A Rule Based Approach for Anaphora Resolution

Varsha Singh, Nisheeth Joshi

Abstract: ***In Natural Language Processing one of the challenges is how to determine which entity is referred in the discourse and how they correlate to each other? Anaphora resolution is one of the most important area in Natural Language Processing. In this paper, we present rule based approach to identify different types of anaphora and its antecedents for Hindi. We use Hidden Markov Model to POS tag the dataset and cases are used for the identification of preferred noun. At last, apply Centering algorithm for identification of pronoun and its referents. We have tested the system with 500 segments to resolve the anaphora.***

Index Terms: Anaphora Resolution, Hidden Markov Model, Case Markers, Centering Algorithm.

I. INTRODUCTION

Natural language processing is provide the platform where communicate the humans and computers. Natural language has various anaphoric expressions and these expressions make it very interesting. Anaphora is an expression which relates to an expression to another expression which preceding it in the discourse. It represents sentences in a very attractive manner without changing its context. The element that refers to anaphora is antecedent. Discourse is the group of collocated sentences or communication between more participants.

Table 1. Resolution of Anaphora

S. No.	Sentences	Antecedents	Anaphora
1.	सीमा बहुत अमीर है, पर वह सुखी नही है।	सीमा	वह
2.	संतोष पढ़ने में बहुत अच्छा है। उसे गणित पढ़ना पसंद है।	संतोष	उसे

In Table 1, 'वह' and 'उसे' is the referent that refers the 'सीमा' and 'संतोष' respectively.

In NLP (Natural Language processing) different types of anaphora available such as one anaphora, pronominal anaphora, zero anaphora, nominal anaphora, definite noun phrase anaphora. Various challenges in NLP like as summarizing text, question answering system, extracting information, machine translation, etc can be resolved by anaphora resolution.

Anaphora resolution is the task for identify and resolve noun phrases that refer same real world entity. It is also responsible for cohesion and coherence. Cohesion is the evident feature of providing text with the reference of the connectedness of the previously defined text part.

Example: 1. वैभव मुम्बई में रहता है। उसे घुमना पसंद है।

In the above example, we observe that second sentence is referred the first sentence and hence the example is cohesive. The referent of pronoun 'उसे' is 'वैभव' concrete entity. Hence this is an occurrence of entity reference. The work of anaphora resolution is identified the legitimate referent of the pronoun 'उसे' which in above example is 'वैभव' .

Anaphora resolution systems are developed for particular languages. In India multitudinal languages used in enormous culture and diversities. Indian languages are morphologically rich and knowledge poor. The research work of anaphora in Indian languages is done very less. In the Indian languages one word has different meaning, it also generate ambiguities in the anaphora resolution. Hindi is independent word order language. In this paper we discuss the rule-based approach of the anaphora resolution by applying centering algorithm. In English 'it' has great importance, but translation in Hindi is quite irrelevant.

Entire paper is categorized into six sections. Section II explains the impact of cases in anaphora. Section III represents the related work on anaphora resolution although in Section IV defines the experimental work for anaphora resolution. In Section V evaluate the accuracy of the model and Section VI concludes the paper with short discussion.

II. GRAMMATICAL FRAMEWORK

A. Cases

B. The relationship of a noun or pronoun with the adjacent word is defined by cases. In Hindi, pronouns do not differentiate gender, for the gender description using the verb and case markers. The case marker analyses referent for pronoun of the anaphora resolution.

Cases plays very crucial role for resolving anaphora. Hindi is a free word order language i.e. a sentence can be structured in different forms.

Revised Manuscript Received on May 06, 2019

Varsha Singh, Department of Computer Science, Banasthali Vidyapith, Rajasthan, India.
Nisheeth Joshi, Department of Computer Science, Banasthali Vidyapith, Rajasthan, India.

