Pronominal Anaphora Resolution Algorithm in Myanmar Text

May Thu Naing, Aye Thida

Abstract— One way to increase the coherence of summary generation system is to derive first the discourse structure of the text and to guide the selection of the sentences. These sentences to be included into the summary are considered by a score according to both the relevance of the sentence in a discourse tree and the coherence of the text, as given by solving anaphoric references. Anaphoric references resolution is a common phenomenon in natural language processing, and has correspondingly received a significant amount of attention in the literature. In this paper, we present a Myanmar pronominal anaphora resolution algorithm which bases on Hobbs' algorithm that works only the surface syntax of sentences in a given text. The purpose of this paper is to implement pronominal anaphora resolution system on computer in java. This study shows that Myanmar anaphora resolution algorithm perform on two kinds of data sets and corresponding results are obtained a substantial accuracy 80% that can be an acceptable resolution performance for Myanmar. In addition, this pronominal anaphora resolution algorithm will be used for summary generation system in Myanmar.

Index Terms—**Anaphora, Pronominal Resolution, Natural Language Processing, Myanmar Pronouns**

I. INTRODUCTION

Anaphora resolution is vital for areas such as machine translation, summarization, and question-answering system and so on. Many of automatic text summarization systems apply a scoring mechanism to identify the most salient sentences. However, the task result is not always guaranteed to be coherent with each other. It could lead to errors if the selected sentence contains anaphoric expressions. To improve the accuracy of extracting important sentences, it is essential to solve the problem of anaphoric references in advance. Anaphoric dependence is a relation between two linguistic expressions such that the interpretation of one, called anaphora is dependent on the interpretation of the other, called antecedent. The problem of anaphora resolution is to find the antecedent(s) for every anaphora.

Anaphora resolution is classically recognized as a very difficult problem in Natural Language Processing Work on anaphora resolution in the open literature tends to fall into three domains: artificial intelligence (as a specialty of computer science, including computational linguistics and natural language processing), classical linguistics (as distinguished from computational linguistics), and cognitive psychology. For our purposes, we are primarily interested in the artificial intelligence linguistics approach [1]. We will only be concerned with computational approaches to pronominal anaphora resolution algorithm that have been implemented on a computer in Java.

The remaining parts of the paper are organized as follows: the relevant works for pronominal anaphora resolution in natural language processing are presented in section 2, about Myanmar language introduced in section 3, section 4 describes Hobbs' algorithm as a basic algorithm, section 5 proposed Myanmar pronominal anaphora resolution algorithm, section 6 explain the performance and data analysis using proposed algorithm and section 7 concludes the paper and identifies future work.

***May Thu Naing**, Ph.D Candidate, University of Computer Studies, Mandalay, Myanmar.*
***Aye Thida**, Reasarch and Development Department, University of Computer Studies, Mandalay, Myanmar.*

II. RELATED WORK

Many approaches on anaphora resolution syntax have been used as an important feature. Some well-known syntax based approaches include Hobbs algorithm [2] and the Centering approach [3]. Various rule based and data driven approaches have been proposed which use syntactic information as an important feature.

Traditionally, anaphora resolution systems rely on syntactic, semantic or pragmatic clues to identify the antecedent of an anaphor. Hobbs' algorithm [2] is the first syntax-oriented method presented in this research domain. From the result of syntactic tree, they check the number and gender agreement between antecedent candidates and a specified pronoun. [4] is one of the most important approach for anaphora resolution in Hindi. They applied a discourse salience ranking to two pronoun resolution algorithms, the BFP and the S-List algorithm. In [5], an algorithm called Anaphora Matcher (AM) is implemented to handle inter-sentential anaphora over a two-sentence context. A statistical approach was introduced by [6], in which the corpus information was used to disambiguate pronouns. A knowledge-poor approach is proposed by [7], it can also be applied to different languages (English, Polish, and Arabic). The main components of this method are so called "antecedent indicators" which are used for assigning scores (2, 1, 0, -1) against each candidate noun phrases. But there does not have any anaphora resolution system in Myanmar.

