Journal of Universal Computer Science, vol. 21, no. 13 (2015), 1830-1848
submitted: 4/5/15, accepted: 6/7/15, appeared: 28/12/15 © J.UCS

Content-based Information Retrieval by Named Entity Recognition and Verb Semantic Role Labelling

Betina Antony J
(Dept of CSE, CEG, Anna University, Chennai 600025, India
betinaantony@gmail.com)

G. Suryanarayanan Mahalakshmi
(Dept of CSE, CEG, Anna University, Chennai 600025, India
gsmaha@annauniv.edu)

Abstract: Tamil Siddha medicine, an ancient medicinal system has yielded us a wide range of untapped information about traditional medicines. In this paper, we explore into the various Natural Language Processing techniques that can be implemented to this syntactically rich corpus. As domain information mostly concentrates on the central concepts, we start our work by identifying the Named Entities and categorizing them. An integrated NER classifier is built which comprises of SVM and Decision Tree classifier with an accuracy as high as 95%. These entities play different roles in different context. Hence their roles are labelled along with the predicates surrounding them. These roles and predicates give rise to a rule based sentence tagging system, trained by an MEM model, to tag different contents in this otherwise unstructured text. These two important techniques are then exploited to develop our Information Retrieval System that combines the methods category tagging done by Named Entity Recognition and content tagging done by Semantic Role Labelling. The system takes full advantage of the rich features of the language and hence can be expanded to other domains.

Keywords: Information Retrieval, Tamil Siddha medicine, Named Entity Recognition, Semantic Role Labelling
Categories: H.3.1, H.3.3, I.2.7

1 Introduction

Siddha System of Medicine (SSM) is one among the oldest traditional systems of medicines discovered by ancient saints called Siddhars. These medicines formulated from herbs are still in use without any degradation in their medicinal value. Siddha System of Medicine is sometimes called “a boon for the rural poor”. Various informations on this venerable system of medicine were written in palm leaves in the age old days in a Tamil purer and more poetic than the current form of it. These informations were lost in time. In the recent times, there are various text documents written on Siddha medicines based on the information gathered from these ancient manuscripts obtained. Few books that are known are ‘Pogar-7000’ and ‘A Scientific Journal from national Institute of Siddha’. Valuable information about Tamil traditional medicines is also available in form of books, magazines and websites. These instructions are however enormous, unstructured and are still being discovered and translated by experts. Natural Language Processing plays a key role in structuring and ordering this silently growing domain information.

Processing of Tamil medical documents will be a challenging task in the field of Natural Language Processing (NLP). This is mainly due to the migration of interest of people from traditional medicines to allopathic or modern medicines. Also high quality studies are essential to compare and evaluate the value of traditional Indian drugs. NLP is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies.

Our work focuses on putting into use the dormant information available in piles by extracting useful information from the different sources. The first task in our Information Retrieval process is to identify named entities pertaining to the field of Tamil biomedicines. The Named Entities (NEs) refer to one or more rigid designators which includes proper nouns as well as certain kinds of natural terms such as biological species and substances. The ability of recognizing previously unknown entities is an essential part of Named Entity Recognition and Classification (NERC) systems. Such ability hinges upon recognition and classification rules triggered by distinctive features associated with positive and negative examples [Nadeau and Sekine, 07].

The next task is to label the different type of sentences based on the roles played by the verbal clause. The primary task of semantic role labelling (SRL) is to indicate exactly what semantic relations hold among a predicate and its associated participants and properties, with these relations drawn from a pre-specified list of possible semantic roles for that predicate (or class of predicates) [Màrquez et al., 08]. In order to accomplish this, the role-bearing constituents in a clause must be identified and their correct semantic role labels assigned.

The main challenge in our system is the domain in itself. Our system is the first of its kind as Tamil BioNLP is entering into the field of Computational Linguistics for the very first time. Both entity identification and role labelling are compelling tasks that need to be performed for Information Retrieval as entities and predicate labels give a definite meaning to the data in hand.

In a nutshell, our work instigates the remarkable use of named entities and semantic roles to retrieve information from this otherwise unregulated corpus. Our system focuses on using named entities and semantic roles to extract and structure information. This is achieved by constructing an Integrated Named Entity recognition (NER) identification module using Support Vector Machine (SVM) and decision tree training to extract information such as name of the medicine, disorder that is treated, ingredients used and preparation techniques from such documents. These named entities are stored in the NE Dictionary along with their categories and post processing heuristic rules. The system also performs semantic role labelling where the constituents of a verb clause are labelled. The role label pattern is further extended to identify the type of statement using a set of rules. These heuristic rules are also added to the dictionary. The NE's and heuristic rules thus obtained are applied for category tagging and content tagging respectively in the Information Retrieval system.

2 Related Work

Information Retrieval is considered a successful language processing technology to obtain information from unstructured text. The basic technique of Information

Retrieval (IR) paradigm is to extract entities by shallow analysis, recognize its references, update database and fill templates [Grishman R, 97].

2.1 Information Retrieval

The retrieval system saw many developments in the 70s and 80s. A number of models were developed and advances were made along all dimensions for document retrieval process. These new models were experimentally proven to be effective on small text collections. Their effect on large corpora was known in 1992 with the inception of Text Retrieval Conference, or TREC1 [Harman, 93]. TREC is a series of evaluation conferences which aims at encouraging research in IR from large text collections. With large text collections available under TREC, many old techniques evolved, underwent modifications and many new techniques were developed.

One common model that gained popularity for IR is Vector space model [Salton et al., 75]. In this model both query and the document are represented as a vector of term and the similarity between the vectors is measured based on the angle between them. As time progressed, the representation of documents and formulation of queries [Bai et al., 05] underwent tremendous changes and better systems were developed. [Billhardt et al., 02] developed a semantically rich representation of document by incorporating term dependencies and building context vectors for terms based on co-occurrences. However this model fails to give uniform results across different data sources and consumes more memory and time. A recent work on IR for Tamil [Premalatha & Srinivasan, 14] deploys vector space model on text processed using a database segregated into 5 components (Vowel – Kuril (Short), Nedil (Long); Consonant - Vallinam (Hard), Mellinam (Soft) and Idaiyinam (Medium)) instead of one. These five components highlight the morphological richness of the language.

