Syntactic Structured Framework for Resolving Reflexive Anaphora in Urdu Discourse Using Multilingual NLP

Jamal A. Nasir* and Zia Ud. Din
ICIT, Gomal University, D.I. Khan
KPK, Pakistan
[e-mail: jamalnasir@gu.edu.pk, ziasahib@gmail.com]
*Corresponding author: Jamal A. Nasir

Received November 17, 2020; revised February 17, 2021; accepted March 13, 2021; published April 30, 2021

Abstract

In wide-ranging information society, fast and easy access to information in language of one's choice is indispensable, which may be provided by using various multilingual Natural Language Processing (NLP) applications. Natural language text contains references among different language elements, called anaphoric links. Resolving anaphoric links is a key problem in NLP. Anaphora resolution is an essential part of NLP applications. Anaphoric links need to be properly interpreted for clear understanding of natural languages. For this purpose, a mechanism is desirable for the identification and resolution of these naturally occurring anaphoric links. In this paper, a framework based on Hobbs syntactic approach and a system developed by Lappin & Leass is proposed for resolution of reflexive anaphoric links, present in Urdu text documents. Generally, anaphora resolution process takes three main steps: identification of the anaphor, location of the candidate antecedent(s) and selection of the appropriate antecedent. The proposed framework is based on exploring the syntactic structure of reflexive anaphors to find out various features for constructing heuristic rules to develop an algorithm for resolving these anaphoric references. System takes Urdu text containing reflexive anaphors as input, and outputs Urdu text with resolved reflexive anaphoric links. Despite having scarcity of Urdu resources, our results are encouraging. The proposed framework can be utilized in multilingual NLP (m-NLP) applications.

Keywords: Natural Language Processing, Reflexive Anaphora Resolution, Linguistic Rules, Urdu Anaphora, Anaphor, Referent

1. Introduction

Humans use natural languages to communicate to and exchange thoughts and information with each other in many contexts in daily life. This communication contains links or references in between various units of text. Humans, having intelligence, can easily understand and interpret these references. Computers cannot interpret these naturally occurring references. The area of research concerned with understanding and generation of natural languages by computers is called natural language processing (NLP)**;** and anaphora resolution (AR) is the most inspiring area of NLP [1]. Anaphora is a Greek word and is the combination of words 'ana', and 'phora'. Word 'ana' is for back while 'phora' stands for carrying, implying anaphora as 'act of carrying backward'. It is a linkage between two language entities called anaphor or anaphoric device and referent or antecedent.

A text in any natural language may contain different types of anaphoric links, indispensable to be resolved. One of these types is reflexive anaphora. Resolution of reflexive anaphora is the central concept of this work. Translation and understanding of the natural languages require interpretation of these links. Hence, a natural language understanding system must have a mechanism for identifying and resolving these links between elements of a text. For the success of any NLP system - Information Extraction (IE), Sentiment Analysis (SA), Question Answering (QA), and Machine Translation (MT) - the comprehension of anaphor is important. Many algorithms for the resolution of anaphoric links are general, claimed to be applicable for any language whereas others make extensive use of linguistic features of a particular genre of language.

An account of representative work in the area of reflexive anaphora resolution is presented. K. Lata et al. [1] reported numerous methods for resolving anaphoric links by using various features. Their outcomes clearly demonstrated that the insights of trends are very supportive for the new researchers of NLP community. Some NLP applications have demonstrated that how AR is capable of improving the performance of that application. Research in anaphora resolution has not been so rapid as in other subfields of NLP because of its complexity. However, it is being used in various vital areas of NLP. Its use in SA is explored by Cambria [2] and Nithya [3]; in MT by Stojanovski [4]; in QA by Zhao [5] and; in IE by Ting et. al. [6]**.**

Urdu is an amalgamative language [7]. It has free word order and rampant pro drop [8]. A very little research work has been done in Urdu, especially in the area of anaphora resolution [9]. For translation and understanding of Urdu text available in the form of newspapers, magazines, historical novels, biographies etc., an NLP system is needed.

All natural languages of the world have diverse syntactic and semantic structure. The challenging task of anaphora resolution requires expertise in syntactic analysis, lexical analysis, discourse analysis, etc. However, in case of Urdu, major challenges in anaphora resolution are:

- Lack of standard datasets
- Lack of preprocessing tools
- Agglutinative nature of language
- Influence of cases

This research work contributes the fragment of reflexive anaphora resolution to an Urdu language understanding and processing system. In this work, a syntactic structure based approach is used to resolve reflexive anaphoric links in Urdu text.

2. Related Work

AR research work started in 1960s with various projects to develop systems to perform this task. These systems were designed to receive instruction in a natural language and identify the pronouns used. A system, with the name STUDENT, developed by Bobrow [10], for understanding of school level algebra lessons and another system, named SHRDLU, developed by Winograd [11], for instructing the robot to move objects around. In these systems, heuristic rules were designed for resolving anaphoric links in a limited domain.

Hobbs [12] proposed the first algorithm by using syntactic approach on the basis of linguistic knowledge. First prominent system, called RAP (Resolution of Anaphora Procedure), was developed by Lappin & Leass [13]. They assigned weights to potential candidates for identification of appropriate referent and final selection was made on the basis of syntax and morphology.

Most important work is done by Ruslan Mitkov for AR. He combined statistical and traditional linguistic methods for AR system in [14, 15]. He presented that various constraints and preferences are not reliable for AR [16]. He proposed various indicators to select appropriate referent out of possible candidates. He presented robust knowledge-poor approach in [17] and then improved and implemented it later in [18].