Therefore, the aim of this paper is to implement a system that is based on Hobbs' algorithm for pronominal anaphora in Myanmar. A syntactic rule based algorithm is run on manually parsed sentences. Hobbs tested his algorithm for the pronouns he, she, it. The algorithm is adapted successfully for those languages (eg. Chinese), which have similar Subject-Verb-Object (SVO) structure and follow a fixed word order. Myanmar language is a free words order. It has to inherent difficulties for the application of Hobbs'

 All Rights Reserved © 2014 IJARCET

algorithm. To apply the syntactic anaphora resolution Hobbs' algorithm for Myanmar texts, the Hobbs' algorithm must be modified to find antecedents for pronouns. Therefore, this paper describes the anaphora resolution algorithm in Myanmar text.

III. ANALYSIS OF MYANMAR LANGUAGE

The Myanmar Language is the official language of Myanmar. It is also the native language of the Myanmar and related sub-ethnic groups of the Myanmar, as well as that of some ethnic minorities in Myanmar like the Mon. Myanmar language is a tonal and pitch-register, largely monosyllabic and analytic language, with a Subject Object Verb (SOV) word order. The language uses the Myanmar script, derived from the Old Mon script and ultimately from the Brahmi script.

The language is classified into two categories. One is formal, used in literary works, official publications, radio bradcasts and formal speeches. The other is colloquial, used in daily conversation and spoken. This is reflected in the Myanmar words for "language": □□ sa refers to written, literary language, and □□□□ sa-kar refers to spoken language. Therfore, Myanmar language can mean either "maran-ma-sa" (written Myanmar language), or "mran-ma-sa-ka:" (spoke Myanmar language). Much of the differences between formal and colloquial Myanmar language occur in grammatical particles and lexical items [8]. Different particles (to modify nouns and verbs) are used in the literary form from those used in the spoken form. For example, the postposition after nouns is □ hnai: at in formal Myanmar language and □□ hma: at in colloquial Myanmar language. In this study, we focuse on written Myanmar language.

Example:

□□□□ □□□□□ □□□ □□□ Mother is at home. **(Formal form)**

□□□□□□□□□ □□□ Mother is at home. **(Spoken form)**

A. Myanmar Earley Parser

In this study, we use the parse trees of sentences as inputs to resolve the anaphora. Therefore, Myanmar Earley parser [9] is applied. The Earley parser is an algorithm for parsing strings that belong to a given context-free language. uses dynamic programming; it is mainly used for parsing in computational linguistics.

As grammar and lexical rules CFGs in chunk for Myanmar sentence "□□ □□□ □□□□□□ □□□ . □□□□ □□ □" (He goes to school) are shown in following.

```
0:[S]
  0:[NP_NOM]
    0:[NP]
      0:[PRON_PERSON]
        > 0:[□□ ]
    0:[PREP_NOM]
      > 0:[□ □□ ]
  0:[NP_ARRIVAL]
    0:[NOUN_PLACE]
      > 0:[□ □□□□□ ]
    0:[PREP_ARRIVAL]
      > 0:[□□□ ]
  0:[VEND]
    0:[V]
      > 0:[□□□□ □□ ]
```

B. Myanmar Earley Parser

Anaphoric reference type can be classified into abstract (event) references where an anaphora refers to an event or a proposition and concrete (entity) references where it refers to a concrete entity like noun phrase (person, place etc) quantifiers etc. In this work, we focus on resolving only entity pronouns; hence the mention detection or anaphoric determination step is not required for our system.

Pronominal anaphora are the most commonly encountered in general usage. Myanmar language has four categories. There are

1. Personal pronoun
2. Demonstrative pronoun
3. Question pronoun
4. Mathematic pronoun

In above pronouns, question and mathematic pronoun do not need for summary generation. Therefore, we ignore the resolving these two types of pronoun. Pronominal anaphora in English and Myanmar are shown in Table 1.

