Corpus & Wordnet Based MMD (Multilingual Manipuri Dictionary)

Y. Bablu Singh
Prof. B. S. Purkayashtha
Computer Science Department, Assam University, Silchar, Assam, India

Prof. Chungkham Yashawanta Singh
Nameirakpam Amit
Linguistics Department, Manipur University, Canchipur, Manipur, India

I. INTRODUCTION

In Automata theory Dictionary can be defined as a Nondeterministic Finite Automaton (NFA), or nondeterministic finite state machine, that is capable of transitioning to zero or two or more states for a given start state and input symbol. This distinguishes it from a Deterministic Finite Automaton (DFA), in which all transitions are uniquely determined and in which an input symbol is required for all state transitions (Martin, John, 2010). Like DFAs, NFAs only recognize regular languages. NFAs were introduced in 1959 by Michael O. Rabin and Dana Scott, [1]. An NFA, similar to a DFA, consumes a string of input symbols. For each input symbol, it transitions to a new state until all input symbols have been consumed. An NFA is represented formally by a 5-tuple, \((Q, \Sigma, \delta, q_0, F)\), consisting of

- a finite set of states \(Q\)
- a finite set of input symbols \(\Sigma\)
- a transition relation \(\delta : Q \times \Sigma \rightarrow P(Q)\).
- an initial (or start) state \(q_0 \in Q\)
- a set of states \(F\) distinguished as accepting (or final) states \(F \subseteq Q\).

Here, \(\Sigma\) denotes the power set of \(Q\). Let \(w = a_1a_2 \ldots a_n\) be a word over the alphabet \(\Sigma\). The automaton \(M\) accepts the word \(w\) if a sequence of states, \(r_0,r_1, \ldots, r_m\), exists in \(Q\) with the following conditions:

\[ r_0 = q_0 \]
\[ r_{i+1} \in \Delta(r_i, a_{i+1}), \text{ for } i = 0, ..., n-1 \]
\[ m \in F \]

Manipuri, a language of Tibeto-Burman (TB) sub-family, is locally known as Meiteiron (the Meetei + Lon ‘language’). It is basically spoken and understood by other communities of about 15.5 lakhs in the state according to 2001 census. There are native speaker also in the neighboring Northeastern states of India, in Assam mainly, Tripura, West Bengal (particularly...
in Brindavan, Radha Kunda). In Assam 3% of the total population speaks Manipuri; in Myanmar 1.5 lakhs speak Manipuri while 0.5 lakhs in Manipuri in Bangladesh according to 1961 census as reported by Manipuri Sahitya Parishad Imphal. The state has officially recognised 32 Scheduled tribes speaking various mother tongues whose second language is Manipur: they use it for their intergroup communication.

II. REVIEW OF LITERATURE


III. MULTILINGUAL MANIPURI DICTIONARY PROCESS

According to Atkins & Rundell [21] the process for building a dictionary is threefold. The three processes for building a dictionary is shown below, the first stage is Analysis stage second stage is Transfer stage and last stage is Synthesis stage.

![Figure 1: Three stage of MMD](image)

As we are developing the features of WordNet in MMD, which are different from other dictionary. WordNet is not merely an on-line thesaurus, however. In order to appreciate what more has been attempted in WordNet, it is necessary to understand its basic design [22]. WordNet is organized by semantic relations. Since a semantic relation is a relation between meanings, and since meanings can be represented by synsets, it is natural to think of semantic relations as pointers between synsets. It is characteristic of semantic relations that they are reciprocated: if there is a semantic relation R between meaning {x, x', . . . } and meaning {y, y', . . . }, then there is also a relation R' between {y, y', . . . } and {x, x', . . . }. For the purposes of the present discussion, the names of the semantic relations will serve a dual role: if the relation between the meanings {x, x', . . . } and {y, y', . . . } is called R, then R will also be used to designate the relation between individual word forms belonging to those synsets. It might be logically tidier to introduce separate terms for the relation between meanings and for the relation between forms, but even greater confusion might result from the introduction of so many new technical terms.

All three kinds of semantic relations- hyponymy, meronymy and antonymy are included; the result is highly interconnected network of nouns. A graphical representation is shown below.

![Figure 2: WordNet Semantic Relationship in English](image)
The four kind of entailment in WordNet is given below in the table format. We developed the same features in our languages too using the WordNet features.