Example: 2(a). स्वागता को लड्डु खाना पसंद है।

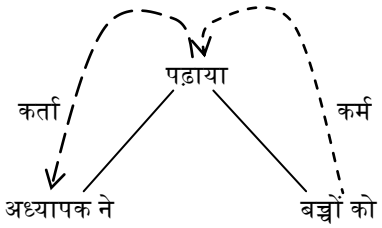
(b). लड्डु खाना स्वागता को पसंद है।

Both the sentences are syntactically correct having same meaning; but construction of sentences is different. Due to abstract nature of Hindi it's difficult to identify the subject of a sentence. For resolving this problem we can use the case markers. The following table defines the case markers in Hindi:

Table 2. Cases in Hindi

S. No.	Cases	Case Markers
1.	कर्ता(Karta)	ने
2.	कर्म(Karma)	को
3.	करण(Karna)	से, के द्वारा
4.	सम्प्रदान(Sampradana)	के लिये
5.	अपादान(Apadana)	से(अलग होने में)
6.	सम्बोधन(Sambandhana)	का, की, के
7.	अधिकरण(Adhikarana)	में

Example: 3. अध्यापक ने बच्चों को पढाया।



C. Pronouns

In a segment, if a noun is used repeatedly it looks offensive. To make the line cohesive pronouns are used in place of noun. Pronouns are flexible in nature. For example 'वह' refers to 'लड़का' , ' सीता' , ' कलम' , ' गौरव' etc.

Table 3. List of Pronouns

Number	First Person	Second Person	Third Person
Singular	मैं, मेरा,मेरी, मुझे, मुझको, मुझसे, मुझमें, मैंने, मेरे, अपना, अपनी, अपने	तुम, तू, तुम्हे, तुमको, तेरा, तेरे, तेरी, तुने, तुम्हारा, तुम्हारे, तुम्हारी, तुमसे, तुझे, तुझसे, तुमपर, तुझपर, तुमने, आप, आपका, आपके, आपको, आपसे, आपकी, आपपर, अपने	यह, यहाँ, यही, वह, वहाँ, वहीं, उस, उसे, उसी, उसका, उसकी, उसको, उससे, उसने, उसमें, उसपर, इस, इसे, इसी, इसको, इसका, इससे, इसमें, इसपर, इसने, इसकी
Plural	हम, हमें, हमको, हमसे, हमारा, हमारे, हमपर, हमने, हमारी		ये, वे, उन, उनका, उन्हें, उनकी, उनसे, उनके, उनपर, उन्होने, उनको, उनमें, इन, इन्हें, इनका, इनसे, इनपर, इनकी, इनको

Table 4. Types of Pronouns

Pronouns	Pronouns in Hindi
Indefinite Pronouns	कोई, किसी, किन्हीं, कुछ, सब, सबने, सबको, सबसे, सबका, सबमें, सबने, सबके, प्रत्येक
Interrogative Pronouns	कौन, कहाँ, क्या, किसने, किसको, किससे, किसके, किसका, किसकी, किसमें, किनमे, किनको, किनसे, किनकी, किनका, किन्होने
Relative Pronouns	जो, जिसने, जिससे, जिसका, जिसमें, जिसको, जिसपर, जिन्होने, जिनका, जिनको, जिनसे, जिनके, जिनमें, जिनपर

III. LITERATURE SURVEY

This section, focused on the previous work that has been done on anaphora resolution in Indian languages. Ashima et al. [1] worked on hybrid approach for resolving the animistic knowledge, co-reference resolution, gender and number agreement in Hindi domain.

Prayalankar and Devi [6] developed the rule based pronominal resolution system, which identifies the various pronominal and its antecedents in Sanskrit. For anaphora resolution used the rule based approach, corpus based approach, knowledge poor approach and discourse based approach in Indian languages [10]. They [4] resolve different aspects of anaphora resolution such as Recency factor, Animistic knowledge, agreement of gender, agreement of number, Named Entity Recognition, resolving Pronouns etc. and various limitations. Anaphora resolution [19] is used to identify the opinion target having prepositions rather than nouns in sentences and analyze the outcome of benchmark datasets. In this paper[18], build the anaphora resolution tagger for tagging all the pronouns and nouns among different criteria. The work is divided into two parts. In first part, define the cluster on the basis of bottom up approach and the second part, defines three techniques of the reverse hierarchical clustering on similar cluster for the identification of Bangla News Recommendation.