A very limited work has been done for AR in Urdu, due to unavailability of standard parsers and other preprocessing tools. Khan et al. [19] presented some important factors for resolution of pronominal anaphora. Kulsoom et al. [20] analyzed anaphora in Urdu and presented various grammatical structures of pronominal anaphora. Khan et al. [9] presented algorithm for distributive anaphora resolution by using syntactic approach.

Some valuable work for AR has been done in Asian languages, particularly in Hindi, which has almost the same syntactical structure as Urdu. Hindi is also called sister language of Urdu. R. Sinha. [21] presented a translation system, from English text to Hindi, called AnglaHindi. This translation system has issues while selecting the appropriate reflexive anaphors. Dutta [22] presented how to solve reflexive and possessive anaphors by using Hobbs algorithm and suggested further improvement subject to the availability of sufficient amount of data. Dakwale et al. [23] used dependency structure to resolve anaphoric links in Hindi text. Lalitha [24] presented a general anaphora engine for resource poor Indian languages by analyzing the similarities and variations between anaphors and their agreement with referent. Lakhmani et al. [25] developed a model, based on recency factor, for resolving pronominal anaphora using Gazetteer method.

K. Wohiduzzaman et al. [26] proposed an AR system for Bangla news articles. In this system, they presented to increase the keywords frequency by exploiting AR. Further, they developed a tagger for AR, that was capable to tag nouns, pronoun with POS, gender, status and number with different criteria. Secondly, they calculated term frequency, inverse document frequency (TF-IDF) for unique words similarity.

It is observed that considerable work has been done for anaphora resolution for many languages like English, Chinese, Hindi, Arabic, Bengali, etc., and several approaches have been used for this purpose. However, still there are many challenges that need to be overcome. One of the challenges is language dependency. Each language has its own syntax, structure, and arrangement of various language elements like nouns, pronouns, adverbs, etc. Each language treats number and gender feature differently. Therefore, a separate syntax analysis is required for each genre of language for NLP. Basic sentence format of Urdu is SOV. It has right to left (RTL) system of writing like Persian and Arabic. A limited number of studies are conducted for anaphor resolution due to the scarcity of Urdu language resources. Therefore, this work has been selected to promote and boost the work needed for Urdu.

3. Reflexive Pronouns/Anaphors in Urdu

In natural languages, grammatical role or function of a noun in sentences is specified by using postpositions or clitics, called noun cases. By changing clitics, function of a noun in the sentence changes. Reflexive anaphors refer to noun phrases and personal pronouns with one or more clitics, affecting the syntax and semantics of the text. In Urdu, postpositions or clitics are used with noun to specify its function in the sentence.

To locate the referent accurately in a discourse and making it translatable and understandable to other natural languages, it is important to deal with and manage the postpositions. There are two different opinions about the noun cases in Urdu: the first, that there are three noun cases [27]; and the second, that there are eight noun cases [28]. According to Butt and King [29], there are seven noun cases. We combine all these cases together, as there is no conflict among them [30]. These noun cases along with clitics are shown in **Table 1**.

Table 1. Noun Cases

Case	Clitics form	Morphological effect
Nominative	Nil	Nil
Oblique	Nil	Nominative/Modified form
Ergative	نے (ne)	Oblique form+ نے (ne)
Accusative	کو ko	Oblique form+ نے (ko)
Dative	کو ko, کے ke	Oblique form+ [کو ko, کے ke]
Instrumental	سے se	Oblique form+ سے (se)
Genitive	کا ka, کی ki, کے ke	Oblique + [کا ka, کی ki, کے ke]
Locative	میں main, تک tak, تلے talle, پر par, تلک tallak	Oblique + [میں main, تک tak, تلے talle, پر par, تلک tallak]
Vocative	اے (A)	اے (A) + Oblique form

Motivated by successful syntactic approaches, specifically Hobbs [12], the first syntax based approach, we aim to analyze and explore the syntactic structures of reflexive anaphors in Urdu. Syntax based approaches follow rules of language, which govern the arrangement of various language elements to make the meaningful sentences. They also ensure the existence of syntax tree and provide help in finding referent [1]. Many approaches for AR use world knowledge. However, use of world knowledge does not explain how the disambiguation process works [31]. Syntax based AR approach requires less information and has a low computational cost [1].

Reflexive anaphors are preceded/succeeded by noun/pronoun in combination with adverb, adjective, etc. to which they refer within the same clause. In generative grammar, a reflexive pronoun is an anaphor that must be bound by its antecedent [32].

In English, reflexive anaphors are “yourself”, “oneself”, “myself”, “himself”, etc. These anaphors end in ‘self’ for singular and in ‘selves’ for plural. They are used, where the object and subject of the sentence are the same. For example:

- Kathrine drove herself.
- Prime minister visits the venue himself.

In first example, *herself* is referring to subject Kathrine and in second example *himself* refers to subject Prime minister. Urdu language has two types of reflexive anaphors.

3.1 Possessive Reflexive (PR) anaphor

They are used to show the possession of an entity within the same clause. Their appearance changes with possessee and is not affected by gender and number of possessor. In Urdu, PR anaphors are "اپنی" (apni), "اپنا" (apna), and "اپنے" (apne). Following section presents the analysis of the syntactical structure of these anaphors.

In most common syntactical structure, PR anaphor follows a noun or personal pronoun (PP), which in turn is its referent. Consider the following example:

علی اپنا کھانا ختم کرچکا ہے۔

'Ali has finished his (own) meal'

In this example, PR anaphor اپنا (Apna) refers to its preceding noun, علی ('Ali'), which is a nominative case of noun. **Fig. 1** shows the syntax tree of the above example.