Table 1. Pronominal anaphora in English and Myanmar

Type of Pronoun		Anaphora in English	Anaphora in Myanmar
Personal Pronoun	Nominative	he, she, it, they	he(*thu*), she(*thuma*), they(*thutoe*)
	Reflexive	himself, herself, itself, themselves,	himself (*thu kotai*), herself (*thuma kotai*), themselves (*thutoe kotai*),
	Possessive	his, her, its, their	his(*thu ei*), her(*thuma ei*), their(*thutoe ei*),
	Objective	him, her, it, them	him(*thu ko*), her(*thuma ko*), them(*thutoeko*)
Demonstrative Pronoun		this, that, these, those, others	ei, the, hto, yin, le' gaung

In Myanmar language, the usage of "it" does not the same of English. Therefore, we consider to resolve personal pronouns without include "it, itself, its" in Myanmar. The resolving of demonstrative pronoun will present in the future study.

Most algorithms in the literature resolve the pronouns 'he', 'she', 'it', 'her', 'him', 'his', 'her', 'its' in English whenever they have an antecedent which is a noun phrase. In this study, we will narrow down the scope of anaphoric phenomena and focus on a sub-problem of anaphora resolution. Therefore, our pronoun resolution system will resolve all four types of

 All Rights Reserved © 2014 IJARCET

personal pronouns except of "it" whenever they have an antecedent which is a noun phrase.

IV. HOBBS' ALGORITHM FOR ANAPHORA RESOLUTION

Hobbs' presents two algorithms of pronominal anaphora resolution. The first approach is a naïve algorithm that works by traversing the surface parse trees of the sentences of the text in a particular order looking for noun phrases of the correct gender and number. In the second approach, it is shown how pronoun resolution can be handled in comprehensive system for semantic analysis of English texts. [2]. We will concentrate on the Naïve algorithm for finding antecedents of pronouns here.

The algorithm assumes that the data is presented in the format of parse trees produced by a particular grammar namely, the one where a Noun Phrase (NP) node dominates an Noun (N-bar) node, to which arguments of the head noun attach. The algorithm traverses the tree, from the pronoun up, stopping on certain S, NP and Verb Phrase (VP) nodes, searching left-to-right breadth-first in the subtrees dominated by these nodes.

It will be necessary to assume that an NP node has an N-bar node below it, as proposed by [13], to which a prepositional phrase containing an argument of the head noun may be attached. Truly adjunctive prepositional phrases are attached to the NP node in English. This assumption, or something equivalent to it, is necessary to distinguish between sentences (1) and (2) in English [5]. It is worth noting that where English has a prepositional phrase we use an NP which has a locative case in Myanmar.

(E1) Mr. Myo$_i$ saw a driver$_k$ in his$_{i,k}$ truck.
(E2) Mr. Myo$_i$ saw a driver of his$_i$ truck.

In sentence (E1) 'his' may refer to Mr. Smith or the driver, but in sentence (E2) it may not refer to the driver. The structures for the relevant noun phrases in sentences (E1) and (E2) are shown in Fig 1.

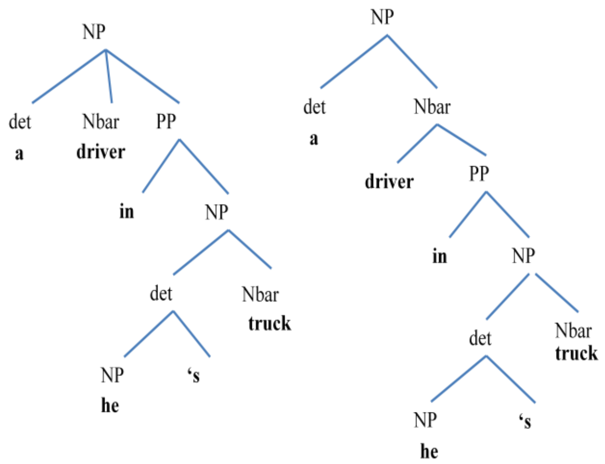

Fig 1. The structures for NPs of sentences (E1) and (E2)

We translate sentence (E1) from English to Myanmar in two possible forms as indicated in sentences (M1) and (M2) in Table 2 and Table 3.

