamil. Projection and induction techniques can be used for POS tagging, base noun-phrase bracketing, named entity tagging and morphological analysis from a resource rich language to a resource deficient language [3]. This has motivated to apply alignment and projection techniques for projecting POS tags, and alignment, lemmatization and morphological induction techniques for inducing root words from English to Tamil. Categorical information and root words can be obtained from POS projection and morphological induction respectively from English via alignment across sentence aligned corpora [4]. English is very rich in terms of resources, tools and techniques especially in statistical side. Since most of the Tamil works in POS tagging and morphological analysis are based only on structural methods, it is motivating to leverage advantages of resources, tools and techniques available for English for the improvement of rule based morphological analysis and POS tagging in Tamil.

In this paper, experiments have been done for the generation of POS tagged sentences and improvement of rule based morphological analysis and POS tagging. Generation of POS tagged sentences has been done using alignment and POS projection from English to Tamil. Improvement of rule based morphological analysis and POS tagging has been done using POS projection and induction from English to Tamil.

A. *NL Processes and Tools in Tamil Language*

Tamil is one of the classical Indian languages which has very strong linguistic base with well defined set of morpho-syntactic rules. Parsing, development of parsing models, chunking, generation of Treebank, POS tagging, morphological analysis, and development of semi-automated and automated tools for these processes in Tamil are at the nascent stage. Morphological analyzers and generators were built based on various techniques and constraints like morpho-tactics, morphological alternations, transliteration, phonology and morpho-phonemics. Some of these works were reported by the authors named Rajendran, Ganesan, Kapilan, Deivasundaram, Vishnavi, Ramasamy, Winston Cruz and Dhurai Pandi, and organizations named AU-KBC (Anna University-KBC) at Madras Institute of Technology (MIT) at Chennai and Resource Centre for Indian Language Technological Solutions-Tamil (RCILTS-T) at Anna University, Chennai. POS taggers were built by Vasu Ranganathan and Ganesan and RCILTS-T. Chunking tools were developed by AU-KBC and RCILTS-T. Syntactic parser was developed by Baskaran, Kumara Shanmugam and RCILTS-T using techniques like finite state automata, grammatical structure formulation and simple rule based method respectively [5]. A simple morphological tagger which identifies suffixes, labels them and separates root words from transliterated Tamil words was reported [6]. Tools for corpus analysis and overview of tagging methodologies were reported in the literature [7]. Syntactic tagging and rule based text analysis tools were also reported in the literature [8] [9].

Due to the constraints, limited coverage of morpho-syntactic and semantic rules, non-availability of methodologies towards large scale development of parsing models, non-availability of standards, non applicability of statistical methods and resource deficiency, reported tools cannot be used directly for all types of NLP applications. Hence development of efficient morphological analyser, POS tagger and phrasing tool is essential. Tools with rule based approaches can be developed with the well defined set of morpho-syntactic rules. This has encouraged the development of rule based tools like morphological analyser, POS tagger and phrasing tool. Hence rule based morphological analyzer, POS tagger and phrasing tool have been developed.

Rule based techniques cannot address all inflectional and derivational word forms and peculiar characteristics like relative free word order, syntax with semantics and long distance relationship to a greater extent. Moderate accuracy can only be achieved in rule based techniques. Hence improvement of rule based morphological analysis and POS tagging through statistical methods like alignment, projection and induction is essential.

II. Issues in Tamil Language

Tamil grammar is agglutinative in nature. Suffixes are used to mark class, number and cases attached to a noun. Tamil word may have a lexical root to which one or more affixes are attached. Most of the Tamil affixes are suffixes which can be derivational or inflectional. Length and extent of agglutination is longer in Tamil resulting in longer words with many suffixes. Some of the other issues are morpho-phonology (sandhi – insertion, deletion and substitution of morphemes like ந், க், ம், ச், த் at word boundaries) rules, complex noun and verb patterns, and out of vocabulary rate due to inflections. Poetry forms are more complex than prose forms.

In Tamil, nouns are classified into rational and irrational forms. Humans come under rational form whereas all other nouns are classified as irrational. Rational nouns and pronouns belong to one of the three classes: masculine singular, feminine singular and rational plural. Irrational nouns belong to one of the two classes: irrational singular and irrational plural. Suffixes are used to perform functions of cases or post positions. Tamil verbs are also inflected through the use of suffixes. The suffix of the verb indicates person, number, mood, tense and voice.

Tamil is consistently head-final language. The verb comes at the end of the clause with a typical word order of Subject Object Verb (SOV). However, Tamil allows word order to be changed making it a relatively word order free language. Other features are plural for honorific noun, frequent echo words, and null subject feature i.e. all sentences do not have subject, verb and object.

A. Need of Rule based Methods

Electronic resources and tools are scarcely available in Tamil. To cater to these challenging needs of Tamil, hybrid parsing models constrained on structural components [10] developed from phrase or dependency structured Treebanks are needed [11]. Generation of Treebank for training parsing models needs laborious man power, time, effort, automated tools, bootstrapping, accuracy and consistency. Phrase and dependency structured Treebanks are to be developed with the process of POS tagging with robust morphological analysis. Indian languages have well defined set of rules in lexical and syntactic structures. When the morphological rules are rich in a language, electronic resources like corpus are not needed in high volume for training process. Even if there is a resource deficiency, rule based morphological analysis and POS tagging can be done with reasonable accuracy [12]. Since Tamil has well defined rules for formation of words, rule based morphological analysis and POS tagging help to a large extent [13] [14]. Generation of new tags, stemming process, creating lookup tables, dynamic memory learning and disambiguation are some of the problems to be handled in an effective manner.