Another very common model for Information Retrieval is the Probabilistic model where *probabilistic ranking principle* holds good [Robertson, 77]. The initial idea of probability started in 1960 [Maron & Kuhns, 60] where the probability of relevance of individual documents to a query is calculated and the document with highest probability is retrieved. Other models include graph based models and ontology based retrieval. In the graph based models [Blanco & Lioma, 12], documents are represented as a text graph and the ranking properties of graph theory such as average path length or clustering coefficient are used to rank results with query.

2.2 Named Entity Recognition

Though there are no actual works on traditional Tamil medicine, a number of NER systems have been developed for Biomedical and Clinical records in English. Current systems in NER employ mainly dictionary based, rule based, Machine Learning based, and hybrid approach. The Machine Learning methods have gained popularity these days due to their accuracy and efficiency to deal with complex annotated dataset [Nadeau and Sekine, 07]. These approaches have also been extended for English clinical and biomedical documents [Cohen & Hersh, 05].

A recent work on English Biomedical documents proposed a two phased approach using Semi- Conditional Random Field (CRF) with novel features to enhance the performance [Yang and Zhou, 13]. The first phase marked the term boundaries for a segment and the second phase involved semantic labeling. The addressing of features as segments was found more effective than single terms. Patrick et al., [05], suggested another Machine Learning approach using Maximum Entropy Model (MEM) where a blend of various linguistic features was incorporated to assign class labels and location within an entity sequence. A post-processing strategy for corrections to sequences of tags was then carried out. The experiment was carried out on the GENIA corpus.

Other Machine Learning approaches for NER include Hidden Markov Model (HMM) and Support Vector Machine. POSBIOTM-NER [Song et al., 04] is a Biomedical NER extraction system that uses SVM Machine Learning approach to build and expand a NER Dictionary by SVM training. This NER system adopts edit-distance measure, an additional input to resolve spelling variant problem.

A multi-strategy approach [Atkinson & Bull, 12] to recognize biomedical entities such as genes and proteins was proposed that combines SVM classifiers and HMM models, and simple linguistic pre-processing methods to automatically recognize gene and protein names from biomedical literature based on the standard Biocreative corpus. This approach does not make use of external knowledge bases such as ontology, lexicons. Hence they can be extended to other domains. Thus as far as NER is concerned, SVM classifiers exhibit a splendid performance when compared to the rest in the field of Biomedicine. Hence we used this classifier for our work to recognize and classify Tamil named entities.

2.3 Semantic Role Labelling

The idea of semantic roles for terms was introduced about 2 decades ago. Since then, the process of labelling these roles has acquired a special degree of interests among researchers. Early works in SRL started with a system which used a statistical classifier that was trained using sentences with hand-annotated semantic roles [Gildea & Jurafsky, 02]. The sentences were then parsed to syntactic trees from which various lexical and syntactic features such as phrase type, grammatical structure and position were obtained. Another SRL system was built using dependency trees which formulated the labelling process as the linear classification of dependency relations [Hacioglu, 04].

Semantic role labelling gained popularity in 2005 as number of models for SRL was developed as part of the CoNLL-2005 shared task. One such system was a joint model that captures dependencies among arguments of a predicate using log-linear model in discriminative re-ranking framework [Haghighi et al., 05]. Another interesting model which was built using different syntactic view used state-of-art based system using SVM classifier to eliminate errors due to parsing [Pradhan et al., 05]. The dependency tree saw few modifications that altered the performance of labelling. In one such model, Tree Conditional Random Field was employed to identify SRL by a system that defined a random field over the structure of each sentence's syntactic tree [Cohn & Blunsom, 05].

Semantic role labelling for biomedical documents was introduced by the system BIOSMILE [Tsai et al., 07] that uses of adverbial and prepositional phrases that are essential to identify biomedical relations. A Maximum Entropy (ME) Machine-Learning model is built to extract biomedical relations with automatically generated template features trained on semi-automatic, annotated biomedical proposition bank. The SRL task was extended to large corpora in SENNA, a fast and accurate neural network based labelling system with better accuracy than other SRL systems [Barnickel et al., 09].

Though there is no published SRL work on Tamil biomedicine, a semantic role labeller for Tamil documents was designed very recently. This system makes use of verb frame and MEM training module and an ensemble evaluator module to identify roles from input system [Pandian & Geetha, 09]. The system is optimized by EM classifier.

After a thorough study of existing work, we discovered that there is absolutely zero computational work being carried out in the field of Tamil biomedicine or Tamil siddha medicinal system. Hence we begin our work with Named Entity Recognition and Semantic role labelling to assist in the task of retrieving essential information from large collection of texts.

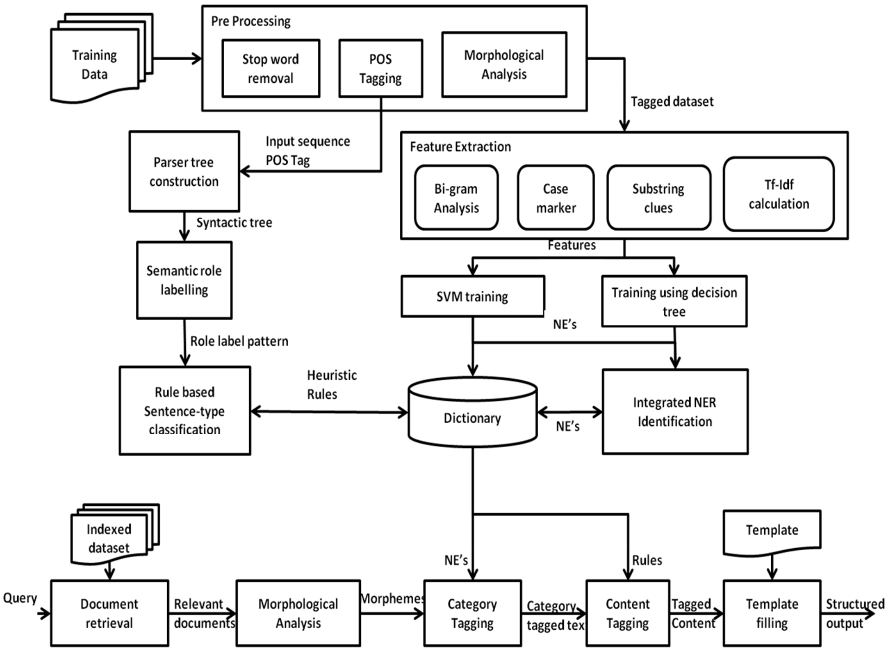


Figure 1: Tamil Biomedical Information Retrieval System – Block Diagram

3 System Description

The Block diagram of the complete Information Retrieval process is shown in figure 1. Information Retrieval in our context involves retrieval of documents relevant to the given query and effective representation of the extracted information. The process of segregating information from the retrieved documents is supported by two main NLP techniques. They are i) Category of Entity tagging done by Named Entity Recognition and ii) Content tagging supported by semantic role identification and labelling.