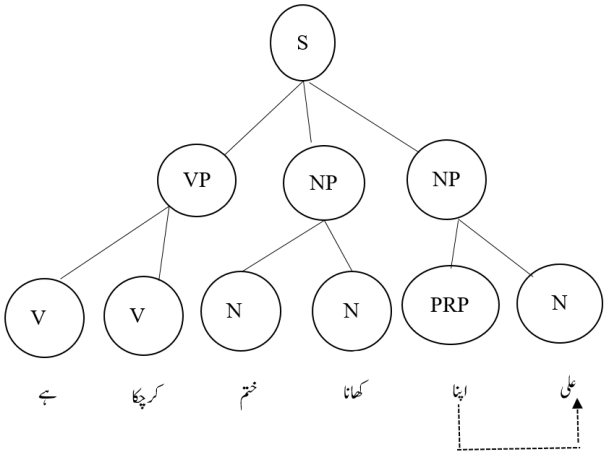

Fig. 1. Syntax tree of PR anaphor

Syntactically, PR anaphor is used with various language entities. Following are various syntactical structures of the use of PR anaphors of Urdu.

3.1.1 PR anaphors preceded by noun/PP

For example:

علی اپنے سکول چلا گیا ہے۔

'Ali has gone to his school

In this case, PR anaphor اپنے is preceded by noun علی, which is its referent.

3.1.2 PR anaphor preceded by ergative case

For example:

اسد نے اپنی جیپ کو آگے بڑھایا ہی تھا کہ بتی سرخ ہو گئی۔

'As soon as Asad moved his jeep forward, the signal turned red'

In this case, PR anaphor اپنی is preceded by clitic نے (ergative case) and noun اسد. The noun before clitic نے is the referent.

3.1.3 PR anaphor preceded by adverb and a noun/PP

For example:

- رحیم بھی اپنی غزل سنانے کے لیئے انتظارکررہا تھا

'Rahim was also waiting to recite his ghazal'

Here, in this case, PR anaphor اپنی is preceded by adverb بھی and noun رحیم, which is its referent.

3.1.4 PR anaphor preceded by a dative case

For example:

- علی کو اپنے گھوڑے کے لیئےگھاس لانی ہے

'Ali has to bring grass for his horse'

Here, in this case, PR anaphor اپنے is preceded by clitic کو and noun علی, which is its referent.

3.1.5 Two PR anaphors connected by word "اور" (Aur)

Two PR anaphors, connected by word **اور** (and)**,** are used together in a clause to refer to a preceding entity and also show the possession of succeeding entity. Consider the following example:

- احمد اپنی اور اپنے دوست کی آئسکریم لے آیا

'Ahmad brought his own and his friend's ice cream'

Here, in this example, the connector اور (Aur) is connecting two PR anaphors اپنی and اپنے. Both PR anaphors refer to preceding noun احمد. And also showing the possession of succeeding entity دوست (friend) to referent.

3.1.6 PR anaphor preceded by vocative case

Vocative case is used to call/identify a person. Clitic اے (A) is used before a noun.

For example:

- اے لڑکے اپنی گاڑی لے آو

'O boy! Bring your car'

In this case, PR anaphor اپنی is preceded by noun لڑکے (vocative case). The noun before PR anaphor اپنی is the referent.

3.2 Non-possessive reflexive (NPR) anaphors

NPR or emphatic reflexive anaphors are personal pronouns. In Urdu, the word خود (self) is preceded by noun/PP to make a compound word as NPR anaphor. For example, میں خود (myself), وہ خود (himself/herself), آپ/تم خود (yourself), ہم خود (ourselves), etc. They are used to show that someone has performed a task without anyone's help. NPR anaphors are mostly used in object position; however, they can be used in any participant position. Consider the following example:

- آصف خود وہاں گیا

'Asif himself went there.'

In above example, reflexive pronoun خود (‘khudd’), preceded by noun آصف, is referring to it and is showing that action done by preceding noun آصف is emphatic. Its syntax tree is shown in **Fig. 2**.

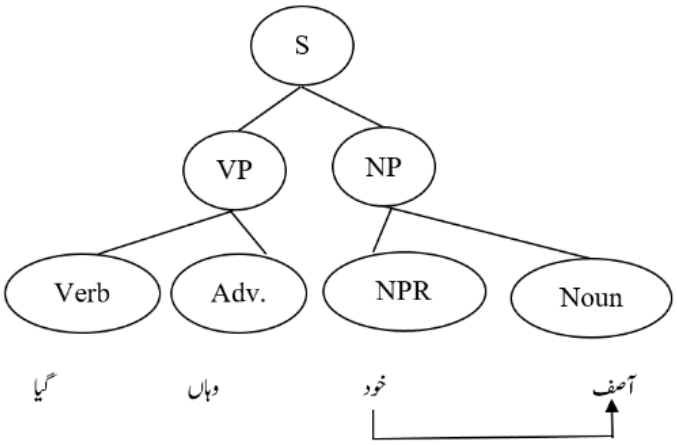

Fig. 2. Syntax tree of NPR anaphor

Syntactically, NPR anaphor is used with various language entities. Following are the syntactical uses of NPR anaphor in Urdu.

3.2.1 NPR anaphor “خود” combined with a noun or PP

For example:

بھائی خود جب سکول میں پڑھتے تھے تو وہ بھی بائیسکل پر آتے تھے۔

‘When brother himself studied in school, he used to come on bicycle’

Here, the NPR anaphor خود is preceded by noun بھائی, which is its referent.

3.2.2 NPR anaphor “خود” preceded by ergative case

For example:

خوشی کی بات یہ ہے کہ فاطمہ نے خود ہی گھر کا کام کر لیا۔

‘Good news is that Fatima herself has done homework’

In this case, NPR anaphor خود is preceded by clitic نے (ergative case) and noun فاطمہ, which is its referent.

3.2.3 NPR preceded by dative case

For this noun case, clitic کو (ko) and کے (ke) are used after the noun. For example:

احمد کوخود رنز بنانے چاہیئں۔

‘Ahmad himself should make runs’

In this structure, NPR anaphor is preceded by clitic کو (dative case) and noun احمد, which is its referent.

