A STUDY ON AN APPROACH TO GENERATE PRONOUN IN ODIA LANGUAGE

Aira Kharvel Parida[1], Pinaki Sankar Chatterjee[2]
[1] School of Computer Engineering, KIIT University, Bhubaneswar - 751024
[2] Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur, India – 721302
Email: {aira_kharvel[1], pinaki_india[2]}@rediffmail.com

Abstract— Natural Language Processing (NLP) is a field which describes how a computer system can understand the language used by human being. Pronoun generation and resolution is an important aspect of NLP. Applying pronoun to a compound sentence is the way to represent it in which the human beings are addicted to use it. With the help of meaningful pronoun we can make a compound sentence more fluent. On the other hand if we use wrong pronoun then the language will lose its original meaning. A considerable amount of work has been done on various non Indian languages in this field. Some of our Indian languages have also developed such system. Odia language processing has just started. In this paper our main aim is to design an algorithm which will meaningfully generate pronoun in compound sentence. So here for our system input will be a compound sentence without pronoun and output will be the meaningful compound sentence with pronoun.

I. INTRODUCTION

Summarization and question answering are both examples of natural language processing systems that produce natural language output, and thus require some sort of text generation module. The degree of sophistication in the text generation can vary widely, but given the high frequency of pronouns in natural language text, it is natural to expect that a proper treatment of pronouns in summaries and responses might lead to better quality output. Here a simple approach is introduced to find out a solution to overcome the problem of pronoun generation in Odia language. As a regional language the use of this language is very limited. But in order to make a perfect system we have to develop a system which can process all most all languages which have some valid syntax, valid grammar and a specific community of people use it for their day to day life communication.

Natural Language Processing is a subfield of Artificial Intelligence and linguistic, developed to make computers understand statements written in human language [1]. So "Natural Language Processing is a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a

range of tasks or applications." A natural language or ordinary language is a language that is spoken, written by human beings for general purpose communication. A language is a system, a set of symbols and a set of rules (of grammar). The symbols are combined to convey new information and the rules governs the manipulation of symbols [1]. Besides the well established problem of pronoun resolution, pronoun generation is now attracting renewed attraction. In the past system generated pronouns without attaching much importance to the problem, one notable exception being the classical algorithm of Dale(1990), loosely based on centring theory [3].

II. MOTIVATION

Now a day we humans are trying to make our life more secure as well as easier. For that we are trying to innovate ideas which can do so. Natural Language Processing (NLP) is one step ahead in this string. The basic purpose of it is not to just use a system rather interact and communicate with the system.

Considerable works have been done in the field of pronoun generation for non-Indian languages, such as, English, Dutch, Spanish, Chinese, and Japanese etc. But the NLP technologies developed for the Indian languages are still in their infancy. But there are some Indian languages where this technology has prove its presence and some have yet to go through it. As Odia is a regional language, so the effort to NLP in this language is a little slower.

III. RELATED WORK

Several works have been done on pronoun generation through Natural Language Processing. But most of works done are in non-Indian Language.

A. *Pronoun Generation through XML*.

Summarization and question answering are both examples of natural language processing systems that produce natural language output, and thus require some sort of text generation module [4]. The degree of sophistication in the text generation can vary widely, but given the high frequency of pronouns in natural language text, it is natural to expect that a proper treatment of pronouns in summaries and responses might lead to better quality output.

As Lingpipe[1] can find all the referents to a specific entity, so by running Lingpipe once on the text, we would have a chain of entities, all with the same referent. The goal is for the latter entities to be systematically replaced by pronouns referring to the former entities. The algorithm can be divided into two phases: *replacement* and *validation*. In the first phase, an appropriate pronoun is chosen and the text is regenerated with the specific entity replaced by this pronoun. Then, Lingpipe is used to validate the replacement. In case of valid replacement, the pronoun will remain in the final text

[1]Lingpipe is a suite of natural language processing tools written in Java that performs tokenization, sentence detection, named entity detection and co-reference resolution on text. The input is plain text and output is an XML file with embedded tags inside the original text.

. In order to suggest a pronoun out of the pronoun set (he, she, his, her, him), we have to deal with gender and case (nominative, accusative, possessive) as part of the replacement phase. Since Lingpipe is not able to guarantee grammaticality, we cannot deal with grammaticality when we use it in the validation phase. Use of XML in pronoun generation is proved a fruitful concept. As it's a scripting language which uses the input which a user defined him/herself. This process generates a unique ID for each noun as well the type of pronoun. According to the ID and type it replaces the pronoun in the place of noun.

The basic disadvantage in this approach is that it works properly for English language not for odia language.

