LANGUAGE IN INDIA
Strength for Today and Bright Hope for Tomorrow
Volume 10 : 9 September 2010
ISSN 1930-2940

Morph-Synthesizer for Oriya Language A Computational Approach

Rudranarayan Mohapatra, M.A., M.Phil., Ph.D
Lipi Hembram, M.A., M.Phil., Ph.D.

Abstract

Dealing with agglutinative languages like Oriya, Morph-Synthesizer plays a vital role for machine translation system in order to increase the output accuracy level. To build a morph-synthesizer for a language, it is necessary to take care of the morphological peculiarities of the language, specifically in Machine Translation (MT). In this paper, we describe our work on rule-based Oriya morph-synthesizer. Here we have concentrated only on the synthesis of the Nouns. Noun synthesis in Oriya depends upon the feature (animacy and honorific) and demands the semantic account and behaviors of the noun-endings.

Key Words: Morpheme, Post positions, PP Decision Maker (PPDMP), gender, Number, person, Animacy

1. Introduction

The present work aims to build a computational model for the analysis and generation of morphological-synthesizer in Oriya language, the language being one of the official languages of the state of Orissa, situated in the eastern part of India. It belongs to Eastern Indo-Aryan group. Some peculiarities of this language - the usages of classifiers, feature base agreement, and semantic contextual agreement, etc., make it morphologically complex and, thus, a challenge in

NLG. Generation of syntactically and semantically correct sentences needs appropriate choice among different forms of words. In this paper, we discuss the features of morph-synthesizer in Oriya language, specially focusing on nominal synthesizer. Through its implementable algorithm, a suitable conclusion is drawn.

2. Oriya Language and Morpheme

In Oriya language, a morpheme termed as the smallest unit in the language that carries and conveys a unique meaning and is grammatically appropriate. A morpheme in Oriya is the most minuscule meaningful constituent which combines and synthesizes the phonemes into a meaningful expression through its form and structure. Thus, in essence, the morpheme is a structural combination of phonemes in Oriya. In other words, in Oriya language, the morpheme is a combination of sounds that possess and convey a meaning. However, a morpheme is not necessarily a meaningful word in Oriya. Morphemes are the smallest units of sentence analysis and include root words, prefixes, suffixes, and verb endings. So in Oriya, every morpheme is either a base or an affix prefix or a suffix. For example, 'apraakrutikataaru' has four morphemes as 'a-praakrutika-taa-ru'.

Again the major morphological contents in Oriya language are Pronoun Morphology, Inflectional Morphology and Derivational Morphology on which our Morphological synthesizer has been built up (Mohapatra Pandit N. and Dash S., 2000, *Sarbasara Byakarana*, New Students Store, Cuttack, Orissa, India). Taking the above content with a hybrid combination of syntax and semantic features the synthesizer exercise is here being tried out.

3. Morph-Synthesizer in Oriya language:

In Oriya, this choice for a Nominal form depends mainly upon the feature of a noun, ending of a noun and its subject or object position. Especially the Gender value in Oriya for Morph analyzer or synthesizer is very less in comparison to grammatical number (singular or plural) and honorific marker. With this GNP feature the post positional case marker looking to the involvement of state/event/action are also played a significant role for a proper Nominal form synthesis For verbs, TAM (tense, aspect, modality), person and verb-root information play a major role.

Nominal form Synthesis can be further classified as Post-position synthesis (e.g. bapa-nkara` of father' vs. bahi-ra `of the book') and Classifier synthesis in Oriya depends upon two aspects. One is honorific or personification marker and another depends upon number. In the case of honorific (pila-mane `children') it requires agreement to synthesize. Bahi-guDika `books') does not need this. Classifier synthesis is again subject to plurality (baLaka-Ti `the boy' vs. chhabi-gudika `pictures').

Verbs also show two types of synthesis depending upon the verb root of active and passive cases. An active verb root takes another form when it is changed into passive (khaIba-khuAiba 'eat'), adding tense and person feature (kha-Ichi `1st P/3rd P-pres & Singular' vs. khaUchhanti `2nd P-Hon-pres' or 3rd P-pres). It is found in Oriya that verb synthesis depends not only on person and tense information but also on the verb root ending. Verb root ending is another specific feature of Oriya which gives the language a proper agreement shape.

3.1 Nominal-Synthesis in Oriya

Oriya is a syntactically head-final and morphologically agglutinative language. It has natural gender (as opposed to grammatical gender). That is**,** gender is dictated by semantic rather than formal characteristics of words. Gender is not specified in the agreement features, nor does it affect other grammatical categories like pronoun or verb.

To handle the synthesis of constructions in Oriya, the nouns and the inter-relationships of the verbal and the nominal forms are followed after karaka formalism in a defined context. The morphological rules written for the synthesizer can be seen through the synthesized output.