3.2.4 NPR "خود" preceded by an adverb and a noun/PP

For example:

احمد بھی خود لاہور جانا چاہتا ہے۔

'Ahmad himself too wants to go to Lahore'

In this case, NPR anaphor خود is preceded by adverb بھی and noun احمد, which is its referent.

3.2.5 NPR anaphor "خود" preceded by word "بذات" and a noun/PP

The word بذات (in person) appears before NPR anaphor (خود) to say emphatically that someone has performed some action in person. For example:

احمد بذات خود اجلاس میں شریک ہوا۔

'Ahmad himself attended the meeting'

In this structure, NPR anaphor خود is preceded by word بذات and noun احمد, which is its referent and showing that some action is performed by noun in person.

3.2.6 PR and NPR anaphors together

Any PR anaphor (اپنا / اپنے / اپنی) and NPR anaphors خود are used together as a compound word to say/show some emphatic activity and also possession of some entity to the referent. For example:

میں خود اپنی مرضی سے کیمسٹری پڑھ رہا ہوں۔

'I am studying chemistry on my own will'

Here, combination اپنی خود (myself + my own) shows that the action is performed emphatically and possession of مرضی (will) to the noun میں (I), which is its referent.

3.2.7 Distributive Reflexive anaphor

In this syntactical structure, a PR anaphor is repeated twice and acts like a reciprocal in the relevant aspect. This twice use of PR anaphor refers to a group or noun (in plural) and show the distribution of succeeding entity to referent. For example, اپنے اپنے (Apne Apne), اپنا اپنا (Apna Apna), and اپنی اپنی (Apni Apni). This twice use of PR anaphor reflects the behavior of both distributive and reflexive anaphors together. For example:

پرنسپل کی تقریر کے بعد طلباء خوش ہو گئے اور اپنی اپنی کلاس میں چلے گئے۔

'The students became happy and went to their classrooms after the speech of principal'

Twice use of PR anaphor اپنی اپنی (Apni Apni) is referring back to noun طلباء (students), which is referent here and also the possession and distribution of entity کلاس (classroom) to them.

3.3 Rules for resolving reflexive anaphors

After exploring and analyzing various syntactical structures of PR and NPR anaphors, the following heuristic rules are formulated for their resolution. Apparently, rules (i) to (iv) in 3.3.1 and 3.3.2 seem identical but their implementation is dissimilar.

3.3.1 Rules for PR anaphors

i) If a PR anaphor is preceded by a noun/PP (Nominative case), select it as referent.

ii) If a PR anaphor is preceded by ergative case, select that noun/PP as referent and mark the case.

iii) If a PR anaphor is preceded by adverb and a noun/PP, select noun/PP as referent.

iv) If a PR anaphor is preceded by dative case, select noun/PP as referent and mark the case.

v) If two PR anaphors together, separated by word اور (and) and preceded by noun/PP, select noun/PP as referent.

vi) If a PR anaphor is preceded by a noun (vocative case), select noun as referent and mark the case.

3.3.2 Rules for NPR anaphors

i) If an NPR anaphor is preceded by a noun/PP (Nominative Case), select it as referent.

ii) If an NPR anaphor is preceded by ergative case, select noun/PP as referent and mark the case.

iii) If an NPR anaphor is preceded by dative case, select noun/PP as referent and mark the case.

iv) If an NPR anaphor is preceded by adverb and noun/PP, select noun/PP as referent.

v) If combination of word بذات and NPR anaphor is preceded by a Noun/PP, select noun/PP as referent, and take the combination (بذات خود) as myself, herself, himself, etc. depending upon gender and number agreement of referent.

vi) If the combination of a PR and NPR anaphor is preceded by noun/PP, select noun/PP as referent. Take the combination as myself, herself, himself, etc. depending upon gender and number agreement of referent.

vii) Twice use of any PR anaphor refers to a group or noun (in plural) in preceding text.

Remaining noun case are either not applicable with reflexive pronouns or covered by the rules stated above. Genitive case is used to show the possession or ownership of something like possessive reflexive pronouns [27]. Locative case is used to indicate the location. Instrumental case is used to indicate that a noun is being used as instrument. Accusative case is used with the object of the sentence [28]. It is dative case when it is used with subject of the sentence. Oblique case is used when the noun or pronoun is the object of a verb/preposition. It is nominative case when it is used as subject of the sentence [29].

4. Proposed Framework

Our goal is to build a framework for an anaphora resolution system which resolves reflexive anaphoric links present in Urdu text document. Here, we present a framework which is based on syntactical rules. Our proposed framework not only resolves the reflexive anaphoric links but also manages noun case for further NL processing. The architecture of the proposed reflexive anaphora resolution system in the form of block diagram is shown in **Fig. 3**.

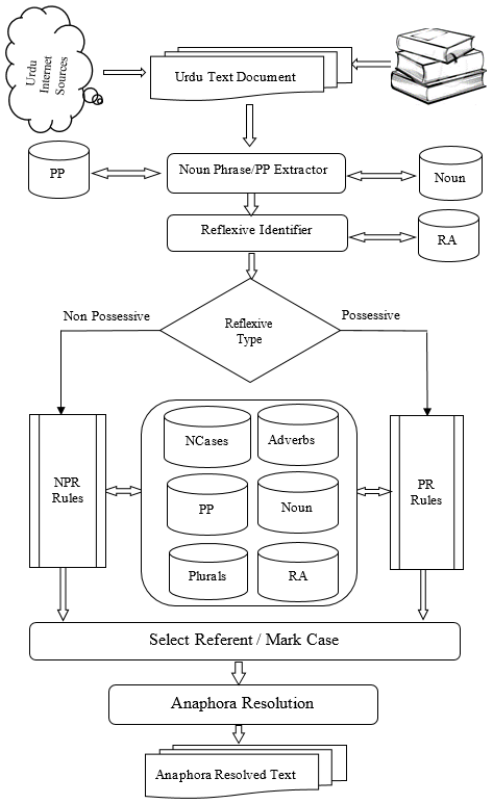

Fig. 3. Architecture of proposed framework

The workflow of the framework starts with input of Urdu text document. This document is prepared by extracting sentences which contain reflexive anaphoric links from various Urdu sources like children story books, newspapers, sports magazines, and various internet sources [33-35], where a huge amount of Urdu text is available. Most of the data collected and organized into discourse units manually in the form of a document, for the purpose of experiment.