This paper [13] presents the Hindi Language Grammar in Backus-Naur Form, for analyzing the grammar or meaning uses the semantic analysis. The pronominal anaphora resolved by the Centering and modified Hobbs algorithm [12]. They have used news articles and children's story for evaluating the system. In [15] the proposed model highly modular toolkit for multilingual co-reference resolution system BART that supports state of the art statistical approaches. It gives the best result in German and in Italian; the performance level is moderate due to various algorithms. Anaphora resolution in Hindi, more emphasize on the use of third person personal pronouns for define the relationship between anaphoric and noun phrases [2].Event anaphora resolution with hybrid approaches for abstract or concrete identification and use the language specific rules. Use Paninian grammar and proximity of events in anaphora resolution as main components. In event anaphora classify the lexical forms in different classes which are this, these, it, its, in this, therefore, etc. [8]. Sobha et al. defined Multilingual VASISTH model, which resolve the two languages Malayalam from Indo-Dravidian family and Hindi from Indo-Aryan family. Accuracy rate is high of the system in the case of Malayalam.

At the time of translation [16] various issues and challenges occur in the context. In resource poor language [11] used differential evaluation based feature selection technique and BART(Baltimore Anaphora Resolution Toolkit) in Bengali. The two stage pronoun reference makeable and links for the anaphora resolution by using data driven approach on three Indian languages: Hindi, Bengali and Tamil [14].

Divergence in English to Hindi translation occurs when similar structured sentences of the source language do not convert into the target language. An EBMT translate the sentences on the behalf of the past similar sentences [3].Hindi Dependency Treebank for the anaphora resolution and identify the issues which occur in the anaphora resolution [5].That model for the annotation used the 162 news items from the Treebank. They contain 2477 sentences with 2122 instances of pronouns, but 1408 pronouns are identified and the remaining 714 pronouns are not identified. They [9] defined GuiTAR off-the-shelf anaphora resolution system for Bengali.

In this paper, Dandpat et. al [22] presented work on Bengali POS tagging and combined both supervised and unsupervised learning for training bigram HMM. Two taggers are implemented, first one is HMM-S i.e. using supervised model and second one is HMM-SS i.e. semi supervised model. In Manipuri Language [21] using Stochastical model i.e. hidden markov model for tagged the data. In this system different models are used which are Tokenizer, Stemmer, Tagged Corpus, statistical Analyzer and Viterbi algorithm is used for finding maximum probability. They [23] defined POS tagger for Bahasa Indonesia, which has been developed using modified HMM to improve the accuracy. Firstly HMM tagger using Affix tree is compared to Baseline tagger which are trained and tested on same dataset, another experiment is done on several configuration for HMM POS tagger. The authors [17] have used HMM model and IL POS tag set which consists of 24 tag sets for the development of this tagger. This paper presents a POS tagger for nepali language using Hidden Markov Model. Annotated tool SANCHAY is used which contain around 1,50,839 tagged words. Including language specific attribute and generic attributes value 42 tags are included in the tag set. Supervised training method is used. For implementing the system NLTK library is used. Accuracy obtained is 96%.

IV. PROPOSED METHODOLOGY

A. POS Tagger

In NLP, Part of Speech (POS) tagging is very essential work in various fields such as machine translation, recommendation system, text summarization etc. The POS tagging is the procedure of allocating part of speech such as pronoun, noun, verb, adjective, preposition, adverb, or other marker of the words in sentences. It is a major part in morphological analyzer. The POS tagger allocates the distinct grammatical category of every word in a sentence.

