A Survey on Odia Computational Morphology

Dhabal Prasad Sethi

ABSTRACT

In natural language processing morphological analyzer, morphological generator and morphological parser is the essential common tool used in developing language software. Morphological analyzer is a computer program which provides the grammatical information of a word. Morphological generator, by giving/providing the root word and grammatical information, it generates all word forms of that word. On the other hand morphological parsing is the process of determining the morphemes from given word? In this article I surveys the different work have done about Odia morphology.

Keyword: Morphology, Morphological Analyzer, Morphological Generator, Morphological Parser, Information Retrieval, Stemmer, IE, Spell Checker, Machine Translation, Question Answering System.

I. INTRODUCTION

Language as a means of human communication is a tool to express the human ideas and emotions. Language has structure and meaning called as syntax and semantics. Instead of human can analyze the language structure and meaning, we should develop such an automated tool for language that will analyze like human and give related information of a word. Before developing a tool every research goes to analyze the grammatical information of a word first then implement it using different methods.

Morphological analysis is essential for morphological rich language like Odia. Morphology is the study of how word forms are built and their inner structure. Examples of morphological tool are morphological generator, morphological analyzer, morphological parser and other. A morphological analyzer is used to analyze the internal structure of the words of a language. Morphological generator is the reverse of analyzer e.g. by giving a root word and grammatical information morphological, generator will generate the particular word form of that word. Morphological parsing or syntactic parsing is the process of analyzing and determining the structure of a text which is made up of sequence of tokens with respect to a given formal grammar. Odia language has inflectional, derivational, compound forms of word .To know/analyze these words a person should have strong linguistic knowledge. Morphological analyzer and generator tool is the first attempt which is commonly used in complex language software like machine translation, pos tagging etc. Morphology is generally two types: inflectional and derivational. Inflectional morphology is the process by which various inflectional forms are formed from a lexical stem. Derivational morphology on the other hand is the process of new lexemes is formed from existing ones by adding affixes to it.

Author Name: *Dhabal Prasad Sethi, lecturer in Computer Science & Engineering, Government College of Engineering,Keonjhar,Odisha,India.*

The rest part is organized as section II describes the literature survey, section III describes the different methods of morphological analyzer and generator, section IV describes the applications and section V describes the conclusion.

II.LITERATURE SURVEY

Itisree Jena, Sriram Chaudhury, Himani Chaudhry, Dipti M. Sharma [3] presented a paper named "Developing Oriya Morphological Analyzer Using Lt-toolbox ".They have developed the morphological analyzer for Odia using paradigm approach. The paradigm approach is a method which defines the all the word form of a given stem and its associated feature structure. Their system handles the inflectional morphology of noun, verb and adjectives.

R.C Balabantray, M.K.Jena S.Mohanty[4] presented a paper named "Shallow Morphology based complex predicates extraction in Oriya" The aim of their paper is to extracting the complex predicates (CPs) for the sentences containing the lexicon pattern {[MMM](n/adj)[NNN](v)} in the shallow parsed sentence where MMM and NNN represent any word. The lexical category of the root word of MMM is either noun (n) or adjective (adj) and the lexical category of the root word is verb (v).

Sanghamitra Mohanty, Prabhat Kumar Santi, K.P.Das Adikary[2] presented a paper named "Analysis and Design of Oriya Morphological Analyzer: Some Tests with OriNet". In this article they have mentioned that there are three types of morphology e.g. pronoun morphology, inflectional morphology and derivational morphology. Then they design and develop the software for odia morphology. They design the architecture of Odia Morphological Analyzer(OMA) which consist of five parts e.g. OriNet data base(OD) which stores the odia lexicon(root words),OMA Engine(OE) which processes the system, Morphological Parser(MP) which parses the word according to orthographic rule, Decision Tree(DT) which decides to classify the morphemes. They also mentions their system is designed on the basis of object oriented approach (OOA).The application of odia morphological analyzer is odia spell checker, odia grammar checker, odia machine translation, word net for odia. They also mention their system development is based on the syntactic approach of Sanskrit language.

Kalyani R. Shabadi [1] presented a paper named "Finite State Morphological Processing of Oriya Verbal Forms". In this article she discusses the morphological processing of verbal forms in Oriya in a deterministic finite state automation. Their work proposes a model for designing a morphological analyzer for odia verbal forms which can provides lexical, morphological and syntactic information for each lexical unit in the analyzed verbal forms.

Rakesh Chandra Balabantray, Sanjaya Kumar Lenka[8] presented a paper named "Computational Model for Reduplication in Odia". In this paper they examine the internal structure if reduplication in odia and possible generation of reduplication word from finite number of lexical items. In this paper a new word if formulated by copying of either a part or whole of the root word.