The first module 'Noun Phrase/PP Extractor' extracts the noun and personal pronoun from the discourse units as candidate antecedent. It uses two resources 'Noun" and 'PP'. We prepared resource 'Noun', which contains most commonly used nouns in daily life. With the help of this resource, we identify and extract the nouns as proposed candidate(s) for antecedent. Similarly, resource 'PP' contains the personal pronouns of Urdu to help in identifying them in Urdu text document. Next module 'Reflexive Identifier' scans the discourse units and identifies reflexive anaphors by using resource 'RA'. This resource contains all reflexive anaphors (possessive and non-possessive).

After identification of reflexive pronoun our proposed system makes decision about the type of reflexive anaphor. It invokes module 'PR rules', when it is possessive reflexive pronoun and invokes the module 'NPR rules' for non-possessive reflexive anaphor.

Module 'PR rules' is developed on the basis of rules stated in section 3.3.1. It applies these rules to locate the referent for PR anaphors in the discourse unit. It utilizes various resources for this task. Resource 'NCase' contains the data about noun cases and their clitics shown in **Table 1**. Resource 'Adverbs' contains list of most commonly used Urdu adverbs. It helps in identifying the adverb and noun/PP combination as referent.

Module ‘NPR rules’ is developed on the basis of rules stated in section 3.3.2. It applies these rules to identify referent for NPR anaphor in the discourse unit. To access singular-plural forms resource ‘Plurals’ is used. It contains a collection of singular-plural words of day to day life. It helps in identifying referent for the case when a PR anaphor is used twice to reflect the possessive and distributive behavior. Next module ‘Select Referent/Mark case’ receives the output from one of the two previous modules. It makes the text ready for resolution and forwards it to last module ‘Anaphora Resolution’, which performs necessary adjustment in the text, if needed, and outputs the resolved text. It also retains the clitic, if any, for any additional NLP activity.

4.1 Algorithm for resolving Reflexive Anaphora

On the basis of rules 3.3.1 and 3.3.2, we developed following algorithm to resolve reflexive anaphoric links in Urdu discourse.

```
/* A list of words of the discourse unit is created. ‘m’ denotes the location of reflexive
anaphor. All method names are self-explanatory */
  /*Algorithm for module PRP*/
  m=search_PRP()
  if is_noun(x[m-1])=”yes” or is_PP(x[m-1])=”yes” then
  antecedent=x[m-1]        /*Nominative case */
  else if x[m-1]=“نے”  and (is_noun(x[m-1])=”yes” or is_PP(x[m-1])=”yes”) then
   {antecedent=x[m-2]       /*Ergative case*/
    Case=”ergative”}
      else if is_adv(x[m-1])=”yes” and (is_noun(x[m-1])=”yes” or is_PP(x[m-1])=”yes”)
 then antecedent=x[m-2]
  else if x[m-1]=” کو” and (is_noun(x[m-1])=”yes” or is_PP(x[m-1])=”yes”) then
  antecedent=x[m-2]        /*dative case*/
  else if x[m+1]= “اور” and is_PRP(x[m+2])=”yes” and (is_noun(x[m-1])=”yes” or
      is_PP(x[m-1])=”yes”) then
       antecedent=x[m-1]
  else if is_noun(x[m-1])=”yes” and x[m-2]=”اے ” then
       {antecedent=x[m-1]     /*Vocative Case*/
       Case=”vocative”}
 else if is_PRP(x[m+1])=”yes”     /*distributive behavior*/
 antecedent=search_plural()
       end /* end of all if */
  /*Algorithm for module NPRP*/
  m=search_NPRP()
  if is_noun(x[m-1])=”yes” or is_PP(x[m-1])=”yes” then
  antecedent=x[m-1]        /*Nominative case */
  else if x[m-1]=“نے”  and (is_noun(x[m-1])=”yes” or is_PP(x[m-1])=”yes”) then
   {antecedent=x[m-2]       /*Ergative case*/
    Case=”ergative”}
  if x[m-1]=” کو” and (is_noun(x[m-1])=”yes” or is_PP(x[m-1])=”yes” then
  antecedent=x[m-2]        /*dative case*/
  else if is_adv(x[m+1])=”yes” and (is_noun(x[m-1])=”yes” or is_PP(x[m-1])=”yes”)
      then
  antecedent=x[m-1]
  else if x[m-1]=” بذات” and (is_noun(x[m-2])=”yes” or is_PP(x[m-2])=”yes” then
```

```
antecedent=x[m-2]
else if is_PRP(x[m+1])=”yes” and (is_noun(x[m-1])=”yes” or is_PP(x[m-1])=”yes”)
    then
antecedent=x[m-1]     /*case of myself, himself,etc.*/
end /* end of all if */
```

5. Experiment and Results

Evaluation of a system in its development process is an important task. It helps in assessing the system and indicates different areas which need improvement to get more accurate results. To apply the anaphora resolution algorithm, a dataset containing anaphoric references is desired. Due to unavailability of precise dataset, authors collected the text from different sources for this particular task. Ashima et al. [36] extracted required text from Hindi stories and news. S. Lakhmani et al. [25] applied their algorithm on Hindi short stories. Hobbs [12] applied algorithm on first chapter of Arthur Hailey's novel Wheels. Similarly, in Malayalam, Pareed et al. [37] picked children short stories to apply their algorithm. Lakhmani et al. [38] used Punjabi story domain; and Palomar, M. et al. [39] applied algorithm on extracted text from Spanish literary books. Uppalapu [40] prepared two datasets containing Hindi short and long stories. For Pashto, Ali, R. et al. [41] extracted required text from Pashto stories and news.

