DEVELOPING MORPHOLOGICAL ANALYZERS FOR FOUR INDIAN LANGUAGES USING A RULE BASED AFFIX STRIPPING APPROACH

Mona Parakh
Reader/Research Officer
ldc-monaparakh@ciil.stpmy.soft.net

Rajesha N
Senior Technical Officer,
ldc-rajesha@ciil.stpy.soft.net

Linguistic Data Consortium for Indian Languages, CIIL, Mysore

Abstract **- The present paper deals with the design and development of morphological analyzers for four Indian languages, viz., Assamese, Bengali, Bodo and Oriya. These analyzers are being developed using the Suffix Stripping Approach.**

The results of the first version of the analyzers using this approach are fairly encouraging. The coverage of the system is directly related to the size of the dictionary. As this is an ongoing work, we hope to expand and make the system more robust, by increasing the dictionary size.

I. INTRODUCTION

Considering the extensive work that is being carried out in the area of Indian Language Technologies, towards building Language Applications for Major Indian Languages it is the need of the hour to develop and generate language resources for a large number of Indian languages, which are of high quality and with distinct standards.

In order to fulfill this long-pending need, the Central Institute of Indian Languages, Mysore and several other institutions working on Indian Languages technology have set up the Linguistic Data Consortium for Indian Languages (LDC-IL), whose main goal is to create and manage large Indian languages databases. One of the many resource building activities that LDC-IL is involved in includes developing Morphological Analyzers and Generators for major Indian languages.

There are two approaches used to build the Morphological Analyzers at LDC-IL, viz., the Word and Paradigm Approach [1] and the Rule Based Affix Stripping Approach. Morphological Analyzers for ten of the thirteen Indian languages mentioned above are being developed using the Apertium – Lttoolbox [2]. and [5]. For four of the languages, viz., Assamese, Bengali, Bodo and Oriya, analyzers are being developed using the suffix stripping approach. Some other research groups have developed analyzers using the Apertium-Lttoolbox for languages like Marathi [6], Telugu and Tamil [3].

The present paper reports the ongoing work of building Morphological Analyzers using the Suffix Stripping method for the four languages – Assamese, Bengali, Bodo and Oriya. Currently the system only handles inflectional suffixes though it will be further modified so as to handle derivation as well as prefixation, in each of these languages. The system is at different stages of completion depending on the availability of the language resources and human resources for the respective languages.

II. RULE BASED SUFFIX STRIPPING APPROACH.

The Word and Paradigm Model (WPM) is unsuitable and inadequate to capture all morphological functions in case of Assamese, Bengali, Bodo and Oriya. The reason for this is that these languages are classifier based languages. Even though the classifiers are finite in number, they can occur in various combinations with nouns. This would increase the manual effort of paradigm creation immensely. Moreover, in these languages morpho-phonemics does not play much of a role. Hence, the Suffix Stripping Approach has been found to be suitable.

As the name suggests, this method involves identifying individual suffixes from a series of suffixes attached to a stem/root, using morpheme sequencing rules. This approach is highly efficient in case of agglutinative languages. However, in languages that display tendency for morpho-phonemic changes during affixation (such as Dravidian languages), this method will require an additional component of morpho-phonemic rules besides the morpheme sequencing rules.

A. ORGANIZATION OF DATA.

The analyzer based on this approach is so modeled that it analyses the inflected form of a word into suffixes and stems. It does so by making use of a root/stem dictionary (for identifying legitimate roots/stems), a list of suffixes, comprising of all possible suffixes that various categories can take (in order to identify a valid suffix), and the morpheme sequencing rules.

The Root Dictionary contains a list of roots, each with its lexical category and features. Following are samples of words from the Assamese, Bengali and Oriya root dictionaries:

1. Assamese

(a) মাহীদেউগৰাকী\NN.sg.fem ‘maternal aunt

(b) পাগলী\ADJ.fem ‘crazy’

(c) কৰ\VM ‘to do’

2. Bengali

(a) স্টার্ট\NN.0 'start'

(b) ওল্ড\ADJ.0 'old'

(c) বলা\VM 'to say'

3. Oriya

(a) ଗଛ/NN.0.0 'tree'

(b) ଗାଢ଼\ADJ.0 'bold'

(c) ଯା\VM 'go'

The Suffix List contains a list of suffixes with their morpho-syntactic feature values like gender, number, person and other relevant morphological information stored in the form of a dual field list. It deals only with inflectional suffixes not derivational. Following are samples of the Assamese, Bengali, Bodo and Oriya suffix lists.