Table 2. Translation of English sentence (E1) to Myanmar sentence (M1)

M1	Mr. Myo saw a driver in his truck.
	[illegible]
	Mr. Myo the driver ko thu ei truck htae twin myin kae the.
	Mr. Myo the / Mr. Myo –NOM driver ko / driver –ACC thu ei / his –PRON-INDEX truck htae twin / truck –POSS-LOC myin kae the / see –PAST

In sentence (M1) from Table 2 "thu ei"(his) may be co-referential with "Mr. Myo" or driver or another person in the previous sentences. The syntactic tree structure for the sentence (M1) is shown in Fig 2.

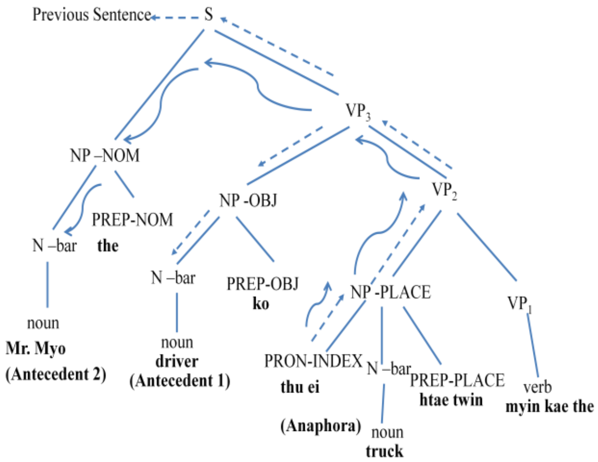

Fig 2. The structures of sentence M1 and the algorithm working on it

The path from thu ei (Anaphora) to driver (Antecedent 1) is PRON –INDEX (Anaphora) –NP3 –VP2 –VP3 –NP2 –Nbar(Antecedent 1, driver).

The path from thu ei (Anaphora) to Mr. Myo (Antecendent 2) is PRON –INDEX (Anaphora) –NP3 –VP2 – VP3 –S –NP1 –Nbar (Antecendent 2, Mr. Myo).

Table 3. Translation of English sentence (E1) to Myanmar sentence (M2)

M2	Mr. Myo saw a driver in his truck.
	[illegible]
	Mr. Myo the thu ei truck htae twin driver ko myin kae the.
	Mr. Myo the / Mr. Myo –NOM thu ei / his –PRON-INDEX truck htae twin / truck –POSS-LOC driver ko / driver –ACC myin kae the / see –PAST

In Myanmar language, reflexive and possessive anaphors are subject oriented. The algorithm traverses from left to right, therefore, driver (noun) is not interpreted as an antecedent in the sentence (M2) from Table 3 in the following Fig 3.

 All Rights Reserved © 2014 IJARCET

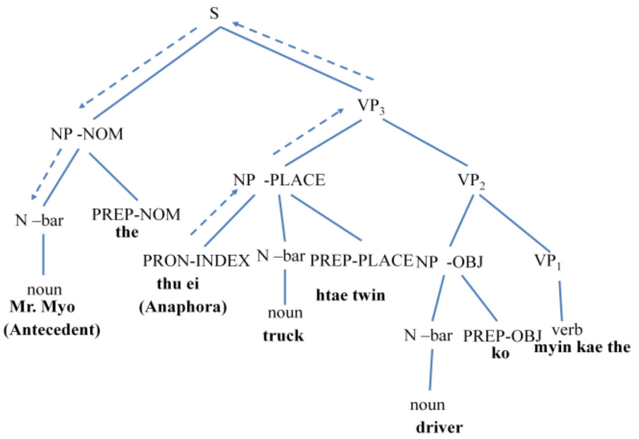

Fig 3. The structures of sentence (M2) and the algorithm working on it

The path from thu ei (Anaphora) to Mr. Myo (Antecedent) is PRON –INDEX (Anaphora) –NP2 –VP3 –S –NP1 –Nbar (Antecedent, Mr. Myo).