There are lots of words having numerous meaning which creates ambiguity at the time of tagging. For example, 'सोना' can be represented as noun or verb. Following sentences shows the different tags for the same word having different meaning:

Example 4(a): प्रीती को सोना पसंद है ।
NN PSP VM NN VAUX PUNC

(b): प्रीती को सोने की अंगुठी पसंद है ।
NN PSP NN PSP NN NN VAUX PUNC

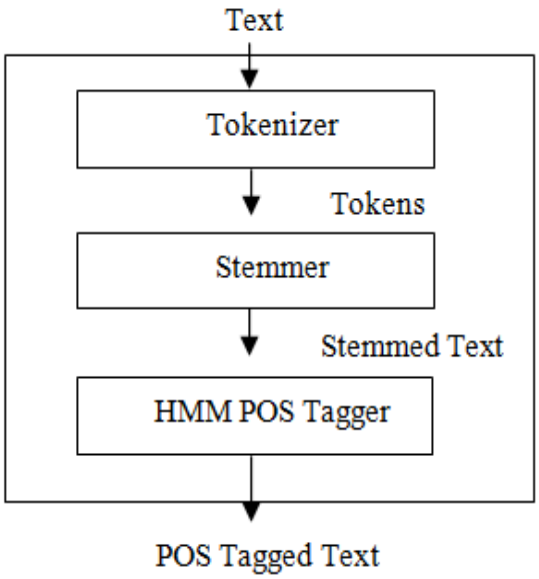


Fig.1. Transition of Sentences

For tagging the segments, firstly take the sentences as an input. Thenceforth, tokenizing the sentence and use any model for the POS tagging. Here, we have used Hidden Markov Model (HMM) for tag the input sentences.

Hidden Markov Model

HMM [17] is a generative model. It is finite set of states, which allocate the joint probability distribution over corresponding observation and tag progression. In the POS tagger based text on HMM sequentially allocate the forward and backward probabilities. The succeeding rule defines the HMM:

$$P(t_i/w_i) = P(t_i/t_{i-1}).P(t_{i+1}/t_i).P(w_i/t_i)$$

For calculating the probabilities use the following formula:

$$P(t_i/t_{i-1}) = \frac{freq(t_{i-1},t_i)}{freq(t_{i-1})}$$

Example: 5. इलाहबाद में गंगा , यमुना और सरस्वती
NNP PSP NNP PUNC NNP CCD NNP
तीनो नदियों का संगम है ।
NNP NN PSP NN VAUX PUNC

Table 3. IL POS Tagset

S. No.	Tag	Description of Tag	Examples
1.	NN	Common Noun	महिला, अंगुठी, किताब
2.	NST	Noun Denoting Spatial and temporal expressions	नीचे, बाहर, ऊपर
3.	NNP	Proper Nouns	सन्तोष, प्रीती, अशोक, मोहन
4.	PRP	Pronoun	तुम, आप, उसे
5.	DEM	Demonstratives	यह, वह
6.	VM	Verb Main	पढ़ना, रोना, दौडना, देखना
7.	VAUX	Verb Auxiliary	रहा, है, हुए
8.	JJ	Adjective	छोटा, लम्बा, पतला, सुंदर
9.	RB	Adverb	तेज, जल्दी, बुरा
10.	PSP	Postposition	से, का, की, ने, को, में
11.	RP	Particle	भी, सा, ही, जी
12.	CC	Conjuncts (co-ordinating or subordinating)	और, या
13.	WQ	Question Words	क्या, किसने, कब, कैसे
14.	QF	Quantifiers	कम, ज्यादा, थोडा,
15.	QC	Cardinals	दो, तीन, चार, सात
16.	QO	Ordinals	तीसरा, पहला,
17.	CL	Classifier	कितने, दो-तीन, लोग
18.	INTF	Intensifier	बहुत, कम
19.	INJ	Interjection	अरे, हाय,
20.	NEG	Negative	ना. नही
21.	UT	Quotative	बोला, कहा
22.	SYM	Special Symbols	$, %,*, +
23.	*C	Compounds	चलना- फिरना, खाना- पीना
24.	RDP	Reduplication	तेज- तेज, हल्का- हल्का
25.	ECH	Echo Words	नीले- पीले, हल्का- फुल्का
26.	UNK	Unknown	English, বাংলা, অসমীয়া