Even though Tamil has well defined morpho-syntactic rules, there are many issues which will not be suitable for applying transformation based learning (TBL) or Hidden Markov Model (HMM) based approach in POS tagging. Since extent of agglutination is larger in Tamil, there are many suffix patterns and their combinations in a word. A Tamil word can take one prefix and seven suffixes at the maximum. There are many exceptions in applying morpho-syntactic rules. Tamil has the nature of relatively free word order in its sentential structure. Ambiguity is also a problem in differentiating noun and verb, verbal noun and verb, and similar suffix patterns for different parts of speech. In order to make the tagger to learn the morpho-syntactic rules in TBL and HMM based techniques with the fore mentioned problems, large size annotated corpus is needed. Hence rule based approach is preferred for POS tagging and morphological analysis in Tamil.

B. Need for Statistical Methods

In rule based morphology, it is very difficult to accommodate all forms of a word. Application of statistical methods like projection and induction techniques will help in generation of resources in Tamil and improving rule based techniques with hybridization. Due to inflections, affix patterns to be checked in morphological analysis and number of POS tags to be generated for POS tagging are more. In English, less than 50 POS tags are used. POS projection from English to Tamil will help in generation of POS tagged sentences and providing the categorical information to morphological analysis. Stemmers and lemmatizers are available with greater accuracy in English [15]. Morphological induction from English to Tamil will help in finding root words for morphological analysis. Further morphological processes can be automated through rule based techniques from the root word or stem created by the induction process.

III. PROPOSED MORPHOLOGICAL ANALYSIS AND POS TAGGING IN TAMIL

Tamil also has eight regular parts of speech like English and some other parts are additionally used. Some of the POS categories in Tamil are listed in table 1.

Table 1. POS categories in Tamil and examples

POS Category	**Examples**
Noun	மரம் (tree)
Pronoun	அவன் (he)
Verb	செய்தேன் (did)
Adverb	உயிருடன் (living)
Adjective	அழகிய (beautiful)
Preposition	வெளியே (out)
Conjunction	ஏனெனில் (because)
Interjection	ஐயோ (alas)
Wh-words	என்ன (what)
Determiner	அந்த (the)
Quantifier	சில (some)
Adjective Participle	உயருகின்ற (growing)
Echo words	கடகட (echo sound)
Complementizer	ஆக (that)
Ordinal Number	மூன்றாவது (third)
Conditional Participle	செய்தால் (if done)

Optative	வைக்க (to keep)
Others	செ.மீ (cm.)

Parts of speech in Tamil take different forms and inflections. Morphological inflections on nouns include gender and number. Prepositions take either independent or noun combined forms with various cases like accusative, dative, instrumental, sociative, locative, ablative, benefactive, genitive, vocative, clitics and selective [16]. Examples of all cases are listed in table 2.

Table 2. Case suffixes used with Noun

Cases	Suffixes
Accusative	ஏ, ஐ
Dative	க்கு, ற்கு
Instrumental	ஆல்
Sociative	ஓடு, உடன்
Locative	இல், உள், இடம்
Ablative	இருந்து
Benefactive	க்காக, ற்காக
Genetive	இன், அது, உடைய
Vocative	ஆலே
Clitics	உம், ஓ, தான்
Selective	ஆவது
Interrogative	ஆ

Verbs in Tamil take different forms like simple, transitive, intransitive, causative, infinitive, imperative and reportive. Verbs are formed from the stem with various suffix patterns. Verbal suffix patterns are shown in table 3.

Table 3. Verbal suffixes in Tamil

Suffix	Categories	Sub categories	Suffixes
Tense	Present	---	கிறு, கின்று, ஆனின்று
	Past	---	த் , ந் , ற் , இன்
	Future	---	ப் , வ்
Person	First	Singular	ஏன்
		Plural	ஓம்
	Second	Singular	ஆய்
		Plural	ஈர்கள்
		Honorific	ஈர்
	Third	Male Singular	ஆன் ,அன்
		Female Singular	ஆள்
		Common Plural	ஆர்கள்
		Honorific	ஆர் , அர்
		Neutral Singular	அது
		Neutral Plural	அன
Others	Causative	---	இ
	Verbal Noun Untensed	---	அல்
	Infinitive	---	உ
	Imperative	Plural	உங்கள்
		Negative	ஆதே, ஆது
	Passive	---	படு
	Future	Negative	மாட், இல்லை
	Optative	---	முடியும், வேண்டும், கூடும், ஆம்
		negative	முடியாது, கூடாது, வேண்டாம்
	Sandhi	---	ந், க், ம், ச், த்
	Plural	---	கள்

Adjective has tense or negative participles. Other parts of speech take simpler forms. Some of the proposed Tag formats which are based on POS forms and their inflections are shown in table 4.