The Block diagram of the complete Information Retrieval process is shown in figure 1. Information Retrieval in our context involves retrieval of documents relevant to the given query and effective representation of the extracted information. The process of segregating information from the retrieved documents is supported by two main NLP techniques. They are i) Category of Entity tagging done by Named Entity Recognition and ii) Content tagging supported by semantic role identification and labelling.

3.1 Named Entity Recognition

As mentioned earlier, NER deals with identifying Named Entities from the given text. Named Entities in our domain includes terms of medicinal plants, herbs and their residues or names of disorders, diseases and related symptoms.

For eg. **முள்ளங்கியை சாறு எடுத்து சாப்பிட சிறுநீரக கோளாறு நீங்கும்.**

[முள்ளங்கி]_Ingredient யை சாறு எடுத்து சாப்பிட [சிறுநீரக]_B-Disorder [கோளாறு**]_O-Disorder**
*[muLLangki]*_Ingredient *yai sARu etuththu sAppita [siRuneeraka]*_B-Disorder *[kOLARu]*_O-Disorder
நீங்கும்.
neengkum.

Translation: The juice of radish can cure kidney stones.

In this example, **முள்ளங்கி (radish) and சிறுநீரக கோளாறு (kidney stones)** are named entities pertaining to Tamil Biomedicine. In our work two broad categories of NE are considered. They are Ingredients and Disorders. The process of Identifying Named Entities is carried out in 3 stages which are i) Pre-processing ii) Feature Extraction and iii) Training module [Betina & Mahalakshmi, 14]. In the training module, in addition to the SVM classifier, a decision tree based classifier is integrated to refine the final NE identification process.

Training module

After identifying features, the machine can now be trained to identify named entities by learning from the training set of data. All the features extracted so far are deployed to recognize entities and classify them. Named Entity Recognition in our system is done by two different Machine Learning methods

- SVM Training – Linear classification. Support Vector Machine is a non-probabilistic binary linear classifier. It is a robust Machine Learning algorithm that is designed for classification tasks based on large margin theory. It ignores the relationships between neighbouring tokens in sequence when applied in sequence labelling problems. This algorithm produces very accurate classifiers.
- Decision Tree Training – Non-Linear Classification. Trees can be used for regression or classification. Unlike linear regression, SVMs, naive Bayes, etc,

trees can fit local models. They are fast and produce interpretable predictions. In a Decision Tree, each internal node tests an attribute, each branch corresponds to attribute value and each leaf node assigns a classification.

We are employing a linear and a nonlinear classifier. The decision tree classifier orders the various features for optimum multistage decision making while the SVM classifier gives accurate results for the small sample dataset. The NER identification module is hence called Integrated NER as the results of both the models are combined based on their error prediction rate.

3.2 Semantic Role Labelling

Semantic roles are the underlying relation that a constituent has with the main verb in a clause. Tamil biomedical text consists of a selective set of statements and hence the roles of neighbouring candidate elements to the verb terms follow specific patterns. Hence to differentiate different sentences, we propose a set of rules based on the role pattern. In our system, the semantic roles are identified and labelled based on the features obtained from POS tags and syntactic tree structure. Some of the commonly identified roles are

- Agent – A0 – doer of action
- Patient – A1 – beneficiary of action
- Co-agent – A3 – actors assisting agent
- Theme – AM – entity affected by action
- Action – AC – action done

Eg. **அகத்திக்கீரையைப் பிழிந்து அதன் சாற்றில் இரு துளி மூக்கில் விட்டால் காய்ச்சல் நீங்கும்**.

[**அகத்திக்கீரையைப்**]_A0 [**பிழிந்து**]_AC [**அதன் சாற்றில்**]_A3 [**இரு துளி**]_AM [**மூக்கில்**]_A0
[*Akaththikkeeraiyaip*]_A0 [*pizhinthu*]_AC [*athan saaRRil*]_A3 [*iru thuLi*]_AM [*mookkil*]_A0

[**விட்டால்**]_AC [**காய்ச்சல்**]_A1 **நீங்கும்**.
[*vittaal*]_AC [*kaaychchal*]_A1 *neengum*.

Translation: Squeeze the leaves of Agati grandiflora, leave two drops of its juice in the nose to cure fever.

Before the learning process, the sentences are converted to a dependency tree (parse tree) [Nivre, 05]. Our tree is a modification of the original dependency tree that used arc standard principles of dependency grammar. In our dependency tree, the nodes represent terms along with their POS tags in a sentence (similar to phrase structure trees). The modification here is that the nodes branch (grow) only if it is a verb phrase. Thus the terms between a predicate term and its parent predicate forms a chunk of constituent phrases for role labelling. This carries the assumption that a verb phrase can act as a median between two different roles. A sample dependency tree is shown in figure 2.