B. Centering Based Algorithm

Centering algorithm[20] is focused on single entity or focused on any given context which is determine from all other entities that have been extract. A discourse model is separate into U_1……..U_m. The two essential representations in the discourse model are U_n and U_{n+1} to be adjacent utterances. In the centering algorithm has following essentials:

- Utterance: Discourse is separate into single sentences or set of sentences, clauses is known as utterances.
- Center: In the discourse focus or referred any entity.
- Forward Looking Centers: $C_f(U_n)$, defines the ordered list that contains exactly realized linguistic expressions in the utterance.
- Backward Looking Centers: $C_b(U_n)$, defines after the utterance U_n is interpreted, the entity that currently focused. $C_b(U_{n+1})$ is the most highly ranked element of the $C_f(U_n)$.
- Preferred Center: $C_p(U_n)$ is the utmost ranked forward looking center of U_n. Preferred center is the first entity of the list.

In this algorithm, defines following transitional relationships between U_n and U_{n+1} on the relation between $C_b(U_{n+1})$, $C_b(U_n)$ and $C_p(U_{n+1})$:

Table2.Transitions in the Centering algorithm

	$C_b(U_{n+1}) = C_b(U_n)$ or undefined $C_b(U_n)$	$Cb(Un+1) = C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) = C_p(U_{n+1})$	Retain	Rough-Shift

The following rules are using the Centering algorithm:

- If any component of $C_f(U_n)$ is get by a pronoun in U_{n+1}, then $C_b(U_{n+1})$ must be getting as a pronoun also.
- Following is order of the transitional state relationships: Continue > Retain > Smooth-Shift > Rough-Shift.

Example: 6(a). हिमांशु बहुत अच्छा लड़का है।(U_1)
C_b: Undefined
C_f: हिमांशु, लड़का

(b). उसका दोस्त सौरभ है।(U_2)
C_b: हिमांशु
C_f: हिमांशु, दोस्त, सौरभ
उसका = हिमांशु
Continuing
(c). वह सबकी सहायता करता है।(U_3)
C_b: हिमांशु
C_f: हिमांशु, सौरभ, सहायता
वह = हिमांशु
Continuing

Using centering based algorithm, for the referent pronoun in U3 there are two choices 'वह' Himanshu and Saurabh. Since Saurabh is recently introduced entity, so choosing referent of 'वह' as Saurabh is 'Smooth-Shift', while since Himanshu is already referred in U2 as use of pronoun 'उसका' choosing it as referent of 'वह' will result in 'Continue' transition. Thus 'Continue' is endorsed over 'Smooth-Shift' as in first scenario, so Himanshu should be selected for the correct referent of 'वह'.