Table 4. POS Forms, Morphological Inflections and Proposed Tag Format

POS & Others	Forms	Morphological inflections	Tag Format
Noun	Simple Noun (NN) Proper Noun (NR) Participle Noun (NP) Adjective Noun (NA) Positive Tensed Verbal Noun	Number ▪ Singular (S) ▪ Plural (P) Gender ▪ Male (M) ▪ Female (F) ▪ Neutral (N) ▪ Common (C) ▪ Oblique (O)	(NN | NO | NR | NP | NA | NVT | NNVT | NVUT) | (S | P) | (M | F | N | C | O)

	(NVT) Negative Tensed Verbal Noun (NNVT) Un-tensed Verbal Noun (NVUT)		
Verb	Simple Verb (V) Transitive Verb (VT) Intransitive Verb (VI) Causative Verb (VC) Infinitive Verb (VIF) Imperative Verb (VIM) Reportive Verb (VRP)	Person ▪ First (F) ▪ Second (S) ▪ Third (T) Number ▪ Singular (S) ▪ Plural (P) Gender ▪ Male (M) ▪ Female (F) ▪ Neutral (N) ▪ Common (C) Tense ▪ Present (P) ▪ Past (A) ▪ Future (F) Passive (P) Honorific (H) Negative (N) Interrogative (Q) Suffix (X)	V (F I S I T) (S I P) (M I F I N I C) (P I A I F) I (P) I (H) I (N) I (Q) VIF I (N) VIM (S I P) I (X)
Adverb	Simple Adverb		ADV
Adjective	Simple Adjective Participle Adjective	Tense ▪ Present (P) ▪ Past (A) ▪ Future (F) Negative (N)	ADJ I (P I A I F I N) (P)
Preposition	Simple preposition Noun + cases	Cases ▪ Accusative (A) ▪ Dative (D) ▪ Instrumental(I) ▪ Sociative (S) ▪ Locative (L) ▪ Ablative (AB) ▪ Benefactive (B) ▪ Genetive (G) ▪ Vocative (V) ▪ Clitics (C) ▪ Selective (SL) Negative (N)	PRP (NOUN) (A I D I I I S I L I AB I B I G I V I C I SL I O I V) (N)
Conjunction	Simple Conjunction Coordinating conjunction	Wh words ▪ What (A) ▪ Who (O) ▪ Whose (S)	CON WH (A I O I S I E I R I M I I I W)
Interjection	Simple Interjection		INT
Others	Echo words (ECH) Determiner Quantifier Complementizer Ordinal Optative	Same (S) Different (D)	ECH (S I D) DET QNT CMP ORD OPT

From the above analysis, more than 600 POS tags were generated for our rule based morphological analysis and POS tagging. Some of the tags are shown in table 5.

Table 5. Sample proposed POS Tags in Tamil Language

Tag	Description	Examples and meaning
ADJ	Adjective	அழகிய (beautiful)
ADJAP	Adjective Past participle	செய்த (done)
ADV	Adverb	வேகமாக (quickly)
CON	Conjunction	அல்லது (or)
CVCN	Verbal Conditional negative	செய்யாவிட்டால்(if not done)
DET	Determiner	இந்த (this)
INT	Interjection	ஐயோ (Alas)
NNSN	Noun singular Neutral	நேர்முகத்தேர்வு (interview)
NOSM	Pronoun singular masculine	அவன் (he)
NAPC	Adjective Noun plural common	நல்லவர்கள் (good people)
ORD	Ordinal	மூன்றாவது (Third)
PRP	Preposition	உள்ளே (inside)
QNT	Quantifier	சில (few)
V	Verb	படி (study)
VC	Verb Causative	கற்பி (teach)

VFPA	First Person Plural Past Tense Verb	சென்றோம் (we went)
VI	Intransitive verb	திரும்பு (turn)
VIF	Infinitive Verb	செய்ய (to do)
VSPAN	Second Person Plural Past Tense Negative Verb	செய்யவில்லை(did not do)
VT	Transitive Verb	திருப்பு (turn – any object)
VTSNFN	Third Person Singular Neutral Future Tense Negative Verb	எடுக்காது (will not take)

Along with POS tags, phrasing tags were suggested for the generation of phrase structure Treebank in Tamil. Every sentence in the corpus is segmented into sequence of tokens and each and every token is applied with POS tag for the application of direct meaning to words. Tags and lexicons are bracketed for all pairs through which phrase structure is imposed. Phrasing is done among the words to form syntactic phrases and constituents. Some of the proposed phrasing tags for Tamil are classified based on the POS are shown in table 6.

Table 6. Proposed Phrases

Phrases	**Descriptions**
NP	Noun Phrase
VP	Verb Phrase
ADVP	Adverbial Phrase
ADJP	Adjective Phrase
PP	Prepositional Phrase
CP	Conjunctional Phrase
IP	Interjectional Phrase
WHNP	Wh-noun Phrase
WHVP	Wh-verb Phrase
WHPP	Wh-prepositional Phrase
SBARQ	Direct Question by Wh-phrases
SBAR	Clause introduced by subordinating conjunction
SINV	Declarative clause with auxiliary inversion
SQ	Sub-constituent of SBARQ excluding Wh-phrase
S	Simple Declarative Clause

IV. IMPLEMENTATION OF RULE BASED MORPHOLOGICAL ANALYSIS AND POS TAGGING

Formation of words in Tamil is based on a well defined set of rules. Adverbs (வேகமாக – fast) and adjectives (அழகான – beautiful) have a closed set of suffix (ஆக, ஆன) patterns. Direct prepositions (உள்ளே – in), conjunctions (அல்லது – or), pronouns (அவன் – he), interjections (அய்யோ – alas) also have a closed set. For all closed set of words, morphological analysis is not needed and POS tags are applied directly with the lookup table. For adverbs and adjectives, morphological analysis is done with the separation of suffixes.