I (Dhabal Prasad Sethi) [5] have (has) presented the paper named Morphological Analyzer for Sambalpuri Odia Dialect Inflected Verbal Forms. Here first time I have presented the morphological analyzer of dialectal language in India. Since dialectal language has separate grammar, syntax in comparison to standard language, it should be developed morphological analyzer of any Indian language. I have used the suffix stripping algorithm to develop the tool.

III.DIFFERENT METHODS OF MORPHOLOGICAL ANALYZER AND GENERATOR

1) Suffix Stripping Algorithm:

The suffix stripping algorithm is an approach used in morphological analysis which requires a root/stem dictionary, a list of suffixes, by comprising of all possible suffixes with that various categories can take, and the morpheme sequencing rules. Once the suffixes are identified, removing the suffixes and applying proper morpheme sequencing rules can obtain the stem.

2) Finite State Automata Based Approach

It is common that most of morphological related tool can be described with regular expressions the use of finite-state techniques. When morphtactics is seen as a simple concatenation of morphs, it can straightforwardly be described by finite automata. The finite state automata consist of states and arcs called transition. It has one initial state and one or more finial states. In between initial and finial states there can be any number of finite states called intermediate states. Transitions are the connection between the states and are moving from one state to another. States are represented as circles and transition between the states are represented as labeled arcs. An arrow is used to indicate the initial states and double circles are used to indicate the finial states. The finial states automata is the best understood as recognizers because they accept a finite set of input string.

A finite state automation that accepts house and houses is shown above. Formally the finite state automation can be defined by the following five parameters:

Q: a finite set of N states q_0,q_1,q_3,q_4,Q_N.

$\sum$: a finite input alphabet of symbols

Q_0: the start state

F: the set of final states $F \leq Q$

Delta (q, i): the transitions function or transition matrix between states

Given a state q€Q. delta is thus a relation from $Q \times \sum$ to Q

3) Finite state Transducer (FST)

A finite state transducer is similar to finite state automata. It consist of states and transitions with labeled arcs .In FST the label can be in a pair of symbols .e.g the relation between two languages instead of simple symbols. When an arc has such a label, it is traversed and the input symbol matches then it transuded to the output symbol. Example: in the upper side is labeled as house+Noun+sg and the lower side are labeled as house. This FST transducers house+noun+sg to house and vice versa. That means input "house"is matched and its outputs house+noun+sg.

4) Two-Level Morphology Based Approach

In 1983 Kimmo Koskenniemi, a Finnish computer scientist first uses this approach for development of computational model of word-form recognition and generation. This approach consists of two levels e.g. surface level and other is lexical level. A word which is in written form or spoken text represents the outer form called surface level or surface form. Or Surface level represents the actual spelling of the word. The lexical level on the other hand displays various kind of information e.g. canonical form or lemma form of word and a set of tags showing its syntactic category and morphological features. Thus the lexical level represents a simple concatenation of morphemes making up a word. The actual arrangements of morpheme are governed by language specific rules. In Odia the surface level word is: BALAKAMANE and the lexical level representation are: BALAKA+NOUN+PL. So two levels is the mapping between surface levels to lexical level.

5) Corpus Based Approach:

Corpus based approach is the statistical based .In this approach, a large sized corpus is needed for training. The machine learning algorithm is used to train the corpus and collects the statistical information and other features from the corpus. The performance of the system depends on the features and size of the corpus. The disadvantage is that corpus creation is time consuming process. This approach is well suited for language software development.

6) Paradigm Based Approach

Paradigm means a set or list of all the inflected forms a word/lexeme or of one of its grammatical category. Example of paradigms is the conjugations of verbs and the declension of noun. Accordingly the word forms of a lexeme may be arranged into tables by classifying them according to inflected categories such as tense, aspect, mood, number, gender or case.

IV.APPLICATIONS OF MORPHOLOGY

Machine Translation, Question Answering System, Information Extraction, Information Retrieval, Spell Checker, Lexicography, Dictionary, Text to Speech System, Pos Tagging, Speech Recognition.