C. Working of the Proposed System

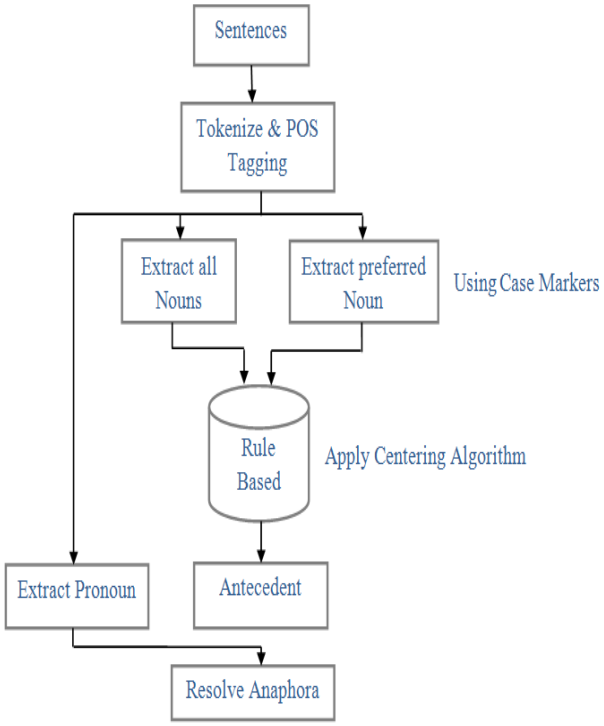


Fig 2. Proposed Model of the System

Initially, the system tag data and excerpt entire pronouns and nouns of the given sentences. After that, it observed the antecedent on the behalf of case markers for referencing expression. The prior noun is preferred as an antecedent and verified from the defined rules of the centering algorithm. Resolving the anaphora system performs the task in subsequent manner:

Proposed Algorithm:

Step 1.Input: Input a sentence
Step 2. Tokenize and tagged the sentences
Step 3. For i=U_1 to U_m
i. Extract nouns & pronouns
ii. Apply case markers to set the preferences of the noun
if ($C_b(U_{n+1})$==$C_b(U_n)$ & $C_b(U_{n+1})$==$C_p(U_n)$)
pro=$C_b(U_{n+1})$
elif($C_b(U_{n+1})$==$C_p(U_{n+1})$)
pro=$C_b(U_{n+1})$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

elif($C_b(U_{n+1})$==$C_b(U_n)$)
pro=$C_b(U_{n+1})$
Step 4. Print "referent of the pronoun="+pro

V. EVALUATION

The aim of this evaluation is to check the performance of the proposed model; as accuracy is directly proportional to the performance of the system. Accuracy of a system depends on the number of rules that are developed. We used 200 segments for evaluation of our system. To measure the accuracy following formula is used:

$$\text{Accuracy}(\%) = \frac{\text{Correct Resolution}}{\text{Total segments}} \times 100$$

Among 500 segments, 405 segments where correctly resolved for noun-pronoun combination. This gave an accuracy of 81% for our system. The system was not able to correctly identify the segments which had similar nouns had ambiguity were system was unable to identify correct noun.

Example: 7. सीता और गीता अच्छे दोस्त है। वो उसको पसंद करती है।

Consequences of Example 1 define that 'वो' and 'उसको' were not able to identify which noun is being referred.

Example: 8. नीलम को पढ़ना बहुत पसंद है। वह प्रतिदिन विद्यालय जाती है।

In the above Example 2, 'वह' is referred to 'नीलम' which is easily identified by our system.

VI. CONCLUSION

In this paper, we have shown an approach of developing a rule based anaphora resolution system for Hindi. For development of this system, we have modified the famous centering algorithm. We have used a POS tagger before applying the centering algorithm. We also evaluated our system for 500 segments which had nouns in one sentence and pronouns in another sentence. Out of these 500 segments, our system was correctly identified 405 segment.