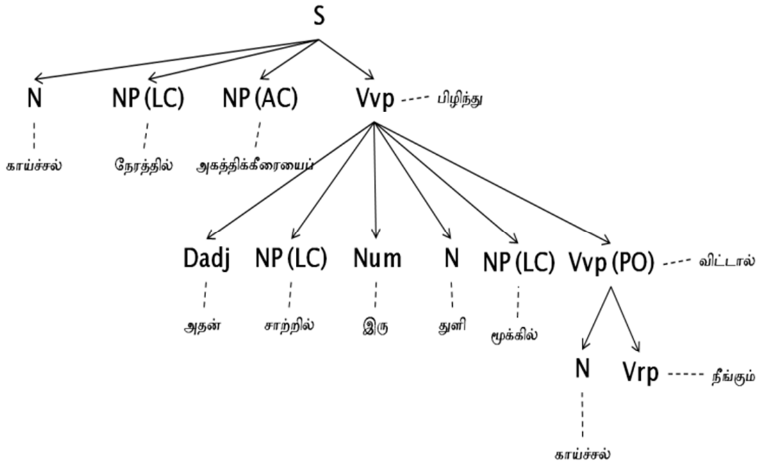


Figure 2: Sample Dependency Tree

A modified system of [Pandian S L and Geetha T V, 2009] 's SRL system is used by us for the field of Tamil Biomedicine. The main contribution of our SRL system is its ability to label verb phrases in addition to the standard roles identified. The verb phrases are usually tagged as a single role 'VBR-PHR' and their part in building the sentence is usually ignored. But in our case of gathering information from text, these verb phrases carry special clues that can be used to differentiate the type of sentences and labelling them accordingly. Thus a verb phrase (splitting node) in the context of Tamil Biomedicine can take up 4 different roles based on the surrounding noun phrase role pattern. The roles are i) Procedural verbs denoting procedure to prepare a medicine; ii) Solution verb to show solution to a particular disorder; iii) Information verb stating the significance of an ingredient or information about a disease and; iv) Verb that do not belong to any of these three. The process of semantic role labeling involves two types of learning. They are feature based and verb frame based learning.

i) MEM Learning

Maximum Entropy classifier is also known as multinominal logistic regression classifier. This is a probability model formed by conditional probability. Dependant variables (semantic role) are predicted based on Independent variables (features).

ii) Verb Frame Learning

Verb frame contains information about verb and the probability of semantic roles that can come with them. In verb frame learning we estimate the probability of a particular verb cluster by analyzing the corpus. All possible verb frames for a given verb class is analysed and a probability is assigned based on their count and slots filled.

Features that were considered for calculating the probabilities are shown in table 1.

Feature	Definition
POS of Term	Parts-of-speech tag of the constituent word is obtained
Headword POS	Parts-of-speech of the first word in the given phrase is obtained
No of Words in phrase	Total number of words in a given phrase
No of Noun Phrases	Total number of noun phrases
Predicate Distance	Distance of the constituent word from the predicate verb
Predicate POS	Parts-of-speech of the predicate verb
Category of Parent	Category of the parent phrase
Category Of Previous Term	Category of the previous term
Verb Frame Probability	Probability of the verb phrase. If the constituent term is a noun phrase, the probability is calculated as probability of the predicate / predicate distance.

Table 1: Features for SRL Learning

3.2.1 Rule based Sentence classification

Among the roles identified, all but one concentrates on noun phrases surrounding the predicate terms. In our system, we assign different roles for different verb phrases as well. This can be considered as an extension of semantic role labelling where the idea that even verb phrases play unique roles in different context is emphasised. For Tamil biomedical texts, almost all the sentences can be put into two broad categories. They are instruction sentences and information sentences. The former directs the reader as to what procedures should be done while the latter educates him about the various resources. These types of sentences can be identified using the verb phrases.

The roles played by these verbs can be identified by their suffixes. Hence we introduce a set of rules for these verb+suffix combinations and the possible roles that can apply to these verbs. For cases where more than one rule applies, a weighted score for the rules is given. For eg.

i) அரைத்து – *araiththu* - crush => <அரை+த்த்+உ> - <*aria+thth+u*>

அரை < Verb & 200 > த்த் < Past Tense Marker & 800 > உ < Verbal Participle Suffix & 900 >

ii) கலந்து – *kalanthu* – mix => <கல+ந்த்+உ> - <*kala+nth+u*>

கல < Verb & 200 > ந்த் < Past Tense Marker & 800 > உ < Verbal Participle Suffix & 900 >

Similarly in most of the activity related verbs, the verb term is followed by a verbal participle suffix. Hence we can label the verb as VRB-PRO denoting the role ‘Procedure verb’. Likewise a number of rules were written to denote four types of verb roles. They are VRB-PRO for ‘Procedure verb’, VRB-SOL for ‘Solution verb’, VBR-INFO for ‘Information verb’ and VBR-NONE for the ones that do not fall into any of these cases. The sentences are now classified based on the different role combination. Eg, “A0 + VBR-PRO + [A3] = procedure statement”. Here the procedural verb is surrounded by an agent and co-agents.

Eg. கோதுமையை வறுத்து பொடி செய்து கொள்ள வேண்டும். – Procedural sentence.

kOthumaiyai vaRththu podi seythu kola vENtum

[கோதுமையை]_A0 [வறுத்து]_VBR-PRO [பொடி செய்து]_VBR-PRO [கொள்ள வேண்டும்]_AC

[*kOthumaiyai*]_A0 [*vaRththu*]_VBR-PRO [*podi seythu*]_VBR-PRO [*kola vENtum*]_AC

Translation: Roast and grind wheat grains.

3.3 Biomedical Information Retrieval

Many works in Biomedical Information Extraction has been attempted in English language. Currently only one official work is recorded for Tamil [Antony & Mahalakshmi, 13]. Unlike English, biomedical acronyms and protein or enzyme names are not present in Tamil biomedical. However ambiguous words are a common challenge in extraction process. NER also plays an important role in Information Retrieval as searches can be made from NE perspective.

The main objective of our retrieval process is to obtain biomedicine based information from Tamil siddha texts based on the query given, extract the name of the medicine along with its cure and the process of preparing it and display the information extracted in a structured format. The steps involved in extracting and presenting query related information are shown below.

```
   Procedure for Information Retrieval from Tamil biomedical queries
 Assumptions:
 ◦   Biomedical documents are tagged based on the term frequency
 ◦   Dictionary contains list of Named Entities along with their
     categories and tag pattern denoting type of sentence
 Input: Query ( eg. இருமல் )
 Output: Medicinal information in structured format
 Step 1: For a given query, retrieve relevant document from indexed
dataset (Tf-Idf indexing)
 Step 2: Pre-process the retrieved document to obtain POS tags and
morphemes
 Step 3: Identify the named entities and tag their category using NE
dictionary.
 Step 4: Tag the sentence types based on the heuristic rules from the
Dictionary
 Step 5: Feed Information to template.
```

4 Experimental results and analysis

The experiment was conducted for a set of documents with an average size of 8 Kilo Byte text documents with approximately 400 terms in each document. The corpus had about 502 files that included 135 ingredients based and 84 disorder based information documents. In addition, an unbiased set of files were also included to avoid favouring one set of data. When a bag-of-words form of the corpus was taken, the number of unique noun terms was more than 1150 despite the poor tagging of the Analyser (Anandhan P et.al., 2002).