TABLE 1: SAMPLE OF ASSAMESE SUFFIX LIST

Affix	Feature	Expansion of Abbreviations
ত	CM.Loc	Case marker, Locative
ও	Prt	Particle
টো	Cl	Classifier

TABLE 2: SAMPLE OF BENGALI SUFFIX LIST

Affix	Feature	Expansion of Abbreviations
য়	CM.loc	Case marker, Locative
টা	Prt.Def	Particle, Definite
য়েরা	Pl	Plural suffix

TABLE 3: SAMPLE OF BODO SUFFIX LIST

Affix	Feature	Expansion of Abbreviations
आव	CM.loc	Case marker, Locative
नो	Prt.emph	Particle, Emphatic
दों	Asp.prg	Aspect: Progressive

TABLE 4: SAMPLE OF ORIYA SUFFIX LIST

Affix	Feature	Expansion of Abbreviations
ରେ	CM.loc	Case marker, Locative
ଯାକ	pl	Plural suffix
ଟାଏ	Prt.def.sg	Particle- definite, singular

The Rule List provides all the possible morpheme sequences for a given category, i.e., for each category, it provides the rules identifying the ordering of suffixes.

TABLE 5: SAMPLE OF MORPHEME SEQUENCING RULES

Rules	**Expansion of abbreviations**
NN+pl+CM.ins	Noun+plural+Case marker: Instrumental
CRD+PART.emp	Cardinal+Particle: Emphatic
ORD+PART.def.sg	Ordinal+Particle: Definite, Singular
PRP+CM.gen+CM.loc	Pronoun+Case marker: genitive+ Case marker: Locative
ADJ+CM.acc	Adjective+Case marker: Accusative
VM+neg+aux.pst+sg	Verb Main +Negative+Auxiliary: Past Tense, Singular

B. THE METHOD.

Following is a Flow Chart diagram of the Morphological Analyser.

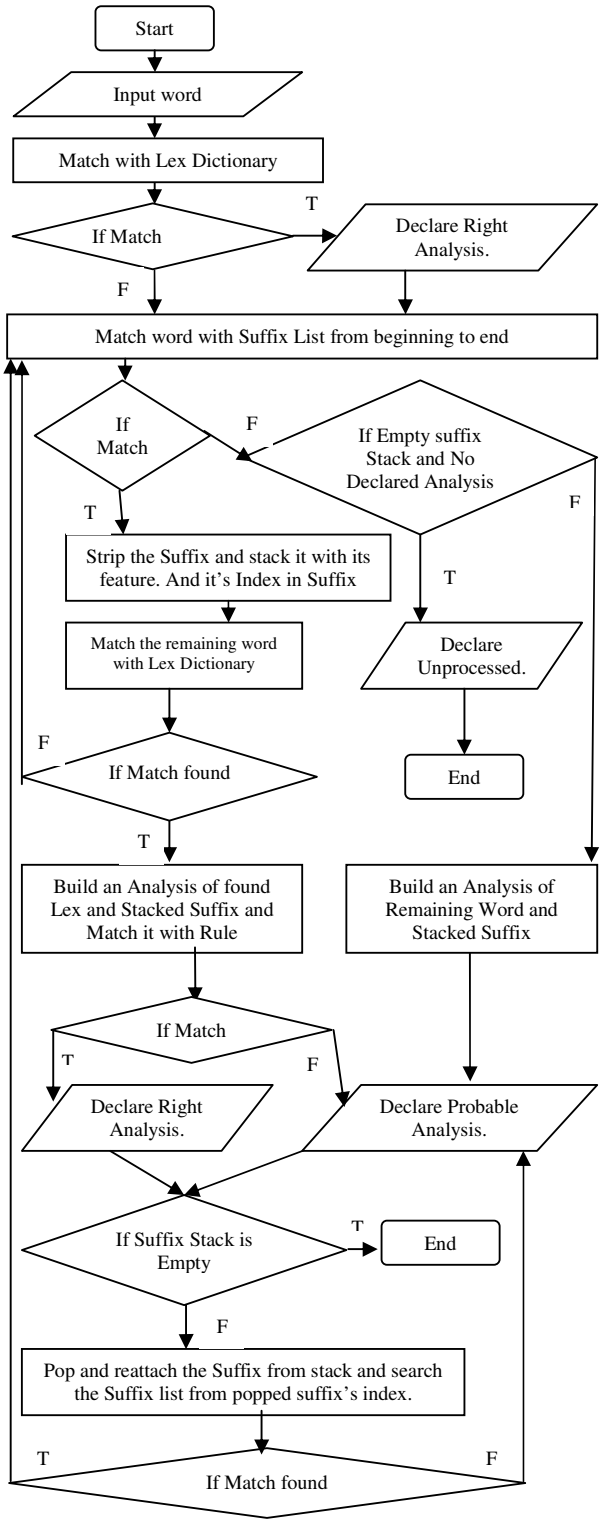

FIGURE 1: FLOW CHART DIAGRAM FOR MORPHOLOGICAL ANALYSER

The suffix stripping algorithm is a method of morphological analysis which makes use of a root/stem dictionary (for identifying legitimate roots/stems), a list of suffixes, comprising of all possible suffixes that various categories can take, and the morpheme sequencing rules. This method is economical. Once the suffixes are identified, removing the suffixes and applying proper morpheme sequencing rules can obtain the stem.