REFERENCES

1. A. Ashima, K. Sukhnandan and C. M. Rajni, "Anaphora Resolution in Hindi: A Hybrid Approach", International Symposium on Intelligent Systems Technologies and Applications, Vol. 530, pp. 815-830, 2016.
2. P. Rashmi and S. Michael, "Discourse Salience and Pronoun Resolution in Hindi", U. Penn Working Papers in Linguistics, Vol. 6.3, pp. 195-214, 2000.
3. G. Deepa and C. Niladri, "Identification of Divergence for English to Hindi EBMT", Proceeding of MT Summit-IX, 2003.
4. M. Rajni, "Anaphora Resolution in Hindi: Issues and Directions", Indian Journal of Science and Technology, Vol. 9, 2016.
5. D. Praveen, S. Himanshu and S. M. Dipti, "Anaphora Annotation in Hindi Dependency Treebank", Pacific Asia Conference on Language, Information and Computation, pp. 391-400, 2012.
6. P. Pravin and D.L. Sobha"Anaphora Resolution Algorithm for Sanskrit", International Sanskrit Computational Linguistics Stmposium, Vol. 6465, pp. 209-217, 2010.
7. L. Sobha and B. N. Patnaik, "Vasisth: An Anaphora Resolution System for Indian Languages", International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications, 2000.
8. M. Vandan, G. Palash and S. M. Dipti, "Pronominal Reference Type Identification and Event Anaphora Resolution for Hindi", International Journal of Computational Linguistics and Applications, Vol. 7, pp. 45-63, 2016.
9. S. Apurbalal and G. Utpal, "Guitar-based Pronominal Anaphora Resolution in Bengali", Annual Meeting of the Association for Computational Linguistics, pp. 126-130, 2013.
10. Y. S. Daya, D. Kamlesh, S. Pardeep and C. Preetika, "Anaphora Resolution for Indian Languages: The State of the Art", National Conference on Recent Innovations in Science and Engineering, pp. 1-7, 2016.
11. S. K. Utapal, E. Asif, S. Sriparna, U. Olga and P. Massimo, "Differential Evolution-Based Feature Selection Technique for Anaphora Resolution", Soft Computing, pp. 2150-216, 12015.
12. L. Priya and S. Smita,"Anaphora Resolution in Hindi Language", International Journal of Information and Computation Technology, Vol. 3, pp. 609-616, 2013.
13. A. Sachin, S. Manaj, A. Pallavi and S. Ratna,"Anaphora Resolution in Hindi Documents", International Conference on Natural Language Processing and Knowledge Engineering, pp. 452-458, 2007.
14. C. Sanjay, D. Arnab, B. Biswanath, PK. Moumita, S. Sudeshma and B. Anupam"Anaphora resolution for Bengali, Hindi and Tamil Using Random Tree Algorithm in Weka", International Conference on Natural Language Processing, pp. 8-12, 2011.
15. B. Samuel, P. Massimo, P. P. Simone, R. J. Kepa, R. Lorenza, U. Olga, V. Yannick and Z. Roberto, "BART: A multilingual anaphora resolution system", International Workshop on Semantic Evaluation, ACL, pp. 104-107, 2010.
16. P. L. Triveni, D. Kamlesh and S. Pardeep, "Anaphora resolution in Hindi: Issues and challenges", International Journal of Computer Applications, Vol. 42, pp. 7-13, 2012
17. J. Nisheeth, D. Hemant and M. Iti, "HMM based POS tagger for Hindi.", International Conference on Artificial Intelligence, Soft Computin, pp. 341-349, 2013.
18. W. Kazi and I. Sabir, "Recommendation System for Bangla News Article with Anaphora Resolution.", 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT). IEEE, pp. 467-472, 2018.
19. S. BiBi, K. Khairullah, K. Aurangzeb, S. Fazali, K. Wahab, and A. Muhammad "Impact of Anaphora Resolution on Opinion Target Identification.", International Journal of Advanced Computer Science and Applications, Vol. 9, pp. 230-236, 2018.
20. J. G. Barbara, K. J. Arvind and W. Scott, "Centering: A Framework for Modeling the Local Coherence of Discourse", Association for Computational Linguistics, Vol. 21, pp. 203-225, 1995.
21. S. Raju, S. P. Bipul and S. Dihren, "Part of Speech Tagging in Manipuri with Hidden Markov Model", International Journal of Computer Science Issues, Vol. 9, pp. 146-149, 2012.
22. D. Sandipan and S. Sudeshna, "Part of Speech Tagging for Bengali with Hidden Markov Model", NLPA Machine Learning Competition, 2006.
23. F. W. Alfan and P. Ayu, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia", International MALINDO Workshop, 2010.
24. P. Abhijit, S. P. Bipul and S. Sunita, "Hidden Markov Model Based Part of Speech Tagging for Nepali Language", International Symposium on Advanced Computing and Communication, 2015.

AUTHORS PROFILE

Varsha Singh is a student at Banasthali Vidyapith. Her areas of interest are natural language processing and machine translation.

Nisheeth Joshi is an Associate Professor at Bnasthali Vidyapith, India. He primiraliy works in Machine Translation, Information Retrieval, Cognitive Computing. He has over 12 years of teaching experimence.