V. MYANMAR PRONOMINAL ANAPHORA RESOLUTION ALGORITHM

With nominative subjects reflexive pronouns (“anaphors”) and possessive pronouns (“pronominal”) are in complimentary distribution when it comes to expressing relation. Anaphors are able to find the antecedent in a local domain. Possessive pronouns look for antecedent farther. Nominative case is an absolute criterion for subject status in English. But, the role of subject and object in Myanmar are found to have significant impact on anaphora resolution .Most algorithms in the literature resolve the pronouns ‘he’, ‘she’, ‘it’, ‘her’, ‘him’, ‘his’, ‘her’ and ‘its’ in English. However, the Myanmar pronominal anaphora resolution algorithm that is based on Hobbs’ algorithm can resolve all personal pronouns that include in “they”, all possessive pronouns and all reflexive pronouns in Myanmar texts. The following algorithm Fig 5 is shown how to resolve anaphora in Myanmar.

Begin

Input: Parse tree of each sentence in Paragraph
Output: Pronoun Resolution
Step 1: Start with NP node of the last parse tree which includes in pronoun
NP, Pronoun € NP;
Step 2: Go up the tree
If (NP is found) then X:= NP;
else if (VP is found) then X:=VP;
else if (highest S is found) then
{ X:=S;
Go to Step 6.}
Step 3: If (X is NP) then Call *funAnti(X)*;
Step 4: If (X is VP) then Call *funAnti(X)*;
Step 5: Go to Step 2.
Step 6: Call *funAnti(X)*;
Step 7: Go to previous parse tree.
X:=Root node of previous parse tee;
Call *funAnti(X).*
If (X is VP) then Go to Step 4.
If (X is NP) then Go to Step 3.

End

funAnti(X)

(a) Do BFS under X.
(b) If (Noun in NP –NOM or Noun in NP –OBJ is found) then
Anti:= Noun Under NP –NOM or NP –OBJ
Else Continue on BFS.

Where,

NP =Noun Phrase
X =variable for node
VP =Verb Phrase.
BFS =Bread first search
NP –NOM =Noun phrase of Nominative
NP –OBJ =Noun phrase of Object
Anti =variable for antecedent

Fig 5. The pronominal anaphora resolution algorithm for Myanmar

Fig 6 and Fig 7 illustrate this algorithm working on the sentences (M3) and (M4) which the translation of the sentence (E2) and (E3) from English to Myanmar for determining the antecedents of each anaphora.

(E2) Ma Ma goes to school.
(E3) She goes by car.
(M3) Ma Ma the school thoe thwar the.
(M4) thuma the car pyint thwar the.

Input: Two parse trees
Output: Pronoun Resolution
Step 1: Start from PRON (***thuma***) in S1
NP, PRON € NP –NOM
Step 2: Go up the tree
X:=S
Go to Step 6.
Step 6: Call *funAnti(X)*
It does not perform any steps in *funAnti(X)*
Step 7: Go to previous tree (S2) of S1.
X:=S1;
Call *funAnti(X).*
1: Do BFS under S2.
2: If (Noun in NP –NOM is found) then
Anti:=***Ma Ma***.

Therefore, according to pronominal resolution algorithm, Ma Ma is antecedent of thuma (anaphora).

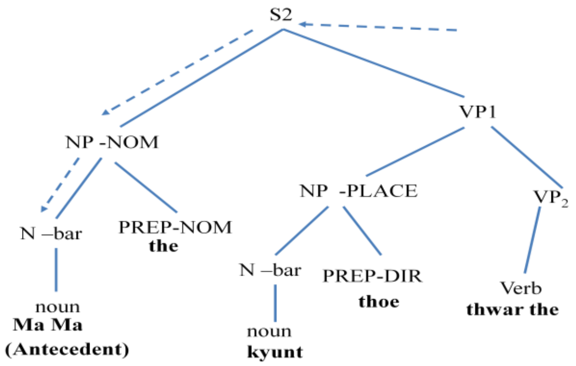

Fig 6. The illustration of the parse tree of sentence (M3), the algorithm working on it and the determination of the antecedent of anaphora ‘thuma’ from S1.