In order to test the proposed framework, we extracted text containing reflexive anaphoric links from children stories [33], Urdu books [34] and sports news [35]. Children story domain contains simple sentences with less complexity, while others contain relatively complex sentences.

For evaluation of AR system, there does not exist a standard scoring method because AR systems are not tested on same dataset [1]. Initially, to evaluate AR system the success rate was introduced by Hobbs [12] as defined in the following equation:

$$\text{Hobbs metric (Success rate)} = \frac{\textit{Number of correct anaphora resolved}}{\textit{Total umber of anaphora resolved by an algorithm}}$$

We used Hobbs metric to evaluate our reflexive anaphora resolution system and Confidence Interval (CI) which tells how well the sample statistic estimates the underlying population and provides a range of values which is likely to contain the population parameter (i.e. success rate in our case) of interest.

Our approach and algorithm work well and the results are satisfactory. Our collected text contains 140 different reflexive anaphoric references. System resolved 120 references accurately and achieved overall success rate of 85.71%. **Table 2** shows the distribution of these 140 anaphoric links over reflexive anaphors along with success rate and CI which is s constructed at confidence level of 90%.

Table 2. Reflexive anaphora resolution results

Anaphor Type	Anaphor	Total	Correctly Resolved	Success Rate%	CI	
					Lower bound	Upper bound
PR	اپنا (Apna)	30	26	86.67	0.77	0.97
	اپنے (Apne)	25	22	88	0.77	0.99
	اپنی (Apni)	22	19	86.36	0.74	0.98

NPR	خود (Khudd)	33	28	84.84	0.75	0.95
	PR + NPR	18	15	83.33	0.69	0.98
	Distributive Possessive	12	10	83.33	0.66	1.00

For PR anaphora, success rate ranges between 86.36% to 88% and for NPR anaphora it is 84.84%. When possessive and non- possessive anaphors are used together, the success rate is 83.33%, and success rate is 83.33% for distributive possessive type.

To find out success rate for PR anaphors, when used with various language entities in a syntactical structure, in-depth analysis is carried out. **Table 3** shows these results.

Table 3. PR anaphors with various language entities

Entity	Total	Correctly Resolved	Success Rate%	CI	
				Lower bound	Upper bound
Noun/PP	30	27	90.00	0.81	0.99
Clitic & Noun/PP	23	20	86.96	0.75	0.99
Adverb+Noun/PP	14	13	92.86	0.82	1.00
PR and NPR + Noun/PP	10	7	70.00	0.46	0.94

Success rate ranges between 86.96% to 92%, when PR anaphors are used with noun/PP, adverbs, and noun cases. Success rate is 70% for combination of PR and NPR with noun/PP. This decrease in success rate is due to the complexity of the syntactical structure, as three entities are combined.

Detailed analysis of NPR anaphors with various language entities in different syntactical structures is shown in **Table 4**.

Table 4. NPR anaphors with various language entities

Entity	Total	Correctly Resolved	Success Rate%	CI	
				Lower bound	Upper bound
Noun/PP	15	13	86.67	0.72	1.00
Clitic & Noun/PP	8	6	75	0.5	1
Adverb + Noun/PP	4	3	75	0.39	1.00
Noun/PP + بذات	6	6	100	1	1

Success rate of NPR anaphors is 75%, when used with noun cases, which is a little lower. The reason behind this decrease is the complications in the use of clitics. Some clitics represent more than one noun cases and also more than one clitics can be used with one noun in Urdu. Similarly, for adverb, success rate is 75%. Multiple usage of an adverb and more than one adverb in a sentence at a time make the process complex. For preceding word بذات (in person) of NPR anaphor, the success rate is 100%, due to its simple syntactical structure.

Confidence Interval (CI) is calculated which shows that there is 90% likelihood that true average model success rate exists between 0.70 and 0.99. It is also obvious from the results that success rate increases with the sample size. **Table 2** shows that the larger the sample size, the more precise the estimate and the smaller the confidence interval. In the first row of **Table 2**, where sample size is 30, estimate is more precise (smaller CI) as compared to sample of size 12, where CI is less precise.

There is no criterion available for comparing the performance of anaphora resolution approaches because they are not evaluated on the same dataset [1]. Also each language has its own set of anaphors, grammatical word order and principles for number and gender handling. **Table 5** shows success rate of different anaphora resolution algorithm. Our success rate is 85.71%, which is pleasing and encouraging.

