ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(23), DOI: 10.17485/ijst/2015/v8i23/79350, September 2015* ISSN (Online) : 0974-5645

Challenges in Morphological Analysis of Tamil Biomedical Texts

J. Betina Antony* and G. S. Mahalakshmi

Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai - 600025, Tamil Nadu, India; betinaantony@gmail.com, gsmaha@annauniv.edu

Abstract

The purpose of a Morphological analyser is to explore the internal structure of the word and retrieve grammatical features and properties of a morphologically inflected word. Breaking down these amalgamated words is in itself a challenging job in the field of Natural Language Processing. The complexity further increases when the analysis is done on a more ancient and morphologically rich dataset like Tamil Siddha Medicinal documents. In this paper we list the different challenges we faced when trying to explore the syntactic and semantic features of Tamil siddha texts for building a Tamil Biomedical NER. We also highlight the different fine tuning that was carried out on the analyser to overcome some of the difficulties and possible changes that can be done to improve the accuracy of the analyser in the given domain.

Keywords: Challenges, Morphological Analyzer, POS Tagging, Tamil Biomedicine

1. Introduction

One of the most ancient medical systems known, Siddha medicines originated in the southern part of India, from Tamil Nadu. It is a part of the trio Indian medicines– Ayurveda, Siddha, Unani. 'Siddhargal' or 'Siddhars' were believed to be the founders of this oldest system of medication. Thousands of texts produced by them laid the foundation for Siddha Medicine. The word 'Siddha medicine' means medicine that is perfect. Based on the mode of application, they are classified into 32 categories of internal medicines and 32 categories of external medicines.

Traditional Knowledge Digital Library is an Indian traditional knowledge repository containing mainly information about medicinal plants and formulations used in the Indian medical system. The main objective of the library is to protect the knowledge about traditional and ancient medical practices from bio-piracy and unethical patents. There are various text documents written on Siddha medicines based on the information gather from the ancient manuscripts obtained. Few books that are known are 'Pogar-7000' that deals with almost all subjects of Siddha medicine especially metals and minerals and 'A Scientific Journal from national Institute of Siddha' containing the scientific research oriented articles on Siddha medicines. Thus there are many sources of traditional medicine information. Hence it is essential to meticulously study the system to efficiently utilize these resources and provide effective usage of the information obtained.

Tamil, a morphologically rich language is agglutinated in nature. That is the root words are blended with postpositional or case markers to give semantically sophisticated words. These blended words carry number of information such as parts-of-speech tags, morphemes etc which are essential for any data analytics procedure. On the other hand, Tamil medical documents are mainly comprised of unstructured text with ample amount of rare terms (names of herbs, medicinal plants, different illness, disorders etc.). Automatic processing of these texts is still a challenging task in the field of Natural Language Processing (NLP). This is mainly due to the migration of interest of the people from traditional medicines

* *Author for correspondence*

to allopathic or modern medicines. Also high quality studies are essential to compare and evaluate the value of traditional Indian drugs.

We take this challenging dataset and start our work by pre-processing it using a morphological analyser[2]. The analyser tool is used to identify the morphological structure or break down of words. It can also be used to find the POS tags of word and case markers. The tool under consideration is a rule based tagger with a labelled dictionary. Our first encounter with this analyser started with the work to extract information from unstructured Siddha texts from blogs and published books[3]. Later in another task, we wanted to extract only noun terms which are the main candidates for finding named entities[4]. The tool though was able to identify most of the words although it started slipping in few main words. The tool was specifically challenging for Tamil siddha documents as most of the noun terms are names of herbs or diseases that are not common in our daily vocabulary. So this paper highlights the challenges that we faced while processing Tamil siddha documents in the context of finding named entity candidates and possible solutions if exits.