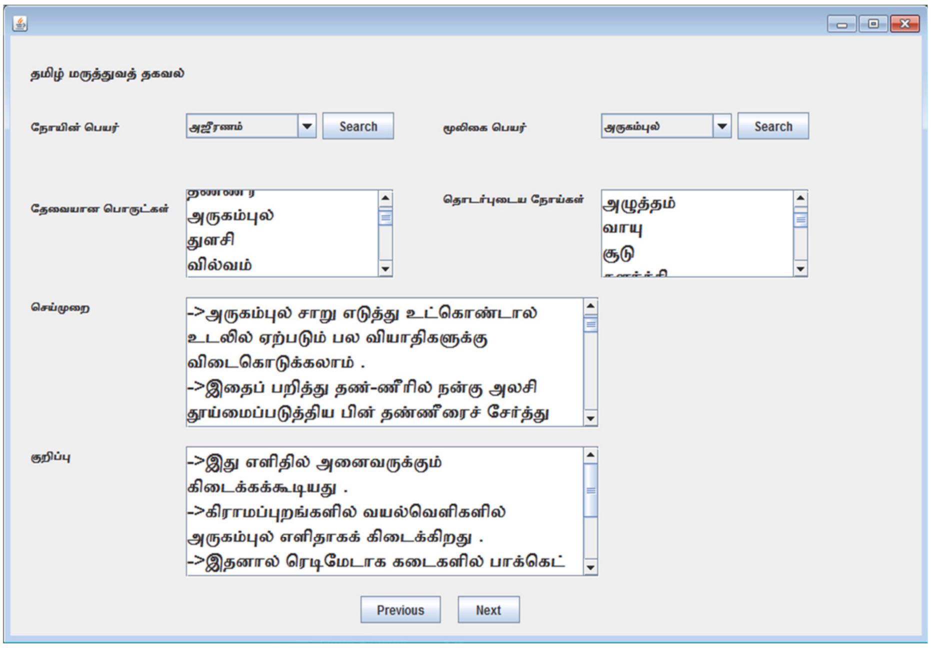
4.1 Biomedical Information Retrieval System

The arduous part in the given framework is tagging of the keywords into their respective categories and classifying sentences. The category and content tagging are assisted by named entity recognition and semantic role respectively. The framework was built to answer about 274 disorder related queries and 299 ingredient related queries. The steps involved in the retrieval process are shown in section 3.3.

Figure 3 gives a sample output screen of the retrieval system. The two types of queries are given as dropdown list to avoid difficulties in typing Unicode symbols. The list of ingredients used and other related disorders that can be treated using the given set of ingredients are shown separately. The preparation procedure is shown in the column செய்முறை: *SeimuRai* (Procedure). Other related information and tips are given in the column குறிப்பு: *kuRippu* (Instructions).

The list of ingredients and disorders tagged are evaluated for random queries based on their precision and recall value. The queries included 65 ingredient and 60 disorder names which included unigrams and bigrams. Each query may return more than one relevant document. To evaluate the whole the system for accuracy of retrieval, the precision values for individual fields such as required ingredients, related diseases and procedural and instruction sentences was calculated. The Average Precision (AP) value is taken as the sum of precision values divided by the number of relevant documents retrieved for a given query. Now the Mean Average Precision (MAP) is calculated by adding the AP's of relevant documents for all queries and dividing it by the number of queries checked. The MAP for our system is 72.2% when evaluated for the above mentioned number of queries.

The precision value of the NE fields was found to be slightly higher than the recall value. This is because of the dictionary based tagging and most of the elements tagged are relevant. However the recall value was found to drop quite heavily. This may be due to 2 main reasons. Firstly the comparison was done only for unigrams. Hence doing the bigram comparison might increase the sensitivity of the system by at the least 20%. Secondly, duplicates were not eliminated from the tagged elements. Hence removing duplicates will definitely boost up the recall. The system can be enhanced to filter non relevant portion from the text as the corpus contains both disease related and ingredient based documents. This in turn reduces the amount of noise added to the result produced thus improving the efficiency of the system. The recall values for random 10 queries each for list of ingredients and disorders are shown in figure 4 (a) and (b) respectively.



தமிழ் மருத்துவத் தகவல்

நோயின் பெயர்: அஜீரணம்

மூலிகை பெயர்: அருகம்புல்

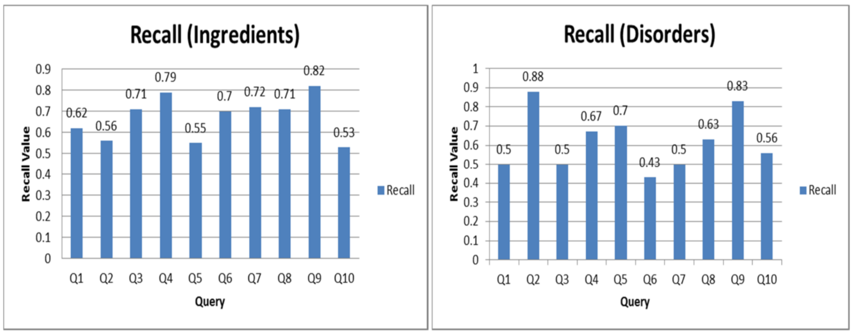
தேவையான பொருட்கள்:
அருகம்புல்
துளசி
வில்வம்

தொடர்புடைய நோய்கள்:
அழுத்தம்
வாயு
சூடு

செய்முறை:
->அருகம்புல் சாறு எடுத்து உட்கொண்டால் உடலில் ஏற்படும் பல வியாதிகளுக்கு விடைகொடுக்கலாம் .
->இதைப் பறித்து தண்-ணீரில் நன்கு அலசி தூய்மைப்படுத்திய பின் தண்ணீரைச் சேர்த்து

குறிப்பு:
->இது எளிதில் அனைவருக்கும் கிடைக்கக்கூடியது .
->கிராமப்புறங்களில் வயல்வெளிகளில் அருகம்புல் எளிதாகக் கிடைக்கிறது .
->இதனால் ரெடிமேடாக கடைகளில் பாக்கெட்

Figure 3: Sample Retrieval screen for an ingredient query அருகம்புல் aRukampul(Bermuda grass)



(a) **(b)**

Figure 4: a) Recall graph for list of Ingredients. b) Recall graph for list of Disorders.

