Patti Vaithiyam — An Information Extraction System for Traditional Tamil Medicines

J. Betina Antony¹ and G.S.Mahalakshmi²

¹Department of Computer Science and Engineering, Anna University
Chennai, Tamil Nadu, India
²betinaantony@gmail.com
²gsmaha@annauniv.edu

Abstract — This work attempts to extract traditional medicine related information from Tamil Siddha documents generally published as ‘health tips’ in websites, blogs and magazines. Sometimes a sub-section of it is also popularly known as ‘Paatti Vaithiyam’. It extracts names of the traditional medicines that are mentioned in the unstructured documents. The idea is to record the name of the items as well as the cure in connection with the process of preparing the home remedy. The output is obtained in a structured format that is easily understandable.

Keywords — Information extraction, Tamil Siddha documents, traditional medicine, Paatti Vaithiyam

I. 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. The 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 also various text documents written on Siddha medicines based on the information gathered 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 orientated articles on Siddha medicines. Thus there are many sources of traditional medicine information. Hence it is essential to create an Information Extraction (IE) system to efficiently utilise these resources and provide effective usage of the information obtained.

Tamil medical documents are mainly comprised of unstructured text and 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 to allopathic or modern medicines. Also high quality studies are essential to compare and evaluate the value of traditional Indian drugs. Our system tries to retrieve these unstructured documents and process them to obtain necessary and valuable information.

This paper is organized as follows: Section II gives an overview on the Information Extraction systems for the Biomedical domain in various other languages. Section III provides details about the approach suggested. Section IV deals with the experiment results and evaluation while Section V gives the conclusion and some discussions for future work.

II. RELATED WORK

Information Extraction is considered a successful language processing technology to obtain information from unstructured text. The basic technique of IE paradigm is to extract entities by shallow analysis, recognize its references, update database and fill templates [6]. The shallow analysis usually involves pattern matching in regular expressions.

Though there are no actual works on traditional Tamil medicine, a number of information extraction works have been carried out for Biomedical documents [10][7] and Clinical records[15][8][16] by string matching or rule based methods. One such information extraction system is MedEx [16], which identifies the name of the medicines and also signature information such as strength, route and frequency from discharge summaries. One important challenge in these clinical records is the heterogeneity of data (narrative or coded).
IE finds its application in many prototypes which are used in the extraction process:

A. Concept Extraction

MetaMap[3], an effective mapping procedure, makes use of knowledge intensive approach and computational linguistic techniques to map biomedical text to the UMLS Metathesaurus. The biomedical texts contain classified concepts and hierarchical relationships. The two main problems faced by this system are detection of idiosyncratic text and resolution of ambiguity. Clashing concepts can also be resolved by classification based on semantic types.

B. Named Entity Recognition (NER)

The concept extraction process is supported by a several Named Entity Recognition (NER) systems. Various approaches for NER include Hidden Markovian Model (HMM) [14][4][17] and Support Vector Machine (SVM)[9][5]. POSBIOTM-NER[13] 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.

C. Word Sense Disambiguation (WSD)

Word Sense Disambiguation (WSD) is the task of selecting the appropriate sense for words in the given context. A number of graph-based approaches have been identified to identify the intended meaning of words in a context [12] that investigates the connectivity measures of the graph and identifies how this method performs comparably to the state-of-art. One such method that makes use of UMLS MetaThesaurus and Personalized PageRank, a state-of-the-art algorithm is explained in [1]. Disambiguation here is done by converting the tables from the thesaurus into graphs and applying the algorithm to select best sense for each ambiguous word.

III. MATERIALS AND METHODS

A. Medical Documents

The medical documents are the text documents containing the remedial instructions for a particular disorder. The documents contain the following tags:

1) \(<\text{சேந்தி}>\) : This tag contains the name of the disorder for which the cure is suggested. The contents of this tag may be more than one.

2) \(<\text{மாந்தி}>\) : This tag provides the remedial measures to be taken for the given disorder. The instruction is composite from which various other informations such as ingredients used, special notes, medicine preparation procedure and dosages instructions can be obtained.

B. Methods

a) Document Retrieval:

Document retrieval is the process of retrieving relevant documents for a given query. The query may be a set of keywords and the relevant documents are text documents whose contents contain these keywords or related words [11].

For a given query (say அஜீரணாள்), the relevant documents are obtained by matching the elements in the \(<\text{சேந்தி}>\) tag. Eg. If \(<\text{சேந்தி}>அஜீரணாள்</\text{சேந்தி}>\) is present for the query, the document is retrieved.

b) Morphological Analysis:

Morphological analysis is the process by which individual words are analysed into their components and non-token words such as punctuations are separated from the words. The process assigns syntactic categories to all the words in a sentence.

For tamil, morphological analysis can be done using a tool called ‘Atcharam’ [2]. Eg. For a word குற்றக்குடியைக், the analysed words will be

\(\text{குற்றக்குடியை} <\text{Noun & 100}>\)
c) **Category Tagging:**

Categorizing a set of text into predefined classes is called text categorization. Text categorization can be done by many machine learning techniques. However, we employ manual mechanism to tag keywords into name of disease, ingredient or other.

Eg. In the sentence ‘நாய் நல்லுச் செய்வோம் ஏலாகா இல்லாத மாடங்கியே’, நாய் falls into the category of disease and ஏலாகா falls into category of ingredient. Rest are not keywords hence are tagged as other.