2. Related Work

The morphological richness of Tamil language caught the attention of researchers only in the last two decades. The analysis of Tamil language started with rule based systems. The initial works on Tamil concentrated on the basic preprocessing steps such as parts of speech tagging and morpheme extraction. Various model based approach were then developed to enhance the performance of efficient tagging. Once such model was a simple HMM based POS tagger[5] using viterbi algorithm were the tag of a word depends only on the previous word and its tag. Compound words were also considered which increased the tag set from 58 to 350. Similar model based tagger for Tamil employed linear programming using a SVM classifier which considered the tagging as a classifier problem. The system was built for 32 custom made tags and gave an overall accuracy of 95.6%. The works on POS tagging hit a new peak when a language model was built to fully utilize the morpheme characteristics of the language[6] with an accuracy as high as 96%.

As tagger started to gain its popularity, researches in Tamil also began to take huge leaps from simple POS features to other features of the language. Hence the complete purpose of a morphological analyzer came to lime light. The Early works in morphological analysis started with similar rule based techniques that were backed up by manually created dictionaries. The main decider in case of these tags were preceding and succeeding words. Their performance was enhanced by deploying projection and induction techniques[8]. Few machine learning based computational models for morphological analyzer were also built taking into account the psycholinguistic feature of the language and the comfort of formal methods[7]. Morphological analysis was perceived as a classification problem in one of the works[1] and was addressed by sequence labeling method using a SVM classifier. The result was found to perform well over CRF++ and Memory Based Tagger (MBT).

3. Challenges in Morphological Analyzer

To identify constituent terms for Named Entities from Tamil biomedical text, we filtered noun phrases extracted by the morphological analyzer[2]. The experiment was carried out for a set of Tamil biomedical documents containing about 32000 words. It is to be noted that the results may vary for the same text in different tools. However, only the outcome of the above mentioned analyzer is taken into consideration. Redundant words are accepted as a single word may be tagged differently in different places. Based on the study on the pattern of tagging, the extraction of noun phrases faces the following types of challenges.

3.1 Different Sense of Tags

The rich dataset of the analyzer allows correct tagging of a number of words. However not all tags fit the context of Tamil biomedicine. Hence few words are incorrectly tagged because of the ambiguous tags obtained. A common example for this condition is உப்பு: uppu (salt). உப்பு, a very common ingredient, has a number of medical values such as intravenous infusion and for killing bacteria. This term is tagged as a verb (to bloat) instead of noun (salt) in all the cases.

<உப்பு>: உப்பு < Verb and 200 > count=0

Similar examples include பால்: paal (milk) tagged as particle instead of noun and கிண்டி: kindi *(to mix)* is tagged a noun where it should be tagged as verb.

A special case found in this tagging is that few words are tagged as expected when they occur in their variant format. For example நீர்: neer (water) is tagged as pronoun whereas நீரை: neerai (water's) is tagged as noun + case. Similarly for பாலை: paalai (milk's) the expected tag here is,

<பாலை>: பால் < Noun and 100 > ஐ < Accusative Case and 500 > count=0

But the observed tag is ,

<பாலை>: பாலை < Noun and 100 > count=0

That is, பாலை: paalai is tagged as noun as in dessert and not noun (பால்: paal(milk)) + case where as பாலில்: paalil (in milk) is tagged correctly as noun + case.

<பாலில்>: பால் < Noun and 100 > இல் < Locative Case and 504 > count=0

3.2 Unknown Tags

Words that do not fit into any category are tagged unknown. This tag is usually found for words that are misspelled or words that are not found in the analyzer dictionary. The observed examples were ஓராயிரம், உண்ண, வப்ப etc. Certain common Tamil biomedical terms are also found to be tagged as unknown. Some of them are வயிற்றுப்புண், புன்னக்காய், சாரப்பருப்பு etc. The analyzer was not able to recognize the words even when they were added to the dictionary as the dictionary category is not known precisely

3.3 Untagged Words

Among the collection of words that are tagged, it was observed that few words were left untagged in between words in all their occurrences. Some of the words left are நோயின்றி, பாதாம், அகத்தி etc. When checked these words were found to be processed. But their tag was ambiguous and hence the analyzer left the words untagged.

Very few other word which were found to be tagged in few places but left untagged in other spots. The main reason behind this is unknown although it may be assumed to be due to preceding and succeeding words.