In the case of content tagging, the sentences are tagged based on the number of predicate terms, their labels and the surrounding agents. The efficiency of the tagging is found to be quite less as most of the sentences fall into both the classes and there is no concrete distinguishing factor to separate them. Also in the current system, all sentences in the retrieved document are displayed. This might contain information not related to the given query. Hence scrutinizing of these sentences can improve efficiency of content retrieval drastically.

4.2 Feature Extraction for NER

The first step of operation is to extract features for Named entity identification and classification. The features used in our system falls into three main categories; Frequency, position and sense based. The four features that were extracted are Bigrams scoring, *Tf-Idf* based scoring, Substring clues and Case markers bases scoring. The results of features extracted [Betina & Mahalakshmi, 14] are shown in table 2.

Feature	Category	Precision	Recall	F-score	Sensitivit y	Specifici ty
Bigram - Analysis	Before Refinement	0.20	-	-	-	-
	After Refinement	0.59	-	-	-	-
Tf-Idf based scoring	Unigram	0.80	-	-	-	-
	Bigram (Ingredients)	0.67	-	-	-	-
	Bigram (Disorder)	0.51	-	-	-	-
Substring clues	Before Refinement (Ingredient)	0.86	0.42	0.56	-	-
	After Refinement (Ingredient)	0.89	0.49	0.61	-	-
	Before Refinement (Disorder)	0.53	0.58	0.55	-	-
	After Refinement (Disorder)	0.53	0.4	0.46	-	-
Case Markers	Ingredient	-	-	-	0.632	0.0018
	Disorder	-	-	-	0.43	0.0047

Table 2: Evaluation Results for Feature Extraction

4.3 NER Classification

In the second part of the experiment, training dataset containing total of 792 instances each provided with values for 5 attributes which are unigram/bigram, case marker score, substring clue, *tf-idf* score and class of entity are given. The instances are training using a SVM classifier and J48 decision tree classifier.

Of the 792 terms, the actual number of ingredients are 404, number of disorders are 331 and number of unknown terms are 57. The Ingredients and Disorders have higher f-score of 96.6% and 95.9% respectively when compared to the unknown terms with f-score 66.2%. This is because the number of training dataset for 'none' terms is very less when compared to the other two. The overall f-score of decision tree is found to be only about 0.2% lesser than that obtained by SVM classifier as both the systems are dealing with mainly binary classification. The average Precision, Recall and F-score of this classifier are 95.0 %, 95.4 % and 95.1 % respectively.

The model was evaluated for 4 types of test document; an ingredient based document, disorder based document, a document having combined information and a document with random information. The two models were tested with nearly 200 constituent terms. The classification was found to be similar in most of the cases that is the same term was tagged either correctly or incorrectly by both the models. Hence it can be observed that for the given features, both SVM and decision tree classification may provide exact same results. However SVM classification was found to have a slight edge over Decision tree in tagging terms correctly proving that SVM is a more accurate classifier for lesser number of classes. The details are shown in figure 5.

The integrated NE model combines the output of both SVM based and Decision Tree based classification and returns entities with an error prediction rate of 80% (threshold).

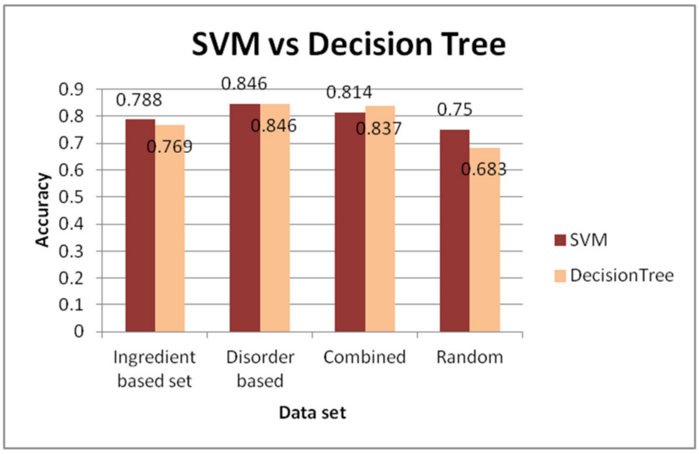
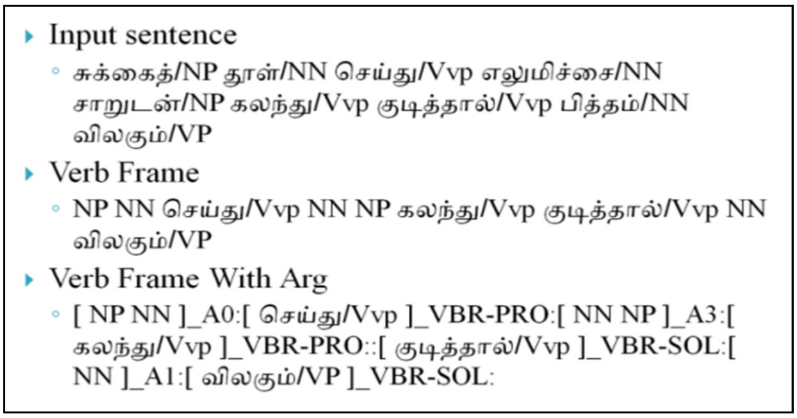


Figure 5: SVM vs Decision Tree Accuracy graph.