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
<th>Adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonym !</td>
<td>Antonym !</td>
<td>Similar &amp; Relational</td>
<td>Derived from \</td>
</tr>
<tr>
<td>Synonym ~</td>
<td>Troponym</td>
<td>Adj \ Also see .</td>
<td></td>
</tr>
<tr>
<td>Meronym #</td>
<td>Entailment +</td>
<td>. Attribute</td>
<td></td>
</tr>
<tr>
<td>Holonym %</td>
<td>Cause &gt;</td>
<td>=</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
<th>Adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonym !</td>
<td>Antonym !</td>
<td>Similar &amp; Relational</td>
<td>Derived from \</td>
</tr>
<tr>
<td>Synonym ~</td>
<td>Troponym</td>
<td>Adj \ Also see .</td>
<td></td>
</tr>
<tr>
<td>Meronym #</td>
<td>Entailment +</td>
<td>. Attribute</td>
<td></td>
</tr>
<tr>
<td>Holonym %</td>
<td>Cause &gt;</td>
<td>=</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
<th>Adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonym !</td>
<td>Antonym !</td>
<td>Similar &amp; Relational</td>
<td>Derived from \</td>
</tr>
<tr>
<td>Synonym ~</td>
<td>Troponym</td>
<td>Adj \ Also see .</td>
<td></td>
</tr>
<tr>
<td>Meronym #</td>
<td>Entailment +</td>
<td>. Attribute</td>
<td></td>
</tr>
<tr>
<td>Holonym %</td>
<td>Cause &gt;</td>
<td>=</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Entailment in WordNet

After analyzing the features and entailment of WordNet here we are making a table of Antonymy in Manipuri Language with gloss in English.

<table>
<thead>
<tr>
<th>Size</th>
<th>Good-small, thick-thin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>Good-bad, love-hated</td>
</tr>
<tr>
<td>Gender</td>
<td>Son-daughter, fisher-mother</td>
</tr>
<tr>
<td>State</td>
<td>Beginning-end</td>
</tr>
<tr>
<td>Personality</td>
<td>Rama-Ravana, David-Celebrity</td>
</tr>
<tr>
<td>Direction</td>
<td>East-west, front-behind</td>
</tr>
<tr>
<td>Action</td>
<td>Give-take, buy-sell</td>
</tr>
<tr>
<td>Amount</td>
<td>Little-much, light-heavy</td>
</tr>
<tr>
<td>Place</td>
<td>Near-far, inside-outside</td>
</tr>
<tr>
<td>Time</td>
<td>Day-night, morning-evening</td>
</tr>
</tbody>
</table>

Table 3: Meronymy Relationship Table

A. GRADATION

Gradation is another semantic relation between adjectives that is considered by WordNet is gradation. For some attributes gradation can be expressed by ordered strings of adjectives going from a weak meaning to a strong one. An example of lexicalized gradation for the lightness attribute in Manipuri would be:

\[
\begin{align*}
A(1) & \quad (1/2) & \quad (0) & \quad (-1/2) & \quad B(-1) \\
| & akba & lakha akhando & kantombana & lakha bai & acerpa \\
& \text{Pre-Intermediate} & \text{Intermediate} & \text{Post-Intermediate} & \\
\end{align*}
\]

Figure 3: Gradation levels

B. MMD CORPUS

A parallel mini corpus have developed from many sources like books, journals, newspaper etc. by categorizing them in different domains name as per in [23] with proper characteristics which should have in a corpus. The MMD corpus used WordNet example and concept texts for developing the corpus. The cleaning of corpus has done by using a Perl programmed. The frequency of word And thus the sanitized corpus are used to stored using WX notation.

From WordNet to MMD Dictionary and MMD Corpus.

C. FREQUENCIES COUNT

Using the Perl program as we have given below we used to calculate the frequencies of words as a unique. And thus we takes a file with words and the corresponding frequencies # and outputs syllables with their frequencies.

```perl
while ($line = <>)
{    chop($line);
    ($frq, $_) = split (/\s+/s, $line);
    while ($_ ne '')  {
        # Do until the line ends.
        if (/^[aieou]*[aiueo]+$/i)  #space+Consonant(cluster)+vowel
            $syllable = $&; $_ = $';
        print $frq, ' ', $syllable, "n";
    } elsif(/^[aieou]*[aiueo]+$/i)
    {     # Conso cluster + vowel $syllable = $&; $_ = $';
        print $frq, ' ', $syllable, "n";
    } else { last;
        # print STDERR "Some non-
        print STDERR "Some non-syllable at the end: 
    } last;
    # Word frequency is counted and addition of new words is
    # process from MMD corpus, which are developed by concept,
    # gloss and synset text from WordNet.
```