Adjective noun (நல்லவன் – good man), verbal noun (செய்தவன் – one who has done) and participle noun (செய்தது – one which has done) are some of the patterns which have pronouns (அவன், அது) in them. They are identified and separated by substring match. POS tagging is done with lookup after separating pronouns, adjectives or verbs or participles, and case endings (செய்தவனக்கு -> செய்த + அவன் + க்கு – to the one who has done).

Perfect and continuous tense verbal affixes (இரு, விட்டு, கொண்டு) are to be separated from words. This is done by substring match. POS tagging is done with separated word patterns (சென்றிருக்கிறேன் -> சென்று + இருக்கிறேன் - have gone, சென்றுவிட்டேன் -> சென்று + விட்டேன் – had gone, சென்றுகொண்டிருந்தேன் -> சென்று + கொண்டிருந்தேன் – was going). Checking pronoun with preposition is done by substring match with case endings. This is done by pronoun and case ending patterns (அவனுக்கு -> அவன் + க்கு – to him). POS tagging is done based on case endings which are considered as inflectional prepositions. Similarly for other noun based inflections, case endings are identified and morphological analysis and POS tagging are done with case endings (பூனையால் -> பூனை + ஆல் – by the cat).

Verbal inflections are checked with possible 80 patterns [17] and some more patterns which are suggested by us. Morphological analysis on verbs is done with verbal suffix patterns described in table 3. POS tagging is done with suffix patterns which cover number, person, gender, tense, voice, honorific, question, etc. There may be an ambiguity in nouns and verbs. Numbers are checked with lookup and partial pattern matching (ஆயிரத்துநூறு -> ஆயிரம் + நூறு – thousand and hundred).

Disambiguation module is needed for identifying nouns and verbs when multiple tags are assigned to a word. In the disambiguation module, nouns are distinguished from verbs by the POS of preceding word like determiner or adjective. Verbs are distinguished by the POS of the preceding word like adverb or verbal participles. If it is not resolved, most frequently applied tag will be used by matching with the dictionary which contains word, tag and frequency. Dynamic memory learning technique is used to add new nouns, verbs, adjectives and adverbs in various lookup tables dynamically when words are not matched with entries of their respective lookup tables. Our proposed rule based morphological analysis and POS tagging are shown in Fig. 1.

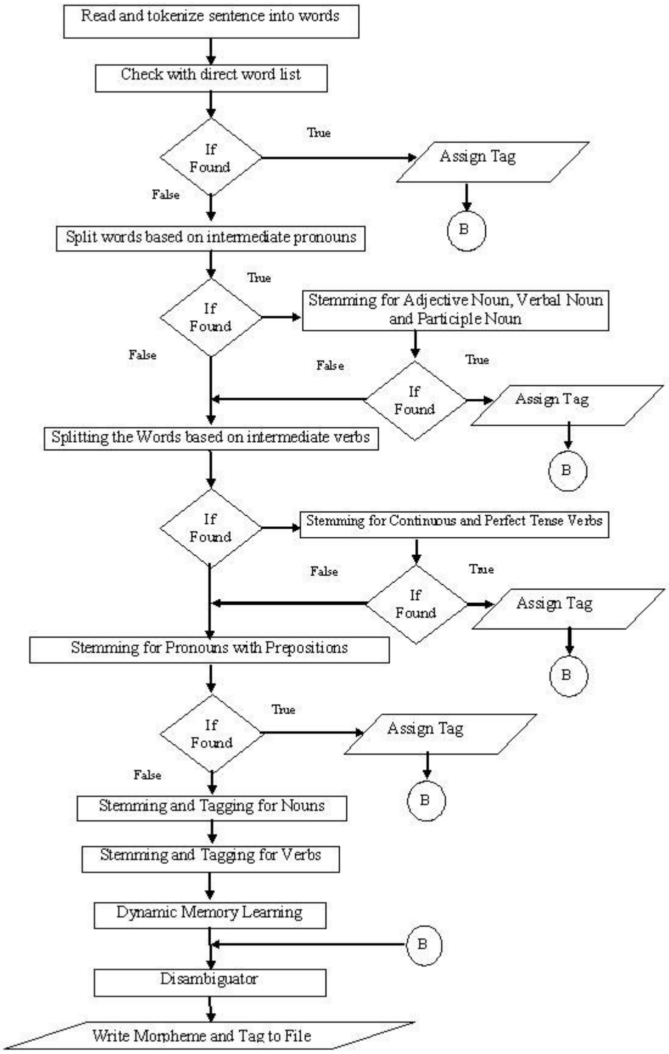
Read and tokenize sentence into words
Check with direct word list
If Found
True
Assign Tag
B
False
Split words based on intermediate pronouns
If Found
True
Stemming for Adjective Noun, Verbal Noun and Participle Noun
False
False
True
If Found
Assign Tag
B
Splitting the Words based on intermediate verbs
If Found
Stemming for Continuous and Perfect Tense Verbs
False
True
If Found
Assign Tag
B
Stemming for Pronouns with Prepositions
If Found
True
Assign Tag
B
False
Stemming and Tagging for Nouns
Stemming and Tagging for Verbs
Dynamic Memory Learning
B
Disambiguator
Write Morpheme and Tag to File