 All Rights Reserved © 2014 IJARCET

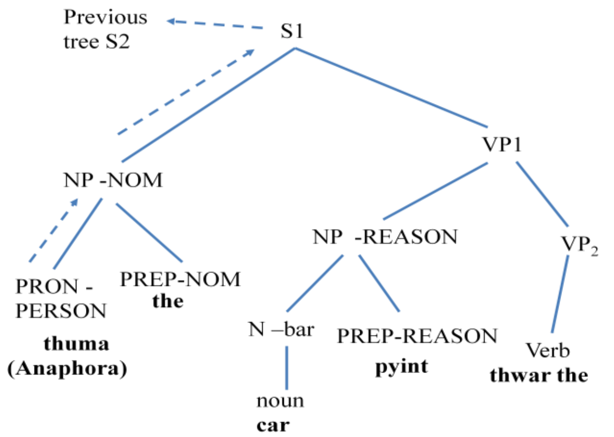

Fig 7. The illustration of the parse tree of sentence (M4), the algorithm working on it.

VI. EXPERIMENTAL RESULT

We have performed two different types of data sets using our proposed anaphora resolution (AR) algorithm. The first data set is short stories in Myanmar. This experiment represents a baseline performance since the story is a straightforward narrative style with extremely low sentence structure complexity. We have taken short stories in Myanmar language from basic Myanmar [11]. Another data set has been taken from basic Myanmar Essays [10]. The comparison of applied two data sets in Myanmar pronominal anaphora resolution system shows in Table 4.

Table 4. Comparison of two data sets

Data Set	Sentences	Words
Short Stories	269	3497
Essays	192	3072

The measurement for success rate on two data sets can be calculated as follows:

$$\text{Success Rate} = \frac{\text{number of correctly resolved anaphors}}{\text{Number of all anaphors}}$$

Table 5 and 6 show the evaluation result of our anaphora resolution algorithm which applies on two different data sets.

Table 5. Accuracy result on Short Stories data set

Type of Pronoun	Correctly resolved	Anaphora to resolve	Accuracy
Nominative	60	69	0.86
Objective	10	15	0.67
Possessive	35	42	0.83
Reflexive	10	15	0.67
Overall	***115***	***141***	0.***82***

Table 6. Accuracy result on Basic Essays data set

Type of Pronoun	Correctly resolved	Anaphora to resolve	Accuracy
Nominative	81	90	0.90
Objective	10	15	0.67
Possessive	60	66	0.90
Reflexive	5	8	0.63
Overall	***156***	***179***	***0.87***

VII. DISCUSSION AND DATA ANALYSIS

Table 5 and 6 present results for two sets of data, i.e short stories and essays. Hobbs tested his algorithm for the pronouns he, she, it and they, successfully 81.8% in English. The accuracy of theirs algorithm in Chinese [12] has been reported to be 77.6%. The overall accuracy of pronominal anaphora algorithm for Myanmar is greater than 80%. We have presented the result on two data sets which contains texts from various domains with average size of 20 sentences. The accuracy of algorithm on both data sets for nominative and possessive pronouns is relatively greater than of other types of pronouns: objective and reflexive pronouns. Fig 8 shows the success rate of personal pronoun on both data sets.

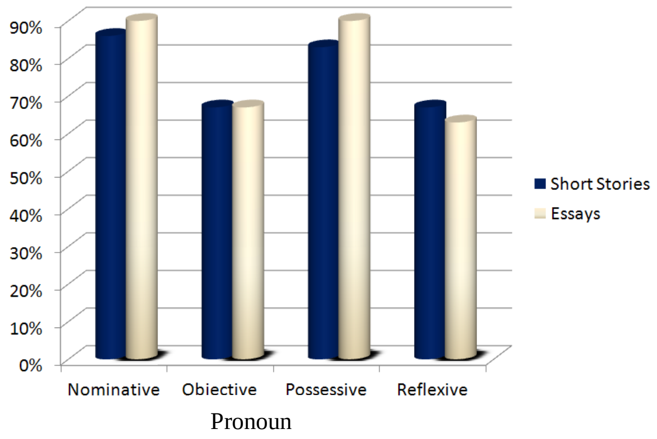

Fig 8. The comparison of success rate of resolving pronoun on two data sets

VIII. CONCLUSION

The purpose of the present work is to implement a syntactic anaphora resolution system that could be used as a baseline to Myanmar anaphora resolution. This paper presents the implementation of pronominal anaphora resolution algorithm for Myanmar by taking into account the free word order and grammatical role in pronoun resolution in Myanmar. The role of subject and object in Myanmar are found to have significant impact on anaphora resolution for reflexive and possessive pronouns. This proposed algorithm has tested for limited set of sentences depend on Earely parser. This algorithm has some limitations. It does not work with not fully parsed sentence. Therefore, the future work will be directed to resolve not only all personal personals but also demonstrative pronouns. And then, we aim to the development of anaphoric fully parsed corpus as the future work.

