Chapter 7 : EVALUATION AND ANALYSIS

This chapter is a key to the answers of second and thid research objectives –

Efficiency & Sufficiency of the technique, and
Coverage of various word forms of different word classes of Manipuri.

To find an appropriate answer to these questions, we carried out a small evaluation of our xfst language model.

7.1 EVALUATION

To evaluate the system model for morphological analysis we have collected 1000 root tokens of animate, inanimate, personal pronouns and kinship terms for nominal category and for verbal category from different sources as mentioned in the following:

1. Newspaper: http://www.hueiyenlanpao.com/ dated 21/4/2014
2. Literature items: Leipərең (poem book by Khwairakpam Chaouba)
3. A collection of children vocabulary
4. Story book: Madhabi (A Novel by Dr. Kamal)

To design a language model for morphological analysis of Manipuri, our different source lexica has been populated with over 300 tokens of nominal roots of animate, inanimate, personal pronouns and kinship terms and with over 130 tokens of process, action and stative verbs. Also we have an adjectives lexicon derived from the verbal roots. The lexicon entries are sampled in such a way that they represent maximum variation in terms of morphonemic, orthographic and spelling change while combining with different affixes. The break-up of the statistics of the lexical entry for each lexicon is in the following table:

Table 7-1: Statistics of the lexical entry and total word forms per entry

Root type	No. of lexical entry	No. of word forms generated per entry	Total word forms	Wrong analysis per entry (Average)
Nominal Category				
Animate	50	35	1750	7
Inanimate	65	35	2275	8
Personal Pronouns	10	30	300	5
Kinship terms	45	33	1485	7
Wh-words	25	31	775	9
Verbal Category				
Process	30	54	1620	12
Action	35	54	1890	11
Adjectives				
Action	30	44	1320	7
Stative	24	44	1056	8

The percentage analysis as against the number of entries in the selected nominal source lexica are shown in the following table:

Table 7-2: Analysis of nominal category word forms

Root type	No. of word forms per lexical entry	Wrong analysis	% of Wrong analysis	% of correct analysis
Animate	35	7	20.00%	80%
Inanimate	35	8	22.86%	77.14%
Personal Pronouns	30	5	16.67%	83.33%
Kinship terms	33	7	21.21%	78.79%
Wh-words	31	9	29.03%	70.97%

The percentage analysis as against the number of entries in the selected verbal and adjective source lexica are shown in the following table:

Table 7-3: Analysis of verbal category word forms

Root type	No. of word forms per lexical entry	Wrong analysis	% of Wrong analysis	% of correct analysis
Verbs				
Process	54	12	27.78%	77.78%
Action	54	11	20.37%	79.63%
Adjectives				
Action	44	7	15.91%	84.09%
Stative	44	8	18.18%	81.82%

An analysis of the results produced by the xfst analyzer has shown that the model mainly failed on two accounts- spelling mismatch in word forms (probably because of lack of a standard grammar on orthographic rules) and missing spelling rules in words. For some portion of the lexicon, a few more combinations of morphemes, i.e. morphotactics are seen to be missing in the word forms.

As we have not considered the case of reduplication and compounding, our model donot have combination of suffixes with roots e.g. *təpnə-təpnə* (slowly slowly, an adverbial complete reduplication) and compounding e.g. *mitinfəm (meeting place).*

So, for the existing language model, the potential for improvement exists especially in the area of:

- Size of the source lexicon- an increase in the number of lexical entry.
- Morphotactics- more morpheme combination and rechecking of morpheme ordering in the verbal word forms.
- Spelling rules- some percentage of the wrong analysis is due to the lack of rules for morphophonemic alternations, addition of more such rules is called for.
- Refinement of and ordering of morphophonemic alternation rule- Our lexicon has some conjunct characters while using Bengali script. These characters when analysed

by xfst analyzer, the component characters are split into their individual forms, so mapping from the lower to the upper symbol is from individual to individual symbol and it gives rise to a wrong analysis. It can either be solved by declaring them as a multicharacter symbol or by defining new spelling change rules.

Also due to wrong ordering of the spelling rules some word forms are analyzed wrongly as in case of personal pronouns and kinship terms when attaching postpositions and particles. So a proper ordering of the rules is required here.

The kinds of irregularities due to homonemy and ideosyncracy for some categories of word forms can be handled by making an entry to respective source lexicon. As in the following example for homonemy

তৌ/tou(do, bound verb root) and

noun তৌ/tou(a kind of grass, free form noun)

The noun should be entered in the inanimate category lexicon as it follows the inanimate morphotactics whereas the other should be entered in the action verb lexicon so that it follows the morphotactics of their respective categories.

Manipuri personal pronouns have free and extended form of first, second and third person forms. They behave differently when inflected with case markers. The adjective category has some irregular forms for expressing colour, etc. e.g. হীগোক/higok (blue), লৈঙাং/leiŋaŋ (saffron), etc. All these can be handled using a different lexicon for such irregularities.

7.2 EFFICIENCY AND SUFFICIENCY

The results of the morphological analysis of Manipuri word forms done, with the help of finite state techniques using the xfst tool proved to be very efficient considering the fact that deterministic finite state systems work linear in the length of the string. xfst tool is capable of handling and analyzing very long input strings which is normal in agglutinative languages, within a matter of seconds. Finite state algorithms look up

routine took less than a second to analyse over 1000 words which is a very impressive rate for a complex natural language processing task like morphological analysis. An interesting point to be noted here is that using finite state technique in modeling morphological analysis is simple as it allows for a straightforward implementation once the linguistic facts are established.

7.3 COVERAGE

Though our source lexicon is not of an impressive size, our sample of tokens for evaluation of the model indeed is good enough to test the model for coverage. Also as per our small evaluation conducted in the last section, and looking at the reasons for failure on some accounts, we can say that finite state techniques are ideally suited to cover a wide range of possible word forms in a morphologically rich and complex agglutinative language like Manipuri.