The input is always a lexical string that is concatenated or synthesized with suffix endings. The grammatical tradition in Oriya follows the demand and merit (i.e., aakaanksha and yogyataa) tradition. Verbs in the language demand the karaka identities and the subject and object parts fulfill the demands according to the yogyataa or level of agreement. And, in a defined context, nouns demand post-position on a semantic account. So, the morph-synthesis in Oriya language not only depends on the feature-based agreement but also it depends on the whole subject or sentence semantic expression. In the flow of text, a noun having animate and non-honorific categories with honorific plural marker, takes the identity of animate and gets honored. So, the nouns like 'Saanga (friend), gay (cow) take both the plural form as 'SAngamAne' or SaangaguDika / gaaiguDika or gaaimAne.

In this paper, we would explain the Noun synthesis in two sections: post-positional synthesis and classifier synthesis. As shown in figure 1, Noun synthesis in Oriya is of two types. We discuss Post-position synthesis below.

3.1.1 Post-position Synthesis in Oriya Language

Oriya has mainly two types of post-positions as opposed to the preposition of English, named as bound post-positions and free post-positions. Bound post-positions do not stand alone and they have agreement with noun by morph synthesizer rules like addition, deletion, insertion, etc., whereas free post-positional markers stand alone irrespective of the condition.

In Oriya, the post positions start from a 'maatraa' (half-vowel) to the post-positional word. This makes the synthesis more complex and four-eyed in nature. It is sometimes only a suffix (e.g. 'e-maatraa' as Lok-e (peoples), sometimes a combination of a suffix and a post-positional word (e.g. pila-nka-madhyae "among the children". Again Oriya Nominal post-positional markers generally follow Paninian case marker principles.

We have listed the Post positions of Oriya and their equivalents in English. Table 1 gives the some Oriya post-positions and their English equivalents.

Post positional Marker	Singular	Honorific (Singular/ Plural)	Equivalent English frequent Prepositions
Nominative	Ø, Ti, Taa,	-e	ø
Accusative	ku	nku, maananku	to
Instrumental	re, dwara, dei	nkaje, maananka-dwara, nka-dei,	by, with
Dative	ku	nku, maananku	to
Ablative	ru, Thaaru, Thun,	maanankaru, maanankaThaaru, Maanankathum	from
Genitive	ra	nkara, maanankara	of
Locative	re, Thaare, Thin	maanankare, maananka-Thaare, Maanankathim	at, in, on

Algorithm

To do the morph synthesis effectively, especially in Nominal Morph-synthesis, we have considered the Noun phrase with its adjacent Post-positional class for simpler, machine learning based Synthesizer.

The three key issues in Noun Morph-synthesis are:
1. Choice of Appropriate suffixes,
2. Ordering of the suffixes, and
3. Surface level changes in the boundary at the time of affixation.

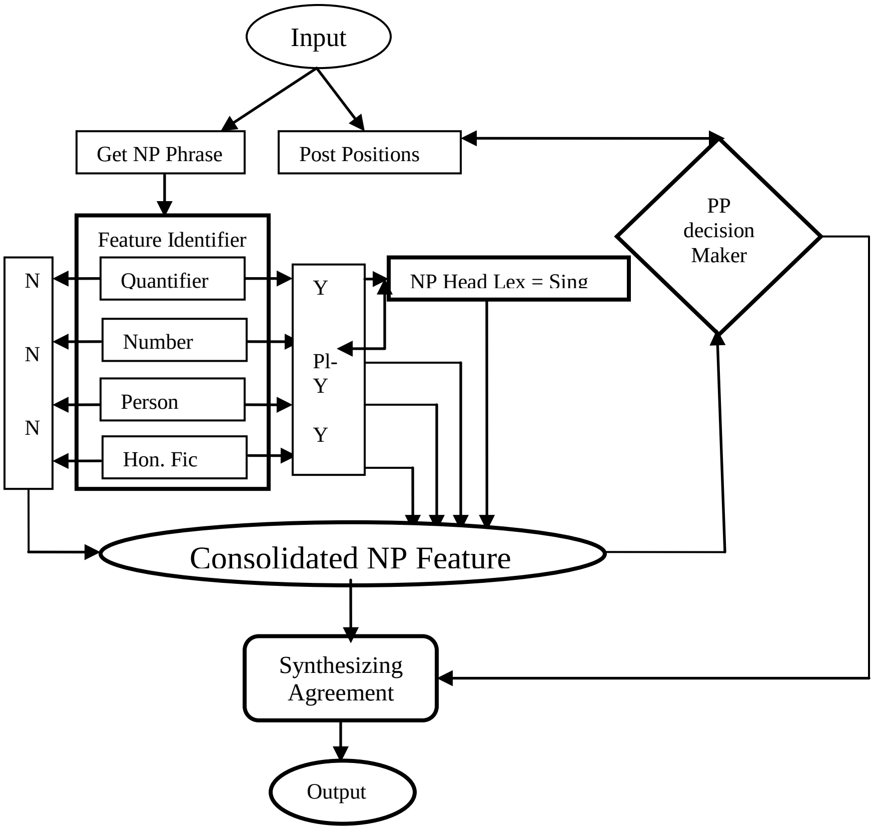
The details of the Nominal morph-synthesizing are given in diagram No 1.1.