4.4 MEM model for Semantic Role Labelling

The semantic role labelling is done in three stages. The first stage is to construct the dependency tree using the POS tags. Next is the learning step where an evaluator model is learnt based on the features and verb frames. Finally a MEM evaluator model is constructed. The steps are shown in algorithm 1. The verb phrases thus obtained are further classified into four types based on their sense. They are procedural, solution related, informational and not applicable. Thus in case of Tamil biomedical information, a sentence falls into any of the four categories based on the position, count and neighbours of these verb phrases.

Verb Frame identification and training involves the following steps. For each verb, its corresponding synonyms are identified from the Wordnet. The verb classes are formed by combining terms with similar meaning. This is to imply that any term 'a' in a class will satisfy a verb frame of every other term in the same class. All the sentences containing a verb are grouped together for each class. For these classes, the possible verb frames are obtained. Verb frames are nothing but sentences with verb and possible combination of other patterns. These patterns are obtained from the training documents. Finally roles are assigned to the slots before and after each predicate term based on the type of verb in question. The value of verb label is obtained from a listing file prepared prior to this operation based on their frequency, position and sense. A sample output of the above steps is shown in figure 6.



- Input sentence
  - சுக்கைத்/NP தூள்/NN செய்து/Vvp எலுமிச்சை/NN சாறுடன்/NP கலந்து/Vvp குடித்தால்/Vvp பித்தம்/NN விலகும்/VP
- Verb Frame
  - NP NN செய்து/Vvp NN NP கலந்து/Vvp குடித்தால்/Vvp NN விலகும்/VP
- Verb Frame With Arg
  - [NP NN]_A0:[செய்து/Vvp]_VBR-PRO:[NN NP]_A3:[கலந்து/Vvp]_VBR-PRO::[குடித்தால்/Vvp]_VBR-SOL:[NN]_A1:[விலகும்/VP]_VBR-SOL:

Figure 6: Sample Role Labelling for a given sentence

The values obtained from both the training process are combined into a single training set and a classifier based on Maximum Entropy model is obtained. The training dataset contained total of 1268 instances each provided with values for 9 attributes and the label. The classification system is expected to classify the terms into A0 (Agent), A1 (Patient), A3 (Co-agent) and verb phrases (VBR-INS, VBR-SOL, VBR-PRO, VBR-NONE). The overall f-score was found to be about 80.2%. The details of the classifier output are shown in table 3.

Class	Precision	Recall	F-score
A0 (Agent)	0.836	0.818	0.827
A1 (Patient)	0.779	0.698	0.736
A3 (Co-Agent)	0.674	0.738	0.705
VBR-INS	0.800	1.000	0.889
VBR-SOL	0.907	0.961	0.933
VBR-PRO	1.000	0.909	0.952
VBR-NONE	0.875	0.977	0.923

Table 3: Classifier output of SRL

The SRL evaluator was tested for 3 different types of input documents. The predicted role labels were compared with the actual label and their precision value was calculated. The results are shown in Figure 7. Of the different datasets, the ingredient based data had a higher precision value when compared to others. This may be due to the biased training set that was used which was mostly ingredient related. Also some of the rules were written with some prediction error. For instance, the sentence format in Tamil is 'SOV' (subject + object + verb) and so the NP preceding the verb will always be 'A1' (patient). Once the roles are labelled, the sentences are put into different categories based on the verb roles. This step is done in the content tagging of information extraction module.

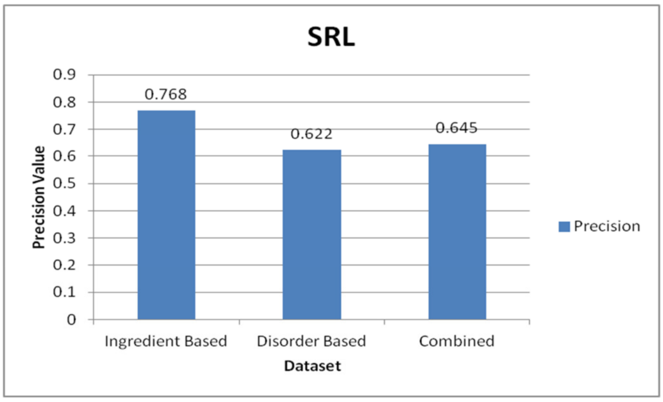


Figure 7: Precision Graph for SRL.

4.5 Challenges

The overall efficiency of the entire system is brought down by one main factor which is the improper tagging of the Analyzer [Anandhan P et.al., 02]. The main reason for the fall in f-score is found to be the absence of the medicinal terms in the analyzer dictionary. For eg. குறிப்பு: *veRRilai* (Betel leaves), a common word is tagged <unknown> by the morphological analyser. Another reason for incorrect extraction is due to the splitting of a single word into two words. For eg. அதிமதுரம்: *athimathuram* (Liquorice) is split into two separate nouns, அதி: *athi* and மதுரம்: *mathuram* by the analyser. Hence a single ingredient is split into two and thereby is tagged incorrectly by the analyser. These kinds of errors can be rectified by adding the unknown words and the compound nouns to the analyser dictionary. Tagging the correct sense of certain words in the given context also reduces the accuracy score. For eg. நீர்: *Niir* (water) and நெய்: *ney* (Ghee) are proper noun denoting elements used for medicine preparation. These are however tagged as pronoun and verb respectively by the analyser.

5 Conclusion

In order to exploit and explore a voluminous growing collection of data, an Information Retrieval system is built that can obtain information from any untrained and untreated dataset. The given system is trained to extract named entities from unstructured biomedical documents using integrated SVM and Decision Tree classifier. The SVM classifier model has a slight edge over other classifiers for the given set of features since the number of classes is less and fairly far apart. The role labels identified roles such as the agent, patient and co-agents involved and classified sentences based on the verb phrases and their roles. Its efficiency can be improved extensively by introducing certain Machine Learning techniques over rule based learning. Also a number of other roles can be identified.

The field of Tamil Biomedicine is new to the field of datamining. Hence a number of other NLP procedures can be carried out for the given dataset. Various other functionalities include resolving non-Tamil words, semantic analysis, anaphoric resolution, co-occurrence analysis etc. All these can be used to distinctly identify named entities and classification. Our system confines to only two type of classes which are Ingredients and Disorders. More NE classes can be identified such as symptoms, measures etc. In addition, Word Sense Disambiguation can be done before the actual NE recognition to avoid ambiguous terms.