Figure 4: Schematic diagram how WordNet data used for MMD Implementation.
IV. SYSTEM DESIGN AND IMPLEMENTATION

As MMD is developed using WordNet features and other data collected from hard copies of Manipuri available. Dictionary as we mentioned in the above literature review we have to enter all words manually as there is no electronic text written in Manipuri i.e. in Meitei Mayek. Figure 5 will represent how we have to process for building/ making database of MMD from different sources and figure 6 will represent the context diagram of web enabled Multilingual Manipuri Dictionary, and it is web enabled the context diagram of MMD is given below

![Diagram of MMD](image)

**Figure 5: Process diagram of MMD**

The context diagram of web enabled MMD is given below

![Context diagram](image)

**Figure 6: Context diagram for www.mmd.com**

A. TOOLS

As we are developing web based using VB.net as frontend, SQL 2000 Server as backend using ASP.net technology as the ASP.net have the advantages features as [24] [25], develops MMD, IIS as web server [26].

B. TRIE STRUCTURE

Structure of storing lexicon in a database can be in trie structure. The trie of a dictionary can be defined as follows:

- The root node is null.
- The first letter of any word appears at the level below the root.
- Each node on level ‘1’ stores a character, which is common to a set of stems that begin with a certain sequence of 1 character. The node specifies an M-way (M=<40) branch depending on the (i+1)th characters.
- Beside the root nodes, there are three other types of nodes depending on the type of pointers associated with the node in addition to the character.
  - Leaf node – contains a pointer to the grammatical information file.
  - Internal node – contains a pointer to the text sibling as the same level.
  - Both internal and leaf node – contains both the above information.

At any level if there is no branching from a node then the remaining letters of the stem are stored consecutively at the node followed by a pointer to the grammatical information file.

![Trie Structure](image)

**Figure 7: Trie Structure of lexicon storing in Dictionary**

A------C—H—A7—ABAA*1 OUBAA*2
K—A7—ABAA*3 HAABAA*4
N—G—A9—ANGBA*5 OUBAA*6
P—18—IKPAA*7 OUBAA*8
T----H—A7—ABAA*9 IIBAA*10

<table>
<thead>
<tr>
<th>Table 4: Example trie in Manipuri</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some pointer address spaces do not get filled up (denoted by ‘-‘) since there is no sibling character at the same level for those characters in the example trie.</td>
</tr>
</tbody>
</table>

C. COMPLEXITY ANALYSIS

Now that we’ve seen the basic operations on how to work with a TRIE, we shall now see the space and time complexities. To measure the complexity we considered two important operations INSERT and SEARCH. For every Node in the TRIE we had something called as Collection where the Collection can be either a Set or a List. If we choose Set, the order of whatever operation we perform over that will be in O (1) time, whereas if we use a Linked List the number of comparisons at worst will be 40 (the total number of alphabets in Manipuri Language). So for moving from one node to another, there will be at least 40 comparisons will be required at each step.

D. MACROSTRUCTURE AND MICROSTRUCTURE OF MMD

A macrostructure denotes the ordering of the dictionary articles. The index pages in MMD contain ordered lists of headwords linking to their corresponding article pages. Most
printed dictionaries are ordered alphabetically. Macrostructure refers to the organization of the dictionary. The macrostructure can be divided into simple and complex, the former referring to the "lexicographic macrostructure which applies to only two macrostructural components"3 and the latter referring to three or more elements of the macrostructure, which every dictionary should ideally have. The macrostructural components of a dictionary may include:

- Table of contents.
- Preface.
- Users guide.
- List of abbreviations.
- Word list.
- Head word list.
- Appendix.
- List of prefix/suffix.

The microstructure in dictionaries, the amount, content and structure of information following the dictionary entry are of interest. Microstructure may include various elements:

- Headword information.
- Part of speech.
- Pronunciation
- Phonetic transcription like “box” of boat
- Tonal mark(for target language)
- Grammatical information.
- Concept.
- Gloss.
- Synset
- Antonyms, Hyponyms, Hyponyma.
- Morphological information.
- Semantic Information.
- Cross lingual information.
- Domain.
- Source.
- Semantic information.
- References.

E. RESULT

Output page for searching by English is shown given below here we are showing only one page of output page we can search in both tri lingual. The module is categorized in two section one for administrative and another for user pages.

![Figure 8: Output page](image)

Total word count in English, Hindi and Manipuri is given below in the table form.