Fig. 1 Rule based morphological analysis and POS tagging

A. Issues in Rule based Methods

Rule based morphological analyzer and POS tagger have been successfully built and tested. Even though it works reasonably well, it has some issues to be resolved. It will not accommodate all rules directly because of some exceptions like more number of patterns in partial stem matching, noun-verb ambiguity, and context based identification of noun and verb. Morphological analysis is mainly done by a lookup with partial stem after substring matching and extraction. More number of partial stems for the same verb or noun, more memory requirement and order of partial stems based on their size in the lookup table are the problems with partial stemming. Sometimes it may not be possible to infer the categorical information from surface or lexical level of words.

B. Need for Projection Techniques

Generation of Treebank with rule based techniques with hand annotation and bootstrapping is time consuming and laborious. It also needs human expertise which is very difficult for high volume corpus and consistency cannot be maintained in the generation of initial training set [18]. Even though bootstrapping is applied for increasing training from the initial training set, corrections are to be made by human experts and only after that training can be done. This can be automated for a large corpus through alignment and projection techniques with sentenced aligned corpora [4]. Even though accuracy is moderate, a resource of large volume of POS tagged Tamil sentences can be generated by projecting POS tags from English to Tamil in the process of generating Treebanks. Morphological analysis and POS tagging in Tamil need a lot of refinement in them. To start with, moderately tagged Treebank is enough. On the other hand if the stem is exactly identified in rule based morphological analysis and POS tagging in Tamil, it will lead to a greater accuracy and human experts are not needed since Tamil has well defined set of rules in analysis and synthesis of words. When POS categorical information for each and every word is provided to rule based morphology and POS tagging, accuracy will be more. These issues can be resolved by projection and induction techniques. This will lead to the identification of noun or verb or noun and verb possibilities. Since sentence aligned corpora, and associated projection techniques and tools are available, POS projection and morphological induction are possible. Through these techniques, categorical information and lemmatized stems are produced which will be required to a large extent for assisting the rule based morphological analysis and tagging.

C. Need for Alignment

Indian languages closer to Tamil are also resource deficient in text corpora and tools. In order to overcome scarcity of resources, alignment and projection from resource rich languages like English through parallel corpora are essential. Availability of bilingual parallel corpora in English-Tamil is moderately possible in the domains like newspaper article, television news, web sites in both of the languages and magazines like India Today. Especially parallel corpora like Bible corpora are available as rich resources which have more than 30,000 sentences in both languages (English and Tamil). For statistical alignment it is reasonably sufficient. Hence projection and induction techniques can be used for POS tagging, base noun-phrase bracketing, named entity tagging and morphological analysis from English to Tamil. Projection techniques need alignment with bilingual parallel corpora. Through alignment, English words are aligned to Tamil words in a sentence. Using this alignment, language processing elements like POS tags, morphological features and noun phrases can be projected. Thus creating large resources like Treebank, parsing models and morphologically related resources in Tamil is highly possible.

D. *Alignment with Parallel Corpora*

With the parallel training corpora, English and Tamil sentences are aligned with the help of the alignment tool and dictionary. Alignment dictionary contains Tamil words and their equivalent English words in the domain. Nouns, verbal stems, adjectives, adverbs and interjections are included in the dictionary. Using the alignment dictionary and tool, English words are aligned to Tamil words. Two examples are given in Fig. 2

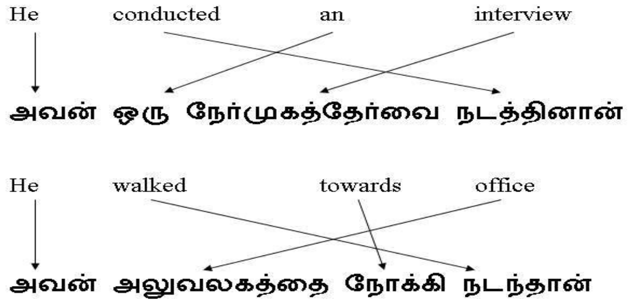
He
conducted
an
interview
அவன் ஒரு நேர்முகத்தேர்வை நடத்தினான்
He
walked
towards
office
அவன் அலுவலகத்தை நோக்கி நடந்தான்

Fig 2. Alignment Examples

Partial morphology applied to Tamil words will reduce lexical gaps between English and Tamil sentences which is explained in detail in section 4.6. Remaining 1-N and N-1 mappings are resolved by the statistical approach used in alignment tool.

E. *Need for POS Projection*

From the rule based morphological analyzer in Tamil it can be understood that nouns and verbs will be analyzed at the end due to complex patterns and ambiguity. Direct mapping can be done with the lookup table for closed set POS like interjection, conjunction, pronoun and direct prepositions and other parts like article, numbers and echo words. Adverbs and adjectives can be identified with well defined and limited set of suffix patterns. Prepositional inflections on nouns i.e. case endings can also be identified through well defined suffix patterns. After separating them, partial noun stems can be obtained. In the verbal inflections also, partial verbal stems can be obtained with various verbal suffix patterns. These suffix patterns decide tense, number, person, gender etc. In both of these cases, a dictionary with possible partial stem and its original stem is needed to have original stems. It is a tedious job. Apart from these patterns whatever words come in the content cannot be identified as noun or verb. Applying the context based disambiguation is also difficult. We need an approach for identifying a word as noun or verb. Once the categorical information is known like noun or verb, further morphological analysis becomes easier due to the well defined affix patterns for nouns and verbs. Hence projection of POS from aligned corpora becomes essential due to POS categorical information provided behind tags. Categorical information will be used for other parts of speech also. This will increase the accuracy of rule based morphology and POS tagging.