B. *Centering approach to pronouns.*

Formalization of the centering approach [3] to modeling attentional structure in discourse and use it as the basis for an algorithm to track discourse context and bind pronouns. The process of centering attention on entities in the discourse gives rise to the intersentential transitional states of continuing, retaining and shifting. The basic idea is that to handle the multiple ambiguity pronouns. A discourse exhibits both global and local coherence. On this view, a key element of local coherence is centering, a system of rules and constraints that govern the relationship between what the discourse is about and some of the linguistic choices made by the discourse participants. Pronominalization in particular serves to focus attention on what is being talked about; inappropriate use or failure to use pronouns causes communication to be less fluent. For instance, it takes longer for hearers to process a pronominalized noun phrase that is *not* in focus than one that *is*, while it takes longer to process a non-pronominalized noun phrase that is in focus than one that is not.

Centering model is based on the following assumptions. A discourse segment consists of a sequence of utterances U1, . . . ,Um. With each utterance Un is associated a list of forward looking centers, Cf(Un), consisting of those dis- course entities that are directly realized or realized by linguistic expressions in the utterance. Ranking of an entity on this list corresponds roughly to the likelihood that it will be the primary focus of subsequent discourse; the first entity on this list is the preferred center, Cp(Un). Un actually centers, or is "about", only one entity at a time, the backward-looking center, Cb(Un). The backward center is a confirmation of an entity that has already been introduced into the discourse; more specifically, it must be realized in the immediately preceding utterance, Un−1. There are several distinct types of transitions from one utterance to the next. The typology of transitions is based on two factors: whether or not the center of attention. Cb, is the same from Un−1 to Un, and whether or not this entity coincides with the preferred center of Un.

Definitions of these transition types appear in below Figure-1.

	$Cb(U_n) = Cb(U_{n-1})$	$Cb(U_n) \neq Cb(U_{n-1})$
$Cb(U_n) = Cp(U_n)$	CONTINUING	SHIFTING
$Cb(U_n) \neq Cp(U_n)$	RETAINING	

(Figure-1)

These transitions describe how utterances are linked together in a coherent local segment of discourse. If a speaker has a number of propositions to express, one very simple way to do this coherently is to express all the propositions about a given entity (continuing) before introducing a related entity (retaining) and then shifting the center to this new entity.

IV. Our Approach

1. Odia Language Analysis:-

As this paper consists of the topic of generation of pronoun in Odia Language, it is much important to understand the various faces of this language. Among the various aspects we basically consider the "noun" and "pronoun" part of this language because to generate a pronoun in compound sentences firstly we have to consider the noun. In our day to day life we do that.

The main work begins with the types of noun, we the normal people use in our life in Odia language. It is observable that we are using basically four types of noun in our day to day life. These four types of pronoun can be identified as

1) Time (ସମୟ)
2) Person (ଲୋକ)
3) Place (ଜାଗା)
4) Object (ଜିନିସ)

If we have 4 types of noun then we have also 4 types of pronoun also. Accordingly we have to place the appropriate pronoun in compound sentences. Considering this fundamental thing the blue print of pronoun generation in Odia language is presenting below.