Table 5. Success rate in different languages

Authors	Language	Anaphora Type	Data set	Success Rate%
Ashima et al. [36]	Hindi	Gender, Number agreement	Stories/News	71
Lakhmani et al. [25]	Hindi	Pronominal	Short Stories	70
J. Hobbs [12]	English	Pronominal	First chapter of a Novel "Wheels"	98
S. Lappin [13]	English	Pronominal	Computer manual	86
Khan et al. [19]	Urdu	Pronominal	Stories/News	79
R. Ali et al. [41]	Pashto	Reflexive	Stories, news	87
Pareed et al. [37]	Malayalam	Pronominal	Children stories	80-84
Lakhmani et al. [38]	Punjabi	Pronominal	Stories domain	64
M. Palomar et al. [39]	Spanish	Third person, demonstrative, reflexive and omitted pronoun	Literary books	76.8
B. Uppalapu et al. [40]	Hindi	Third person pronoun	Long and Short Stories	61

6. Conclusion

This paper deals with a novel syntactic structure-based framework for resolving reflexive anaphora in Urdu discourse. The exclusive feature of the framework is designing of algorithm, based on Hobbs syntactic approach and a system developed by Lappin & Leass. The algorithm is designed based on rules formulated as a result of structural analysis of reflexive anaphors. In our realm of knowledge, based on the available literature, this is the first syntactic structure-based framework for resolving reflexive anaphora in Urdu discourse, which is the main contribution of this work. This framework can be incorporated as an integral part of a natural language processing system to improve its performance in areas like Information Extraction (IE), Sentiment Analysis (SA), Question Answering (QA), and Machine Translation (MT). Structured analysis of reflexive anaphors and formulated rules can be utilized in other approaches of anaphora resolution. Similar frameworks can be designed for other types of anaphora to develop a large-scale system for Urdu anaphora resolution. To test the framework, we conducted an experiment on a textual dataset prepared from children's stories, Urdu books, etc. containing reflexive anaphoric links. We developed a module to identify reflexive anaphors and extract noun as potential antecedent candidate. Our developed algorithm selects

the appropriate antecedent and resolves the anaphoric links. We used success rate metric for system evaluation and Confidence Interval (CI) to find out the probability that a population parameter will fall between a set of values for a certain proportion of time. The success rate and CI range obtained are significant and encouraging. Urdu is a vast and free word order language affecting the structure of sentence, which in turn affects the success of the system. Also, Urdu is an accommodative language for the words of other languages, having a scope for refinement in building the rules for resolution of the reflexive anaphoric links in this case. Success rate can further be improved by increasing/improving the set of resources required; and by incorporating more novel linguistic and syntactic rules. Currently our focus is to develop systems for other anaphora types in Urdu as well as in other Pakistani languages like Punjabi, Pashto, Saraiki, etc.

References

[1] K. Lata, P. Singh, and K. Dutta, "A comprehensive review on feature set used for anaphora resolution," *Artificial Intelligence Review*, pp. 1-90, 2020. Article (CrossRef Link)

[2] E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102-107, 2016. Article (CrossRef Link)

[3] R. Nithya, "Need for anaphoric resolution towards sentiment analysis-a case study with scarlet pimpernel (Novel)," *International Journal of Education Management Engineering(IJEME)*, no. 1, pp. 37-50, 2019. Article (CrossRef Link)

[4] D. Stojanovski and A. Fraser, "Improving anaphora resolution in neural machine translation using curriculum learning," in *Proc. of Machine Translation Summit XVII Volume 1: Research Track*, pp. 140-150, 2019. Article (CrossRef Link)

[5] W. Zhao, P. Haiyun, S. Eger, E. Cambria, and M. Yang, "Towards scalable and reliable capsule networks for challenging NLP applications," *arXiv preprint arXiv:1906.02829*, 2019. Article (CrossRef Link)

[6] M. Ting, R. A. Kadir, A. Azman, T. M. T. Sembok, and F. Ahmad, "Named entity enrichment based on subject-object anaphora resolution," in *Proc. of Intelligent Computing-Proceedings of the Computing Conference*, pp. 873-884, 2019. Article (CrossRef Link)

[7] A. Siddiqui, "Jamia Ul Qawaid," Markazi Urdu Board, Lahore, Pakistan: Markazi Urdu Board, 1971. Article (CrossRef Link)

[8] V. Dayal and A. Mahajan, The Status of Case: Clause Structure in South Asian Languages, vol. 61, Dordrecht: Springer, Kluwer Academic Publishers, 2004. Article (CrossRef Link)

[9] M. A. Khan and J. A. Nasir, "Distributive anaphora resolution in Urdu discourse," in *Proc. of the 4th International Conference on Emerging Technologies*, pp. 38-43, Oct. 2008. Article (CrossRef Link)

[10] D. G. Bobrow, "A question-answering system for high school algebra word problems," in *Proc. of Fall Joint Computer Conference, part I*, pp. 591-614, 1964. Article (CrossRef Link)

[11] T. Winograd, "Understanding natural language," *Cognitive Psychology* 3, no. 1, pp. 1-191, 1972. Article (CrossRef Link)

[12] J. R. Hobbs, "Resolving pronoun references," *Lingua 44*, no. 4, pp. 311-338, 1978. Article (CrossRef Link)

[13] S. Lappin and H. J. Leass, "An algorithm for pronominal anaphora resolution," *Computational Linguistics* 20, no. 4, pp. 535-561, 1994. Article (CrossRef Link)

[14] R. Mitkov, "An integrated model for anaphora resolution," in *Proc. of the 15th Conference on Computational Linguistics*, pp. 1170-1176, 1994. Article (CrossRef Link)

[15] R. Mitkov, "Anaphora resolution: A combination of linguistic and statistical approaches," in *Proc. of the Discourse Anaphora and Anaphor Resolution(DAARC'96)*, 1996. Article (CrossRef Link)

[16] R. Mitkov, "An uncertainty reasoning approach for anaphora resolution," in *Proc. of the Natural Language Processing Pacific Rim Symposium(NLPRS'95)*, vol. 25, pp. 149-154, 1995. Article (CrossRef Link)