F. *POS Projection*

Sentences in English corpus are given to the parser and phrase structure parse trees (Treebank) are obtained. Phrases are eliminated in order to get words and their POS tags. After the alignment, every word in English sentences is mapped with one or more Tamil words. For the mapped Tamil words, POS tag of the aligned English word is projected. Accuracy depends upon the alignment process. This POS projection is not sufficient since an English word may be mapped to more than one word. NULL may be assigned to few words. For the same noun or verb in English various surface level combinations are possible due to suffixes in Tamil. In order to overcome this problem, partial morphological analysis in Tamil is needed for noun or verbal combination which suits direct mapping to English words. This reduces the big gap in number of words in parallel sentences which leads to the reduction of fertility [19]. This can be understood from the number of unique words in Tamil-English parallel corpus. Bible parallel Corpora contain 16,546 unique words in English and 95,116 unique words in Tamil. In the partial morphological process of Tamil, common patterns are achieved in verbs by omitting suffixes for number, gender etc except tense. Similarly for nouns in Tamil, case endings are separated as different words to match the English prepositions. After the partial morphological analysis in Tamil sentences, number of unique words has been reduced to 33,176. Through this process, alignment can be improved for functional and content words. Tamil words which are not mapped to English word are normally available in a separate list which is assigned with NULL in every sentence. Direct mapping can be done for functional words which are available in the NULL word list using English-Tamil functional dictionary. Projection with partial morphological analysis in Tamil is used for the resource generation i.e. POS tagged sentences in Tamil. After alignment, original Tamil words are mapped with partial morphologically separated Tamil words, and POS and categorical information are applied. This is explained below

Words in English	:	He	conducted	an	Interview
Aligned Tamil words	:	அவன்	நடத்தினான்	ஒரு	நேர்முகத்தேர்வை
POS Projected	:	PRP	VBD	DET	NN
Categorical Information	:	NOUN	VERB	DETM	NOUN

Words in English	:	He	walked	towards	office
Aligned Tamil Words	:	அவன்	நடந்தான்	நோக்கி	அலுவலகத்தை
POS Projected	:	PRP	VBD	IN	NN
Categorical Information	:	NOUN	VERB	PREP	NOUN

G. *Need for Morphological Induction*

Rule based morphology in Tamil needs greater level of accuracy in stem identification. There are no accurate stemmers or lemmatizers in Tamil. Since Tamil morphology is based on stems, lemmatization plays an essential role in morphological analysis. For example, finding the stem using substring match from the beginning of the words "நடத்தினான்" (conducted) and "நடந்தான்" (walked), returns the same stem "நட" (walk). For the word "நடத்தினான்", the correct stem is "நடத்து" (conduct). Finding the partial stem after matching with suffix patterns at the end of words is also a problem. A dictionary with entries of partial stem and exact stem is needed for further morphological analysis. Creating such a dictionary is a tedious task. Due to these problems, accuracy of morphological analysis is reduced. Correct stem can be induced via alignment, lemmatization and induction processes through aligned corpora.

H. *Morphological Induction*

For the morphological induction, parallel dictionary is created between English and Tamil root words in the domain which is shown in table 7. It is created with the help of English-Tamil dictionary. Initially a word is matched with the functional word dictionary. If a match is found, the word is directly used in morphological analysis. Aligned English word for a Tamil word is selected and its exact stem or root word will be obtained with the help of the lemmatizer in English. Tamil root word will be induced by the simple lookup with the English stem in the parallel dictionary [4]. When multiple Tamil stems are matched with the English stem, substring match is done between actual Tamil word and induced Tamil stems. From that, exact Tamil stem is selected and used for further morphological analysis [20].

Table 7. Sample entries in root word dictionary

English Stem	Equivalent Tamil Stem
he	அவன்
conduct	நடத்து
interview	நேர்முகத்தேர்வு
walk	நட
office	அலுவலகம்

These alignment, lemmatization and induction processes are explained with examples shown below.

Tamil word	Aligned English word	Lemmatized English stem	Induced Tamil Stem
நடத்தினான்	→ conducted	→ conduct	→ நடத்து
நடந்தான்	→ walked	→ walk	→ நட
நேர்முகத்தேர்வை	→ interview	→ interview	→ நேர்முகத்தேர்வு
அலுவலகத்தை	→ office	→ office	→ அலுவலகம்

As an example, the Tamil word "நடத்தினான்" is aligned to the word in English "conducted". English lemmatizer produces its equivalent stem "conduct". Induction process induces the Tamil stem "நடத்து" for the English stem from parallel dictionary. Thus for the Tamil word "நடத்தினான்", stem "நடத்து" is obtained successfully through this morphological induction via aligned English-Tamil Corpora. After induction, the resultant stem can be matched partially with the word to ensure correctness. If it does not match then partial matching and lookup are done with the dictionary which contains partial verb or noun with actual.