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>POS</th>
<th>English</th>
<th>Hindi</th>
<th>No. of Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Noun</td>
<td>57000</td>
<td>145000</td>
<td>87000</td>
</tr>
<tr>
<td>2</td>
<td>Pronouns</td>
<td>39</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>Verb</td>
<td>21000</td>
<td>25884</td>
<td>31,000 +</td>
</tr>
<tr>
<td>4</td>
<td>Adjectives</td>
<td>834</td>
<td>834</td>
<td>19500</td>
</tr>
<tr>
<td>5</td>
<td>Adverbs</td>
<td>19500</td>
<td>31302</td>
<td>843</td>
</tr>
<tr>
<td>6</td>
<td>Prepositions/Postposition</td>
<td>189</td>
<td>189</td>
<td>189 (English)</td>
</tr>
<tr>
<td>7</td>
<td>Interjections</td>
<td>8</td>
<td>17</td>
<td>8 (English)</td>
</tr>
<tr>
<td>8</td>
<td>Conjunction</td>
<td>199</td>
<td>17</td>
<td>17 (Hindi)</td>
</tr>
</tbody>
</table>

Table 5: Word count of MMD

F. ADVANTAGES OF MULTILINGUAL DICTIONARIES

- MMD is dictionary independent and language dependent.
- It has intelligent access to various sorts of dictionaries bilingual corpora, monolingual and multilingual corpora.
- Simultaneously an unlimited number of dictionaries can be held open, thus by a single interrogation step, all the dictionaries (with translations, explanations, synonyms, etc.) can be surveyed.
- The translators own glossaries built with the help of the system may also be disseminated among other users, if needed.
- It has an open architecture and a well-defined API.
- It has been implemented and is available with a gradually increasing number of dictionaries for numerous language pairs.
- It has reputable Internet equipment including on the Internet Multilingual dictionaries, systems for grammar examining, as well as other processing equipment.
- The uniqueness of language printed dictionaries, as a result the need for on the Internet.
- Online English dictionaries are auto-updated; in actual fact here is the primary benefit of on the web dictionaries above printed kinds.
- An on-line Multilingual dictionary is actually a great tool of data, especially if we know how you can distinguish a fantastic on line dictionary from the bad a single.
- This may be an essential resource for understanding an international language; here I used English source language so one can translate from this source language easily.
- One among the very best traits of these Multilingual Manipuri dictionaries would be the translation feature, e.g. an online dictionary can translate an English word into Hindi, Manipuri and vice versa.
- Dependability in the resource.
- It means that every new phrase or terminology might be extra during the databases of on the net dictionary.
- Development of inter-linguistic indices for multilingual conceptual equivalence, with Machine Translation.
- New ideal tool to optimize the retrieval capacity of existing systems: natural language interfaces for search.
engines; automatic generation of tools for semantic disambiguation of concepts (corpora, dictionaries, directories, thesauri) and the creation of knowledge summaries from expanded queries.

- A design, which support of grammatical categorizations designed to classify information by aspects and traits, but in particular to design and classify semantic ontologies that organize web data.
- Language teaching and translation application.
- Domain classifications of each word.

V. FUTURE WORK AND CONCLUSION

WordNet other features and more advantages can be add as a future work, and for corpus one can start a corpus based on voice also, which may work very advantageous methods in future. In researching of words from any languages one can add picture as search result but so far my knowledge we can add Noun as the best, but it will be a tremendous work for dictionary for kind of profession. As a future work for the entries of data in dictionary one can extend the number of entries as the listed below:

A. ENTRY LIST

Word entry for one dictionary might be the following categories for more information of a dictionary to help user for more information like as follows: Lemma, Homonym number, Sense form, Citation form, Phonological representation, Sound, Phonological variants, Orthographic variants, Dialects, Sociolect, Style, Stage, Proper name, Syntactic category, Morphological structure, Word formation, Derivation, Morphological categories, Irregular inflection, Construction, Phraseology, Meaning definition, Gloss, Semantic class, Semantic relations, Encyclopedic info, Picture, Origin, Entymology, Cognates, Comments, Problems, Date.

ACKNOWLEDGMENT

We would like to thanks Consortium leader of Indian Language Corpora Initiative (ILCI), JNU, Dr. Girish Nath Jha, IndoWordNet, Consortium Leader, Prof. Pushpak Bhattacharya, IIT Mumbai and IndowordNet P.I. of Manipuri Language.

REFERENCES