[17] R. Mitkov, "Robust pronoun resolution with limited knowledge," in *Proc. of the 18th International Conference on Computational Linguistics(COLING'98/ACL'98)*, pp. 869-875, 1998. Article (CrossRef Link)
[18] R. Mitkov, R. Evans, and C. Orasan, "A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method," in *Proc. of International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 168-186, 2002. Article (CrossRef Link)
[19] M. A. Khan, M. N. Ali, and M. A. Khan, "Pronominal Anaphora Resolution in Urdu Discourse," in *Proc. of IEEE ICET 2nd International Conference on Emerging Technologies*, pp. 543-548, 2006. Article (CrossRef Link)
[20] B. Kulsoom and R. Begum,"Urdu Anaphora resolution in monologue," M.S. Thesis, Department of Computer Science University of Peshawar, Pakistan, 1993.
[21] R. M. K. Sinha, and A. Jain. "AnglaHindi: an English to Hindi machine-aided translation system," *Machine Translation Summit IX*, p. 497, 2003. Article (CrossRef Link)
[22] K. Dutta, N. Prakash, and S. Kaushik, "Resolving pronominal anaphora in Hindi using Hobbs algorithm," *Web Journal of Formal Computation and Cognitive Linguistics*, vol. 1, no. 10, pp. 5607-5611, 2008. Article (CrossRef Link)
[23] P. Dakwale, V. Mujadia, and D. M. Sharma, "A hybrid approach for anaphora resolution in hindi," in *Proc. of the 6th International Joint Conference on Natural Language Processing*, pp. 977-981, 2013. Article (CrossRef Link)
[24] S. Lalita Devi, V. S. Ram, and P. Rao, "A generic anaphora resolution engine for Indian languages," in *Proc. of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1824-1833, 2014. Article (CrossRef Link)
[25] P. Lakhmani, S. Singh, and P. Mathur, "Gazetteer Method for Resolving Pronominal Anaphora in Hindi Language," *International Journal of Advances in Computer Science and Technology*, vol. 3, no. 3, 2014. Article (CrossRef Link)
[26] K. Wohiduzzaman and S. Ismail, "Recommendation system for bangla news article with anaphora resolution," in *Proc. of the 4th International Conference on Electrical Engineering and Information & Communication Technology(ICEEiCT)*, pp. 467-472, 2018. Article (CrossRef Link)
[27] R. L. Schmidt, Urdu: An Essential Grammar, 1st Edition, London: Psychology Press, 1999. Article (CrossRef Link)
[28] A. Siddiqui, "Jamia Ul Qawaid," Markazi Urdu Board, Lahore, Pakistan: Markazi Urdu Board, 1971. Article (CrossRef Link)
[29] M. Butt and T. H. King, "The status of case," in Clause Structure in South Asian Languages, Dordrecht: Springer, pp. 153-198, 2004. Article (CrossRef Link)
[30] M. Humayoun, H. Hammarström, and A. Ranta, "Urdu morphology, orthography and lexicon extraction," *Chalmers Tekniska Högskola*, 2006. Article (CrossRef Link)
[31] V. J. Leffa, "Anaphora resolution without world knowledge," *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, vol. 19, no. 1, pp. 181-200, 2003. Article (CrossRef Link)
[32] Reflexive pronoun, Wikipedia. Article (CrossRef Link)
[33] Urdu Point Kids بچوں کی دنیا Bachon Ki Dunya. Article (CrossRef Link)
[34] Urdu Books اردو کتابیں . Article (CrossRef Link)
[35] کھیل - BBC News اردو. Article (CrossRef Link)
[36] A. Ashima, S. Kaur, and C. Rajni Mohana, "Anaphora Resolution in Hindi: A Hybrid Approach," in *Proc. of The International Symposium on Intelligent Systems Technologies and Applications*, pp. 815-830, 2016. Article (CrossRef Link)
[37] A. A. Pareed and S. M. Idicula, "An Integrated Framework for Pronominal Anaphora Resolution in Malay-alam," *Special Issue on Innovation in Computing, Engineering Science & Technology*, vol. 4, no. 5, pp. 287-293, 2019. Article (CrossRef Link)
[38] P. Lakhmani, S. Singh, P. Mathur, and S. Morwal, "Pronominal Anaphora Resolution In Punjabi Langugage," *International Journal of Computer Science & Technology*, vol. 4, no. 4, pp. 99-105, 2014. Article (CrossRef Link)

[39] M. Palomar, A. Ferrández, L. Moreno, P. Martínez-Barco, J. Peral, M. Saiz-Noeda, and R. Muñoz, "An algorithm for anaphora resolution in Spanish texts," *Computational Linguistics*, vol. 27, no. 4, pp. 545-567, 2001. Article (CrossRef Link)
[40] B. Uppalapu and D. M. Sharma, "Pronoun resolution for Hindi," in *Proc. of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009)*, pp. 123-134, 2009. Article (CrossRef Link)
[41] R. Ali, M. A. Khan, and M. Ali, "Reflexive anaphora resolution in Pashto discourse," in *Proc. of Accepted in the 2nd Conference on Language and Technology (CLT2009)*, 2009. Article (CrossRef Link)

Jamal Abdul Nasir received Master degree in Computer Science from University of Peshawar, Pakistan and is a Ph.D. scholar in Gomal University, D.I.Khan, KPK, Pakistan. He is serving as Assistant Professor in Institute of Computing and Information Technology (ICIT), Gomal University, D.I.Khan, KPK, Pakistan. His research interests include Natural Language Processing, Machine Learning, Knowledge Engineering, and Object Oriented Analysis and design.

Zia ud Din earned his Master degree in Computer Science from University of Peshawar and Ph. D. in Software Engineering from ICIT, Gomal University, D.I.Khan, KPK, Pakistan. His research interests include Software Engineering, Software Cost Estimation, Software Project Management and Machine learning. Presently, he is working as the Director of Institute of computing and Information Technology (ICIT), Gomal University, D.I.Khan, KPK, Pakistan.