References

[Anandhan P et.al., 02] A Anandan, P., Saravanan, K., Parthasarathi, R., Geetha, T.: “Morphological analyzer for Tamil”; International Conference on Natural language Processing; 2002.

[Antony & Mahalakshmi, 13] Antony, J. B., Mahalakshmi, G.: “Patti vaithiyam—an information extraction system for traditional Tamil medicines”; Proceedings of the Twelfth International Tamil Internet Conference, INFITT [2013], 125-131.

[Atkinson & Bull, 12] Atkinson, J., Bull, V.: “A multi-strategy approach to biological named entity recognition”; Expert Systems with Applications; 39 [2012], 17, 12968–12974.

[Bai et al., 05] Bai, J., Song, D., Bruza, P., Nie, J.-Y., Cao, G.: “Query expansion using term relationships in language models for Information Retrieval”; Proceedings of the 14th ACM international conference on Information and knowledge management; 688–695; ACM, 2005.

[Barnickel et al., 09] Barnickel, T., Weston, J., Collobert, R., Mewes, H.-W., Stümpflen, V.: “Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts”; PLoS One; 4 [2009], 7, e6393.

[Betina & Mahalakshmi, 14] Betina Antony, J., Mahalakshmi, G. S., "Named entity recognition for Tamil biomedical documents." In *Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on*, [2014], 1571-1577.

[Bhakkad et.al., 13] Bhakkad, A., Dharamadhikari, S., Kulkarni, P.: “Efficient approach to find bigram frequency in text document using E-VSM”; International Journal of Computer Applications; 68 [2013], 19, 9–11.

[Billhardt et al., 02] Billhardt, H., Borrajo, D., Maojo, V.: “A context vector model for Information Retrieval”; Journal of the American Society for Information Science and Technology; 53 [2002], 3, 236–249.

[Blanco & Lioma, 12] Blanco, R., Lioma, C.: “Graph-based term weighting for Information Retrieval”; Information retrieval; 15 [2012], 1, 54–92.

[Cohen & Hersh, 05] Cohen, A. M., & Hersh, W. R.: "A survey of current work in biomedical text mining." *Briefings in bioinformatics* 6.1 [2005], 57-71.

[Cohn & Blunsom, 05] Cohn, T., Blunsom, P.: “Semantic role labelling with tree conditional random fields”; Proceedings of the Ninth Conference on Computational Natural Language Learning; 169–172; Association for Computational Linguistics, 2005.

[Gildea & Jurafsky, 02] Gildea, D., Jurafsky, D.: “Automatic labeling of semantic roles”; Computational linguistics; 28 [2002], 3, 245–288.

[Grishman R, 97] Grishman, R.: “Information extraction: Techniques and challenges”; Information extraction a multidisciplinary approach to an emerging information technology; [1997] 10–27.

[Hacioglu, 04] Hacioglu, K.: “Semantic role labeling using dependency trees”; Proceedings of the 20th international conference on Computational Linguistics; 1273; Association for Computational Linguistics, [2004].

[Haghighi et al., 05] Haghighi, A., Toutanova, K., Manning, C. D.: “A joint model for semantic role labeling”; Proceedings of the Ninth Conference on Computational Natural Language Learning, [2005], 173–176.

[Harman, 93] Harman, D.: “Overview of the first text retrieval conference (trec-1)”; Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, USA, (1993), 36–47.

[Maron & Kuhns, 60] Maron, M. E., Kuhns, J. L.: “On relevance, probabilistic indexing and Information Retrieval”; Journal of the ACM (JACM); 7 [1960], 3, 216–244.

[Màrquez et al., 08] Màrquez, L., Carreras, X., Litkowski, K. C., Stevenson, S.: “Semantic role labeling: an introduction to the special issue”; Computational linguistics; 34 [2008], 2, 145–159.

[Nadeau and Sekine, 07] Nadeau, D., Sekine, S.: “A survey of named entity recognition and classification”; Lingvisticae Investigationes; 30 [2007], 1, 3–26.

[Nivre, 05] Nivre, J.: "Dependency grammar and dependency parsing." *MSI report*5133.1959 [2005], 1-32.

[Pandian & Geetha, 09] Pandian, S. L., Geetha, T.: “Semantic role labeling for Tamil documents”; Int. J. Recent Trends Eng; 1 [2009], 1.

[Patrick et al., 05] Patrick, J., Wang, Y.: “Biomedical named entity recognition system”; Proceedings of the Tenth Australasian Document Computing Symposium (ADCS 2005); [2005].

[Pradhan et al., 05] Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H., Jurafsky, D.: “Semantic role labeling using different syntactic views”; Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; [2005], 581–588.

[Premalatha & Srinivasan, 14] Premalatha, R., Srinivasan, S.: "Text processing in Information Retrieval system using vector space model"; Proceedings of International Conference on Information Communication and Embedded Systems (ICICES), [2014], 1–6.

[Robertson, 77] Robertson, S. E.: "The probability ranking principle in ir"; Journal of documentation; 33 [1977], 4, 294–304.

[Salton et al., 75] Salton, G., Wong, A., Yang, C.-S.: "A vector space model for automatic indexing"; Communications of the ACM; 18 [1975], 11, 613–620.

[Song et al., 04] Song, Y., Yi, E., Kim, E., Lee, G. G., Park, S.-J.: "Posbiotm-ner: a Machine Learning approach for bio-named entity recognition"; Korea; 305 [2004], 350.

[Tsai et al., 07] Tsai, R. T., Chou, W.-C., Su, Y.-S., Lin, Y.-C., Sung, C.-L., Dai, H.-J., Yeh, I. T., Ku, W., Sung, T.-Y., Hsu, W.-L.: "Biosmile: a semantic role labelling system for biomedical verbs using a Maximum-Entropy model with automatically generated template features"; BMC bioinformatics; 8 [2007], 1, 325.

[Yang and Zhou, 13] Yang, L., Zhou, Y.: "Exploring feature sets for two-phase biomedical named entity recognition using semi-crfs"; Knowledge and information systems; 40 [2014], 2, 439–453.