3.4 Incorrect Tags

A number of words were tagged with erroneous tags. The words however do not have same tag in all places they were tagged. Example அருகம்புல்: arukampul (Bermuda grass), though tagged correctly as noun in most of the places, had tag 'null அருகம்புல் < Noun and 100 >' if it occurred in the beginning of the document.

Few words were only partly tagged, that is only part of the word is tagged. For instance,

<குறுஞ்செடி>: செடி < Noun and 100 >

Here only the part செடி: chedi (plant) is tagged whereas the part குறும்: kuRum (small) is left untagged. This problem could not be resolved as there was no pattern in the words being wrongly tagged.

3.5 English Words

When it came to English words in between Tamil words or English words spelt in Tamil, different tags are observed. There is a separate dictionary having a list of words for Non-Tamil entities. But adding words simply to the dictionary did not solve the problem. Some of the common words and their tags are;

Table 1. Type of tags for English words

Word	Tag
<ஆஸ்துமா>:	ஆஸ்துமா < Noun & 100 > count=0
அலர்ஜி	No tag
<Sesbania>:	<Error>
<டீ>:	டீ < Non Tamil Noun & 107 > count=0
<வைட்டமின்>:	unknown

3.6 Split Tags

Very few words which were expected to be singly tagged were split into two. These words though had a meaningful tag when spilt did not give the expected biomedical named entity.

Adding these words to the entity list of the dictionary did not make any difference. Some very common words that encountered this problem are

- <அதிமதுரம்>: அதி < Noun and 100 > மதுரம் < Noun and 100 >
- <பூனிஏப்பம்>: பூனி < Noun and 100 > ஏப்பம் < Noun and 100 >

4. Conclusion

Rule based systems, though simple to construct, often suffer the problem of low accuracy as the rules have low coverage over previously unknown domain. It is hence crucial to move to model based approaches for improving performances. The condition holds good even in the case of morphological analyzer. However they do not perform badly in common sectors such as newswire or tourism. Hence we have addressed the open challenges faced in the given rule based analyzer in the context of finding named entities in Tamil biomedicine.

Few challenges were solved by ignoring the defects of the analyzer and taking the orthographic features of the language into consideration. Although these saw some tremendous improvement in accuracy of the system, few words ignored by the analyzer was never identified by any other features. Also ambiguity among words for the right tense still remains a problem. Hence the tags by the analyzer are not taken as the only criteria to segregate entity candidates. Another popular suggestion was to employ an ontology or dictionary to identify entities. This however broadens the problem as no known ontology for Tamil Biomedical texts is available as of now. Nevertheless works are carried out to exploit the tool at hand to its maximum capacity.

5. References

1. Anand Kumar M, Dhanalakshmi V, Soman KP, Rajendran S. A sequence labelling approach to morphological analyser for Tamil language. International Journal on Computer Science and Engineering. 2010; 2(06):1944–195.
2. Anandan P, Parthasarathy R, Geetha TV. Morphological Analyser for Tamil. ICON 2002; RCILTS-Tamil: Anna University, India; 2001.
3. Antony JB, Mahalakshmi GS. Patti Vaithiyam - An Information Extraction System for Traditional Tamil Medicines. 12th Tamil Internet Conference; 2013.
4. Antony JB, Mahalakshmi GS. Named entity recognition for Tamil biomedical documents. 2014 International Conference on Circuit, Power and Computing Technologies (ICCPCT); IEEE; 2014. p. 1571–7.
5. Palanisamy A, Lalitha Devi S. HMM based POS Tagger for a Relatively Free Word Order Language. Research in Computing Science. 2006; 18:37–48.
6. Pandian SL, Geetha TV. Morpheme based Language Model for Tamil Part-of-Speech Tagging. 2008. Available from: http://www. gelbukh.com/polibits/38_02.pdf
7. Ramaswamy V. A morphological analyser for Tamil. 2011.
8. Selvam M, Natarajan AM. Improvement of rule based morphological analysis and POS Tagging in Tamil language via projection and induction techniques. International Journal of Computers. 2009; 3(4).