In order to identify the legitimate roots/stems, the dictionary of root/stem needs to be as exhaustive as possible. Considering this fact, the analyzer is designed to provide three types of outputs such as:

The Correct analysis: This is obtained on the basis of a complete match of suffixes, rules and the existence of the analyzed stem/root in the root dictionary.

Probable analysis: This is obtained on the basis of either a matching of the suffixes and rules, even if the root/stem is not found in the dictionary or a matching of the suffixes, but not any supporting rule or existing root in the dictionary.

Unprocessed words: These are the words which have remained unanalyzed due to either absence of the suffix in the suffix list or due to the absence of the rule in the list.

C. *Increasing the coverage (Phase 1).*

In order to increase the coverage of the system the root dictionary had to be made robust. To this end, a module has been introduced in the system, so that the roots of the probable analyses can be manually added to the root dictionary after validating them and automatically checking whether they already exist in the dictionary or not. Also, the list of unprocessed words, are manually checked and validated, after which they are added to the dictionary, with their corresponding feature values. In phase 1, this process was repeated over larger and random test corpora and with every repetition the dictionary size increased, thereby resulting in the increase in the number of correct analyses.

D. *Towards increasing the coverage (Phase 2).*

In the second phase a method has been devised to ensure that the coverage of the root/stem dictionary increases faster. Hence, the test data has been replaced by a frequency wise word list (FWL) generated from the entire available corpus of a given language. The FWL has been run on the system in blocks of 10,000 each, starting with the most frequent words to the less frequent ones in the descending order. The words which remain unanalyzed or fall under the probable analysis are first entered in the root/stem dictionary before the next block of 10,000 words are given to the system.

The logic here is simply that by first adding the most frequently occurring words in a language the overall coverage of the system shoots up manifold as compared to when entering words randomly from a corpus.

E: *Suffix and Dictionary coverage for Indian Languages.*

Details of the system coverage and the coverage of the rules and the root/stem dictionary for each of the above Languages are given below in table 6.

TABLE 6: LANGUAGE WISE COVERAGE OF THE SYSTEM

Language	Lex Dictionary Entries	Suffix-Feature pair	Rules	Coverage
Assamese	15452	216	1040	56.338 %
Bengali	12867	187	227	48.326 %
Bodo	16784	131	4379	65.82 %
Oriya	22532	127	536	70.39

CONCLUSION

The paper is about the design and development of morphological analyzers for four Indian Languages, using the suffix stripping approach. The results of the first phase of the suffix stripping approach have been fairly encouraging. It was observed, that with an average of 7000 to 8000 root entries, the affix stripping approach gives around 50% coverage. As is evident from the table 6, the coverage of the system is directly related to the size of the dictionary. We hope to expand and make the system more robust by increasing the dictionary size.

ACKNOWLEDGEMENT

We wish to thank the LDC-IL team for their support and help; but our special thanks are due to Ms. Ashmrita Gogoi,Mr. Farson Dalmari , Mr. Pramod Kumar Rout and Mr. Sankarsan Dutta for rigorously taking up the task of resource building for Assamese, Bodo, Oriya and Bengali, respectively.

REFERENCES

[1]. Bharti, A., V. Chatanya, and R. Sangal. *Natural Language Processing: A Paninian Perspective*. New Delhi: Prentice Hall. 1995.

[2]. M. L. Forcada, B. Bonev, Ortiz S. Rojas, et. al., "Documentation of the Open-Source Shallow-Transfer Machine Translation platform Apertium". 2007. Available online at: http://xixona.dlsi.ua.es/~fran/apertium2documentation.pdf.

[3]. Parameswari K. "An improvised Morphological Analyzer for Tamil: A case of implementing the open source platform Apertium". Unpublished M.Phil. Thesis. Hyderabad: University of Hyderabad. 2009.

[4]. S. Mohanty, P.K.Santi ,K.P.Das Adhikary. "Analysis and Design of Oriya Morphological Analyser: Some Tests with OriNet". *Proceeding of symposium on Indian Morphology,phonology and Language Engineering*,IIT Kharagpur. 2004.

[5]. Tyers, F. M. and Sánchez-Martínez, F. and Ortiz-Rojas, S. and Forcada, M. L. "Free/open-source resources in the Apertium platform for machine translation research and development". *The Prague Bulletin of Mathematical Linguistics. Vol.* 93. pp 67—76, 2010.

[6]. Vaidhya, Ashwini and Dipti Misra Sharma. "Using Paradigms for Certain Morphological phenomena in Marathi". *7th International Conference on NLP (ICON-2009).* New Delhi: Macmilan Publishers India Ltd., December 2009.

This page is left blank deliberately because of formatting problem.