2. System Architecture:-

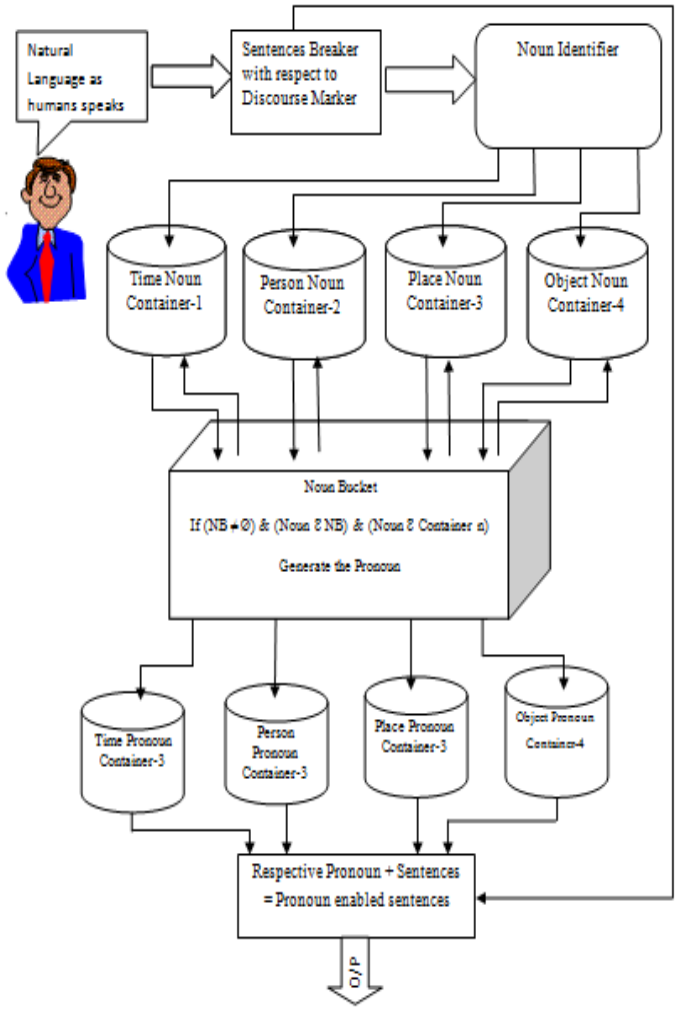


Figure-2 (System architecture)

3. Methodology: - In this section it is described how this blue print will work.

Step-1: Take an appropriate compound sentence where a pronoun can be generated. (In Odia)

Step-2: Break the total sentence with respect to the discourse marker [3] present within it. Count the resulting sentences.

Step-3**:** Give these sentences to the noun identifier. Here the types of noun are identified and it will store the nouns in its appropriate noun container (NC).

Step-4: Here noun will go into the noun bucket (NB). Here some logic has to implement.

(i) If SC=1,

(ii) Store the noun in NB.

(iii) Increment SC with 1.
Now SC=2

(iv) Check the following condition for generating pronoun.
If (NB≠Ø) && (Noun Ɛ NB) && (Noun Ɛ NC n)
Where n=1, 2, 3, 4
Then pronoun will generate.

Step-5: The noun bucket will map the corresponding pronoun bucket and get the appropriate pronoun [4].

Step-6: Combining the output of Step-2 with Step-5 the sentence with present with correct pronoun.

Another problem may appear which is called "ambiguity" [5]. This problem may same in Odia as well as in English language. It states that if we generate a pronoun in the deeper inside a paragraph the system may get confused for which noun this pronoun is generated, which seems a big problem. So for our approach we will overlook these types of sentences for processing.

4.A simple example:-

Let's take a simple example in Odia language

Step-1: Rama jauchhi o Rama kheliba (ରାମ ଯାଉଛି ଓ ରାମ ଖେଳିବ) (Rama is going and Rama will play)

Step-2: As "o" i.e. "ଓ" is a discourse marker in Odia language, so we have two sentences

I) Rama jauchhi (ରାମ ଯାଉଛି)

II) Rama kheliba (ରାମ ଖେଳିବ)

Step-3: As given to the noun identifier it will find out the noun. For sentence (i) it is Rama and for (ii) it is Rama.

Step-4: In both sentences the noun Rama comes under the "person" noun.

(i) SC=1

(ii) Store Rama in NB.

(iii) SC=SC+1
SC=2

(iv) (NB≠Ø) -- Condition Satisfied.
(Rama Ɛ NB) -- Condition Satisfied
(Rama Ɛ NC 2) -- Condition Satisfied
Then pronoun can be generated.

Step-5: NB will connect to the PB i.e. NB 2→PB2 where the corresponding pronoun was stored earlier. In this case it is "se" or "ସେ ".

Step-6: OP Step-2: Rama jauchhi (ରାମ ଯାଉଛି)

OP Step- 5: Se ସେ

OP Step-2: Rama Kheliba (ରାମ ଖେଳିବ)

Now the resulting sentence with proper discourse marker is:

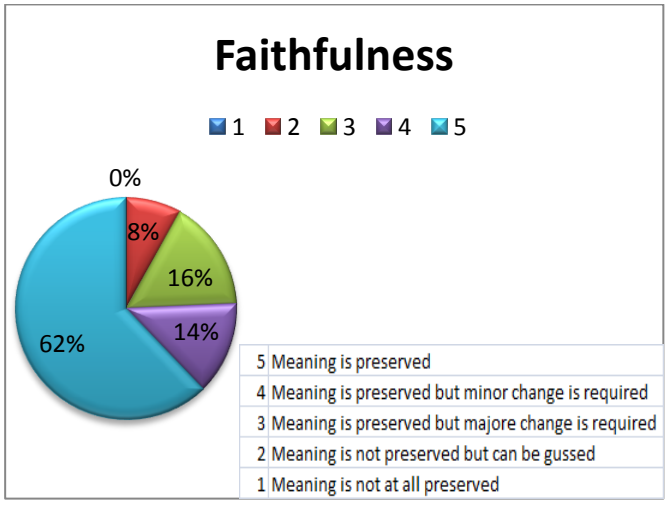
Rama Jauchi o se Kheliba (ରାମ ଯାଉଛି ଓ ସେ ଖେଳିବ)