 All Rights Reserved © 2014 IJARCET

APPENDIX

Some abbreviations used in Table 2, Figures 1-7

S = sentence
NP=–Noun Phrase
NP –NOM =Noun Phrase of Nominative
NP –PLACE=– Noun Phrase of Place
NP –OBJ =Noun Phrase of Object
VP =Verb phrase
V =Verb
N –bar = Noun
PREP –NOM = Preposition of nominative
PREP –OBJ =Preposition of object
PREP –CAU =Preposition of cause
PRON =Pronoun
PRON –INDEX =Pronoun index
PP =Preposition Phrase
det =determinator

REFERENCES

[1] Tufekci, P. and Kilicaslan Y, "A Computational Model for Resolving Pronominal Anaphora in Turkish Using Hobbs' Naïve Algorithm", Trakya University, 2006.
[2] Jerry Hobbs, "Resolving pronoun references. In Readings in natural language processing", Morgan Kaufmann Publishers Inc, 1986.
[3] Susan, E., Marilyn, W, and Carl J Pollard, "A centering approach to pronouns", In Proceedings of 25th ACL, 1987.
[4] Rashmi Prasad and Michael Strube, "Discourse salience and pronoun resolution in hindi", U. Penn Working Papers in Linguistics, 2000.
[5] Denber, Michel., "Automatic resolution of anaphora in English," Technical report, Eastman Kodak Co, 1998.
[6] Dagan, Ido and Alon Itai, "Automatic processing of large corpora for the resolution of anaphora references," In Proceedings of the 13th International Conference on Computational Linguistics (COLING'90), Vol. III, 1990.
[7] Mitkov, Ruslan, Richard Evans and Constantin Orasan, "A new fully automatic version of Mitkov's knowledge-poor pronoun resolution method," In Proceedings of CICLing- 2000, Mexico City, Mexico.
[8] Thet Thet Zin, Khin Mar Soe, and Ni Lar Thein "Myanmar Phrases Translation Model with Morphological Analysis of Statistical Myanmar to English Translation System", University of Computer Studies, Yangon, Myanmar.
[9] Soe Lai Phyue, 2012 "Lexical Analyzer for Myanmar Language", in Proceedings of the 10th International Conference on Computer Applications (ICCA2012), Yangon, Myanmar,
[10] Thar Ra, P, "Basic Primary Myanmar Essays", 2004.
[11] Min Thu, W, "Stories for Babies", Myanmar.
[12] Converse,S.P, "Resolving Pronominal References in Chinese with the Hobbs algorithm", Proceedings of the Fourth SGHAN Workshop on Chinese Language Processing, 2005.
[13] Robert C.B, Paul., Beracach Y.,Chomsky N., "Poverty of the Stimulus Revisited", 2011.

May Thu Naing
She received the B.C.Sc. (Hons) degree from Computer University University (Mandalay) in 2008 and M.C.Sc degree from Computer University (Taunggyi) in 2010, respectively. Now, she is a PhD candidate at University of Computer Studies, Mandalay. Her interested fields are data mining and natural language processing.

Aye Thida She is an Associate Professor of Research and Development Department at University of Computer Studies, Mandalay (UCSM), Myanmar. She was a leader of Natural Language Processing Project. Her team has developed Myanmar to English Translation System in 2011. Her research interests include Distributed Processing, Queuing and Natural Language Processing. She is currently working Myanmar to English Translation System Project. Dr. Aye Thida received B.Sc(Hons) Maths degree from the Mandalay University, Myanmar and her M.I.Sc and Ph.D degrees in Computer Science from the University of Computer Studies, Yangon(UCSY), Myanmar.

 All Rights Reserved © 2014 IJARCET