Once categorical information and root word are obtained for a Tamil word, rule based morphological analysis and POS tagging can be done as described in section 4. This is shown with two examples below.

Tamil words	:	அவன்	ஒரு	நேர்முகத்தேர்வை	நடத்தினான்
Stem induced	:	அவன்	ஒரு	நேர்முகத்தேர்வு	நடத்து
Categorical Information	:	NOUN	DETM	NOUN	VERB
Refined Morphology	:	அவன்	ஒரு	நேர்முகத்தேர்வு+ஐ	நடத்து+இன்+ஆன்
Refined POS in Tamil	:	NOSM	DET	NNSNA	VTSMA

Tamil words	:	அவன்	அலுவலகத்தை	நோக்கி	நடந்தான்
Stem induced	:	அவன்	அலுவலகம்	நோக்கி	நட
Categorical Information	:	NOUN	NOUN	PREP	VERB
Refined Morphology	:	அவன்	அலுவலகம் +ஐ	நோக்கி	நட+த்(ந்)+த்+ஆன்
Refined POS in Tamil	:	NOSM	NNSNA	PRP	VTSMA

V. EXPERIMENTS

Sentence aligned Bible corpora contains 66 documents which contain 30,152 sentences each in both English and Tamil. By using Carnegie Mellon University statistical modeling toolkit, unique words in Tamil and English have been obtained. English has 16,546 words and Tamil has 95,116 words. This difference is due to high level inflections of Tamil.

In the rule based morphological analysis and POS tagging, 29,152 sentences were given for dynamic memory learning. 2000 sentences were taken as two test cases from Bible and CIIL corpora each with 1000 sentences and obtained the output. Gold standard has been prepared for these 2000 sentences and used for evaluation. Since sentence aligned corpora is available only for Bible domain, projection and induction processes have been done only with Bible sentence aligned corpora.

Alignment of 29,152 English and Tamil sentences from Bible parallel corpora has been done with GIZA++ toolkit which is mainly used for Statistical Machine Translation [19]. Alignment dictionary has been prepared from Bible Corpus which has more than 6000 words in Tamil and English. Lemmatization for finding the English stem for all unique words has been done with MBLEM Lemmatizer. Functional word dictionary has also been created. English-Tamil root word dictionary used for morphological induction has been created with more than 8000 entries using English-Tamil dictionary. Induction process has been done with in-house induction tool. Lookup process in various dictionaries is done with hash maps in order to reduce searching time. All the resources used for experiments are tabulated in table 8.

Table 8. Resources used

Details	Description
Corpora used	Bible aligned corpora (English-Tamil) CIIL corpus (Tamil)
No. of documents	66 (Bible corpora) 2 (CIIL corpus)
No. of sentences for alignment	29,152 (Bible corpora)
No. of sentences in test cases	1,000 (Bible corpora) 1,000 (CIIL corpus)
Unique words in aligned corpora	16,546 (English) 95,116 (Tamil) 33,176 (Tamil – after partial morphology)
No of sentences in gold standard	1,000 (Bible corpora) 1,000 (CIIL corpus)
Size of alignment dictionary	More than 6000 entries
Size of root word dictionary	More than 8000 entries
Tools used	GIZA++ (alignment) MBLEM (lemmatization)

VI. RESULTS AND DISCUSSION

In rule based morphological analysis and POS tagging, accuracy of 85.56% and 83% has been achieved for the test cases of Bible and CIIL corpora respectively. This moderate accuracy is due to partial stemming process and misidentification of exact root words. 50 test sentences with 288 words have been taken for testing alignment accuracy. Alignment accuracy has been obtained up to 72%. POS projection to Tamil words for all 30,152 sentences has been done successfully and categorical information has been obtained. Stems for Tamil words from English have been induced for all the words of Bible Tamil corpus. By using same 50 test sentences induction accuracy has been measured. Induction has achieved 70% accuracy. Again shortfall in this accuracy (30%) is due to the errors in alignment process and out of vocabulary in root word dictionary. When alignment accuracy is more, induction accuracy will also be more. This alignment accuracy is due to the morphological nature of Tamil.

A. Improvement in Rule based Morphology

In the rule based morphological analysis and POS tagging, 92.48% accuracy has been achieved from the same 1000 sentences used as test case by the use of categorical information and root words obtained from POS projection and morphological induction respectively.7% accuracy was improved in this hybridization of rule based approach and statistical approach of using alignment, projection, lemmatization and induction techniques. Results and improvement have been indicated in table 9.

Table 9. Results and Improvement

Experiment	Accuracy
Rule based morphological analysis and POS tagging	85.56% (Bible corpus test case) 83% (CIIL corpus test case)
Alignment	72%
Induction	70%
Improvement in rule based morphology and POS tagging	92.48 % (Bible corpus test case) (7% improvement)

VII. CONCLUSION AND FUTURE WORK

Rule based morphological analyzer and POS tagger have been built and tested successfully. POS projection through Bible sentence aligned corpora has been done and POS tagged Tamil sentences have been obtained as a resource. Rule based morphological analysis and POS tagging have been improved through stems and categorical information successfully. Significant improvement has been achieved through this hybridization of rule based approach with projection and induction techniques. The performance depends upon the alignment accuracy and root word dictionary. Since English Lemmatizer accuracy is very high in finding the root words in English it does not pose any problem to accuracy. Some of the issues to be resolved are disambiguation of participle noun and verb, similar adverb and preposition suffixes, adjective and noun clashes, multiple prepositional suffixes added in a direct preposition, wh-words with prepositional suffixes, and handling of prefix and circumfix in a word in Tamil. In future, context based disambiguation techniques can be used for further improvements.