5. Evaluation:-

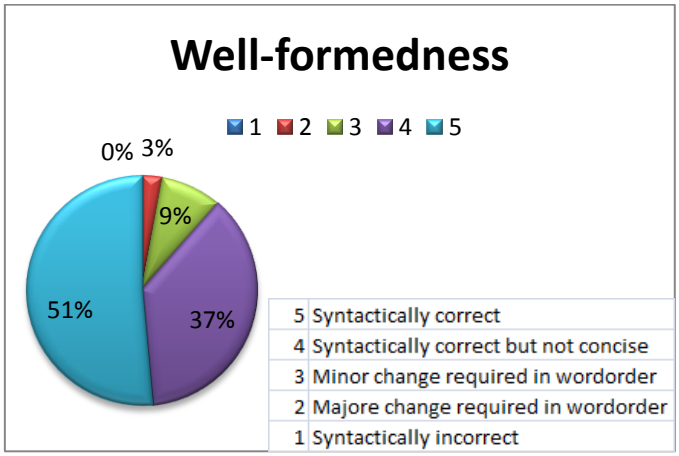
We have developed a system which generates the pronoun in compound sentences following the approach mentioned above. We performed a user based evaluation to validate our approach. The system outputs were shown to twenty three human evaluators, all of them were native speaker of Odia, and they were asked to rate the output Sentences in a scale of 1 to 5 based on some parameters. The evaluation is performed on the basis of

- *Well-formedness:* We define the well-formedness of an output sentence by its grammatical correctness and conciseness. The grammatical correctness measures the accuracy of the syntax, word in the sentences.
- *Faithfulness:* The faithfulness of an output measures how well the communication goal is preserved by the generated output.

The overall performance of our system for well-faithfulness and faithfulness are shown in Figure-3 and Figure-4 respectively.



(Figure-3)



(Figure-4)

V. CONCLUSION AND FURTHER WORK

Natural Language Processing is a vast field which is under research. This report basically focuses on a simple approach for pronoun generation in Odia Language. As a regional language the work of NLP in odia language is very few. The approach which we have described needs some further refinery. The next step is to polish this algorithm and to make these four types of pronoun database, so that its efficiency will increase.

Another problem ambiguity will take into consideration for this process for further work.

ACKNOWLEDGMENT

We take the opportunity to express our sincere gratitude to Mr. Pinaki Sankar Chatterjee (School of Computer Engineering), School of Technology, KIIT University and School of Computer engineering, KIIT University for providing valuable guidance, encouragement and a great environment throughout the thesis work. Also thanks to the people who participate in the evaluation procedure

REFERENCES

[1] Natural Language Processing:AI Course Lecture 41 www.myreaders.info/ RC Chakraborty, e-mail rcchak@gmail.com, June 01,2010

[2] Mehdi M. Kashani, Fred Popowich "*Pronoun Generation for Text Summarization and Question Answering"* ,School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada {mmostafa, popowich}@sfu.ca

[3] Susan E. Brennan, Marilyn W. Friedman, Carl J. Pollard "*A CENTERING APPROACH TO PRONOUNS*", Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA

[4] Kalyanamalini Sahoo,"*The Generic Subject Pronoun in Oriya"*,Handout:Impersonal Pronouns, Paris 20 Sept. 2011.

[5] Kathleen F. McCoy, Michael Strube "*Generating Anaphoric Expressions: Pronoun or Definite Description?"*

[6] L. Danlos & F. Namer, **"** *A Global Approach for Pronominalization in Text Generation"*

[7] Sanghamitra Mohanty, Prabhat Kumar Santi, K.P. Das Adhikary"*Analysis and Design of Oriya Morphological Analyser : Some Tests with OriNet",* Utkal University,Bhubaneswar, Orissa, India- 751004

[8] B.J. Grosz, A.K. Joshi, and S. Weinstein. "*Providing a unified account of definite noun phrases in Discourse"*.In Proc., Blst Annual Meeting of the ACL, Association of Computational Linguistics, pages 44-50,Cambridge, MA, 1983.

[9] Charles B. Callaway, *"Pronominalization in Generated Discourse and Dialogue",* Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002

[10] Anil Kumar Singh, *"Natural Language Processing for Less Privileged Languages:Where do wecome from? Where are we going?"* Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 7–12, Hyderabad, India, January 2008.