According to this, we have considered the Noun Phrase ‘Nx’ and adjacent avail Postpositions is ‘Px’ as the input. Where Nx = {w0, w1, w2…..wn} and Px = {p1, p2, p3…Pn}

After input in Nx & Px, the Nx part will enter our feature identifier module to get the individual features of w0 to wn where every token i.e. w0 to wn are the set of nouns or adjectives in the set. Nx and wn would be the head word.

In feature identifier at present, we have considered a minimal feature at an initial level and set for every possible feature true as value ‘0’ and false as ’1’. If any set of Nx has the bullion value set ‘0’ for the feature Quantifier, then the wn would consider its singular lexeme form, irrespective of the plural features.

Again if the features are not available for an unknown lexeme, then we consider its ending to get shallow feature information to smooth the process ahead. Then the Consolidated NP feature module will finalize the Noun lexicons and the consolidated feature of the Nx and the output of this module would be supplied to the ‘PP Decision Maker (PPDMP’ to gain the decisive Post positions. ‘PP Decision Maker’ will make coordination with possible postpositions and the decisive post-positions will be sent to the Synthesizing Agreement. The synthesizing agreement with its internal rules makes an agreement with the Nx outcome from Consolidated NP feature to get the final synthesized output.



Nominal Synthesizer Diagram 1.1

The Synthesizer agreement will work in the principle of Finite State Automata. Finite State Automaton (FSA) is a set of principles that receives a string of symbols as input, reads the string one symbol at a time from left to right, and after reading the last symbol halts and indicates either acceptance or rejection of the input. The automaton performs computation by reacting on a class of inputs. The concept of a state is the central notion of an automaton. A state of an automaton is analogous to the arrangement of bits in the memory banks and registers of an actual computer. Here, we consider a state as a characteristic of an automaton which changes during the course of a computation and serves to determine the relationship between inputs and outputs.

By this **NP = {[ά -*x] + *z + p},**
Where,

ά = Availed Noun form,
*x = possible deletion of morpheme in conditional environment
z* = possible addition of morpheme in conditional environment
p = decisive post position.

The rules are implemented by seeing the exact features positions and numbers. Certain constraints are also added to these rules. For example, in a set of input we get (keteka pilaamaane + ku). Here ‘keteka’ is a quantifier used to modify the head word ‘pilaamaane’ whose set of features are: Noun, 3rd third person, plural, honorific, Neuter gender. So, irrespective of plural feature we consider the lexeme of ‘pilaamaane’ as ‘pilaa’ only as its singular base form. And taking the agreement principles, ‘pilaa + ku’ will call an extra addition of morpheme, i.e., ‘ή’ and make the final synthesized output as: keteka pilaanku.

Hence, the proposed Oriya morphological synthesizer not only covers the syntactic and semantic aspects including case-ending, noun-class and noun-ending but also the socio-linguistic parameter in its matrix. Except this, the whole thought boundary subject and Verb agreement, etc., play important roles in increasing the Nominal morph-synthesizer accuracy which continue to under our observation and consideration for further improvement.

The application of this synthesizer in a real MT system has enhanced the accuracy of the generation noticeably. We will also observe by taking text from daily news papers. The operation shows that more than 50% of nominal synthesizer would be handled effectively by this method without the micro level higher order semantic and pragmatic features.

4. Conclusion

In this paper we have tried to describe analysis of Oriya morphological features to develop a Morph-synthesizer for Oriya language. We identified that some voluntary particles in the context of subject and object, and other classifiers have also played a significant role and have been exceptionally handled by sophisticated morpho-syntactic analysis. This process will open a new door in the area of Natural language Generation especially in Oriya language and would increase the output accuracy level of the concerned language.

Reference

Mohapatra Pandit N. and Dash S. (2000). *Sarbasara Byakarana*, New Students Store, Cuttack, Orissa, India

Mohanty S. et al. "Object Oriented Design Approach to OriNet System: On-line Lexical Database for Oriya Language", IEEE Proceedings of LEC-2002, University of Hyderabad, Hyderabad, India.

Mishra S.K., et. al., ‘Identifying Verb Inflections in Sanskrit Morphology’, Proceedings of ‘simple05’, IIT, Kharagpur.

Robert A. Dooley. Source-Language versus Target-Language Discourse Features in Translating the Word of God. Journal of Translation, Volume 1, Number 2 (2005).

Rudranarayan Mohapatra, Ph.D.
C-DAC
Pune 411007
Maharashtra, India
Rudra1979@gmail.com

Lipi Hembram, M.A., M.Phil, Ph.D
Utkal University
Bhubaneswar 751004
Orissa, India
Kamalakanta2007@gmail.com