ACKNOWLEDGEMENT

The authors would like to thank Central Institute of Indian Languages (CIIL), Mysore, India for providing the Tamil text corpora.

REFERENCES

[1] Eugene Charniak, “Immediate-Head Parsing for Language Models”, Proceeding of ACL, 2001.

[2] Ryan McDonald, Fernando Pereira, Kiril Ribarov and Jan Hajic, “Non-projective dependency parsing using spanning tree algorithms”, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, , 2005, pp. 523 - 530

[3] David Yarowsky and Grace Ngai, “Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora”, Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, 2001, p.1-8.

[4] David Yarowsky, Grace Ngai and Richard Wicentowski, “Inducing multilingual text analysis tools via robust projection across aligned corpora”, Proceedings of the first international conference on Human language technology research, San Diego, 2001, pp. 1 – 8.

[5] Rajendran. S, “Parsing In Tamil: Present State of Art”, Language in India, Vol. 6:8, 2006.

[6] Vasu Renganathan, “Development of Part-of-Speech Tagger for Tamil”, Tamil Internet 2001 conference, 2001.

[7] Rajan K, “Corpus analysis and tagging for Tamil”, Symposium on Translation Support systems, I.I.T Kanpur, 2002.

[8] Rajan K, Ganesan M and Ramalingam V, “Syntactic tagging of Tamil Corpus”, Tamil Internet 2002 Conference, 2002.

[9] Rajan K, Ganesan M and Ramalingam V, “Tamil Text Analyser”, Tamil Internet 2003 conference, 2003 .

[10] Daniel Jurafsky and James H. Martin, “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”, 2nd Edition, Pearson Education, 2006.

[11] Selvam.M, Natarajan A.M, and Thangarajan R "Lexicalized and Statistical Parsing of Natural Language Text in Tamil using Hybrid Language Models", WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 7, 2008, pp. 1362-1374

[12] Sandipan Dandapat, Sudeshna Sarkar and Anupam Basu, “Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario”, Proceedings of the ACL, 2007, pp. 221–224

[13] Singh. S, Gupta. K, Shrivastav. M and Bhattacharyya. P, “Morphological Richness Offset Resource Demand – Experience in constructing a POS Tagger for Hindi”, COLING/ACL, 2006, pp. 779 - 786.

[14] Manish Shrivastava, Nitin Agrawal, Bibhuti Mohapatra, Smriti Singh, Pushpak Bhattacharya, “Morphology Based Natural Language Processing tools for Indian Languages”, Workshop on Morphology, IIT Mumbai, 2005.

[15] Antal van den Bosch and Walter Daelemans, “Memory-based morphological analysis”, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99, University of Maryland, USA, June 20-26, 1999, pp. 285-292. (http://ilk.uvt.nl/mbma/)

[16] Rajendran S, “Strategies in the formation of compound nouns in Tamil”, Languages of India, Volume 4: 6, 2004.

[17] Rajendran. S, S.Viswanathan, and Ramesh Kumar, “Computational Morphology of Tamil Verbal Complex”, Language in India, Vol. 3:4,. 2003.

[18] Marcus, M. P., Santorini, B. And Marcinkiewicz, M. A. “Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics”, Vol. 19, 1993.

[19] F.J. Och and H. Ney, “Improved statistical alignment models”, in Proceedings of ACL-2000, 2000, pp. 440-447.

[20] Aniket Dalal, Kumar Nagaraj, Uma Sawant, Sandeep Shelke, Pushpak Bhattacharyya, “Building Feature Rich POS Tagger for Morphologically Rich Languages: Experiences in Hindi”, ICON, IIT Bombay, Mumbai, 2007.

M. Selvam received the B.E. degree in Computer Science and Engineering from Bharathidasan University, Trichy in 1990 and the M.E. Degree in Computer Science and Engineering from Bharathiar University, Coimbatore in 2002. He is currently pursuing Ph.D. Degree at Anna University, Chennai. He is currently working as Assistant Professor in department of Information Technology at Kongu Engineering College, Perundurai, Erode. He is working in number of projects on speech recognition and synthesis, statistical natural language processing and machine translation at Speech and Language Processing Lab at Kongu Engineering College. He is also a team member of research project sponsored by Tamil Virtual University, Chennai. His areas of interest are Speech and Language Processing, Computational Linguistics, Machine Translation and Enterprise Computing.

A. M. Natarajan is chief executive and professor in the department of Electronics and Communication Engineering at Bannari Amman Institute of Technology, Sathiyamangalam, Erode. He obtained Ph.D in Systems Engineering from P.S.G College of Technology, Coimbatore in 1984. He has published more than 100 papers in national and international journals and conferences. He was awarded “The Best Engineering College Principal Award” for the year 2000 by ISTE, New Delhi. He is the member of various scientific and professional societies. He has guided more than 75 M.E and M.C.A students. Presently he is guiding many Ph.D and M.Phil students. His research areas include Software Engineering, Soft Computing, Operating Systems, Software Project Management and Networking.