Machine Translation: It is a branch of computational linguistic which concerns with building application software which will translate text or speech from one language to another language**. Question Answering System:** It is a computer program which will automatically answers the question like human produced natural language. **Information Extraction:** It is a type of information retrieval whose goal is to automatically extract structured information from unstructured documents. Or it refers to the machines ability to automatically extract structured information. Unstructured information refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Data mining technique are applied in unstructured data. Structured information involves the manual tagging with metadata or part-of-speech tagging for text-mining-based structuring. **Information Retrieval:** It is the interaction between human and computer that happens when we use a machine to search relevant information from large content information that match our search query**. Spell Checker:** It is a tool that will check the spelling of words in a document, validate them and in case the checker finds some error then list out the correct form of that word. The spell checker detects the mistakes and prompts the user with a set of suggestions, which will aid the correction of the misspelled word. Spell checking deals with detection and automatic correction of spelling errors in an electronic document. **Lexicography:** It is the discipline of analyzing and describing the semantics, syntagmatic and paradigmatic relationship with the lexicon (vocabulary) of a language**. Dictionary:** It is a collection of words in one or more specific languages, listed

alphabetically with usage information, definitions, phonetics, pronunciation and other information. **Text to Speech System:** A text-to speech system takes text as input and produce speech form from it. Morphological analysis helps to solve two different tasks in such system. One is to guide the grapheme-to-phoneme conversion. Characters are generally ambiguous with respect to their translation into phonemes. Finding the morphological structure is necessary for solving the task correctly. The sequence *th, is* pronounced as */D/ or/T/*in English .The word <hothouse> we need to know the morpho structure <hot+house> to correctly pronounce the *th* sequence as */th/*. **Pos Tagging:** POS tagging is an important field of natural language processing. Part of speech tagging means assigning grammatical classes e.g. appropriate part of speech tags to each word in a natural language sentence (noun, pronoun, verb etc). **Speech Recognition:** Speech Recognition (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or"speech to text" (STT).Speech recognition is a field where morphological analysis plays an important role. At the time most systems make use of full form lexicons and perform their analysis on a word basis. Increasing demands on the lexicon size on the one hand and the need to limit the necessary training time .On the other hand will make morpho-based recognition systems more attractive.

V.CONCLUSION

In this article I surveyed the different algorithms used in the Odia computational morphology and discuss the applications of morphology.

REFERENCES

[1]Finite State Morphological Processing of Oriya Verbal Forms by Kalyani R.Shabadi from Resource center for Indian language technology Solutions, Department of Management Studies, Indian Institute of Science,Banglore-560012,india
[2]Analysis and Design of Oriya Morphological Analyzer: some test with orient by Sanghamitra Mohanty,Prabhhat Kumar Santi, K.P Das Adhikari from PG Department of Comp.Sc and Application, Utkal University, Bhubaneswar, Orissa
[3]Developing Oriya Morphological Analyzer Using Lt-toolbox by Itisree Jena,Sriram Chaudhury,Himani Chaudhry,Dipti M.Sharma.
[4]R.C Balabantray, M.K.Jena S.Mohanty presented a paper named "Shallow Morphology based complex predicates extraction in Oriya
[5]Morphological Analyzer for Sambalpuri Odia Dialect Inflected Verbal Forms by Dhabal Prasad Sethi at international of advance research in computer and software engineering October,2013
[6]ccl.pku.edu.cn/doubtfire/nlp/lexical_analysis/word_lemmatization /introduction/computational%morphology.htm
[7] Computational morphology and natural language parsing for Indian languages: a literature survey by Antony P J,DR k P Soman from AMRITA Viswa Vdyapeetham University, Coimbatore, India at international journal of computer science & engineering technology,vol.3No.4 April2012.
[8]"Computational model for reduplication in odia" by Rakesh Chandra Balbantray ,Sanjay Kumar Lenka at international journal of computational linguistic and natural language processing volume2Issue2February2013,ISSN2279-0756
[9]"A computational analysis of Nepali morphology: model for natural language processing "by Balaram Prasain a PhD dissertation Submitted to the Faculty of Humanities and Social Sciences of Tribhuvan University in Fulfillment of the Requirements for the Degree Of doctor of philosophy in linguistic.
[10] Designing Hybrid Approach Spell Checker for Oriya by Harihar Padhy and Sanghamira Mohanty at International Journal of Latest Trends in Engineering and Technology volume2, issue4, july2013, issn2278-621X
[11]two level morphology: a general computational model for word recognition and production by kimmo koskenniemi,university of Helsinki ,Department of General Linguistics,Hallituskatu11-13,SF-00100 HELSINKI 10 FINLAND
[12] http://en.wikipedia.org/wiki/Lexicography
[13] http://en.wikipedia.org/wiki/Speech_recognition

BIOGRAPHY

Dhabal Prasad Sethi is currently working as a lecturer in Computer Science & Engineering at Government College of Engineering, Keonjhar, Odisha. He has completed his Bachelor of Engineering in CSE from BIET, Bhadrak in 2006 then completed his Master of Engineering in CSE from PG Department of Computer Science and Application, Utkal University, Bhubaneswar, Odisha in 2011. Before he has presented 4no.s Paper at different International Journals. This is his 5th no at international level. His area of interest is natural language processing, information retrieval, software engineering, data mining.