**C. Assumptions**

Few assumptions to be considered in this system are:

i. Biomedical documents are manually tagged and stored based on the name of the disease and name of the core medicinal ingredient.

ii. NE Dictionary contains list of named entities that include name of the disorders and also names of a variety of medicinal elements.

iii. Anaphoric pronouns are avoided in the instructions to avoid co-reference ambiguities.

**D. Algorithm:**

The algorithm for the proposed system is as follows

**STEP 1:** Retrieve the relevant tagged documents on receiving a query.

**STEP 2:** For each document retrieved,

2a: Do Morphological Analysis using an Analyser to obtain morphemes.

2b: Do Category Tagging using the NE Dictionary.

**STEP 3:** From the category tagged document, Assign

3a: the element in tag <DD> to name of the disorder

3b: the elements in tag <MM> to the list of ingredients required

3c: the contents of tag <மாடங்கியே> from original document to the procedure.

**STEP 4:** Feed the Information to the template. Output is obtained in a structured format.

Fig. 1 gives the overall working procedure for the system.

---

Fig. 1 Traditional Medicine Information Extraction
IV. RESULTS AND DISCUSSION

A. Dataset
The information extraction routine was carried out for nearly 160 documents covering over 30 different common disorders ranging from simple disorders like hiccups (விக)` or dry cough (வறงาม)` to diabetes (சகைர or நீங்கலை இழிகேணலை).

B. Sample Query:
The information extracted for a given query is depicted in the following snapshots.

For a given query - அஜீரணா, the relevant tagged document is retrieved (Fig.2).

![Fig. 2 Retrieve relevant document for the given query](image)

The instructions in the document are set as input to the Morphological Analyser. Fig.3 shows the analysis output and the morphemes obtained.

![Fig 3. Morphological Analysis](image)

The keywords are now extracted and tagged into their categories (Fig 4 & Fig.5).

![Fig 4. List of Morphemes](image)
The remedial instruction from the original document from the tag <மᾞᾸᾐ> is obtained (Fig. 6).

The information about the ingredients used and the remedy preparation instruction are fed to the template. The output is displayed in a structured format (Fig. 7).

C. Evaluation of Results and discussion

The arduous part in the given framework is tagging of the keywords into their respective categories. For experimental purpose, the whole process was run for 40 files, each containing a remedial instruction for a particular disorder. The f-score was found to be 0.782. The fall in the value of f-score is mainly due to the lower recall (0.72). The precision however was found to be quite high as manual category tagging in involved (0.858).

TABLE 1

<table>
<thead>
<tr>
<th>Incorrectly tagged words by the Morphological Analyser</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of word</td>
</tr>
<tr>
<td>Compound noun [noun + noun]</td>
</tr>
<tr>
<td>Partly tagged words</td>
</tr>
<tr>
<td>Noun tagged as blended word</td>
</tr>
</tbody>
</table>
The fall in recall value is due to various reasons. Table 1 shows the types of words that were tagged incorrectly by the Analyser. The main reason for fall in f-score is found to be the absence of the medicinal terms in the analyzer dictionary. For eg. பாைல (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. அதிமᾐரΆ (Liquorice) is split into two separate nouns, அதி and மᾐரΆ by the analyser. Hence a single ingredient is split into two, thereby preventing it from being tagged. These kinds of errors can be rectified by adding the unknown words and the compound nouns to the analyser dictionary.

Since the keyword extraction involved extracting only proper nouns, certain ingredient elements that were tagged as entity were not retrieved. Eg. ச῾ᾱகைர (sugar) was placed in the category of entities. Hence to overcome this error, words tagged as <Entity> are also considered as a keyword and are subjected to category tagging.

A very common difficulty that was observed in most of the text is two words denoting the same item. Eg. அ앇தி அ앇தி அ앇தி அ앇தி பழΆ பழΆ பழΆ பழΆ (Fig) has two words but denotes the same fruit. However when analysing, they are tagged as two separate nouns. Though அ앇தி denotes the fruit fig, பழΆ here is a common noun. Hence it gives an ambiguous output. Similar error can be found in கமலா கமலா கமலா கமalah and மேதᾱகாᾼ மேதᾱகாᾼ மேதᾱகாᾼ மேதᾱகாᾼ. This factor can be resolved by considering common nouns and the adjacent proper noun and tagging them as a single word. This however may not be successful in all cases.

Another difficulty that was observed was to identify the correct sense of certain words in the given context. For eg. நீ᾽ (Water) and ஌னᾼ (ghee) are proper nouns denoting elements that are used to prepare the remedy. But they are tagged as pronoun and verb respectively.

V. CONCLUSION AND FUTURE WORK

Thus the given system can extract traditional medicine information from tamil health tip documents and display the essential output in a structured format. Traditional medicine terms are extracted and tagged. The system may be extended to support disease based or ingredient based queries or both. To avoid the errors due to missing terms in the category tagged data, the NE dictionary can be built using machine-learning algorithms. Also to avoid the incorrect understanding of terms in the context, the Word Sense Disambiguation (WSD) can be addressed.
ACKNOWLEDGMENT

This project is funded by Centre for Technology Development and Transfer, Anna University, Chennai under Sanction No: 2984/CTDT-1/RSS SIP/CEG/2013 Batch-5 for “Student Innovative Project under Research Support Scheme”.

REFERENCES