An Optimal Order of Factors for the Computational Treatment of Personal Anaphoric Devices in Urdu Discourse

Mohammad Naveed Ali
Department of Computer Science, University of Peshawar, NWFP, Pakistan
Naveed_asadecp@yahoo.com

M. A. Khan
Department of Computer Science, University of Peshawar, NWFP, Pakistan
m.abid6@gmail.com

Muhammad Aamir Khan
Department of Computer Science, University of Peshawar, NWFP, Pakistan
bitvox@yahoo.com

Abstract

Handling of human language by computer is a very intricate and complex task. In natural languages, sentences are usually part of discourse units just as words are part of sentences. Anaphora resolution plays a significant role in discourse analysis for chopping larger discourse units into smaller ones. This process is done for the purpose of better understanding and making easier the further processing of text by computer.

This paper is focused on the discussion of various factors and their optimal order that play an important role in personal anaphora resolution in Urdu. Algorithms are developed that resolves pronominal anaphoric devices with 77-80% success rate.

1 Introduction

In written text, cohesion occurs when some elements in a discourse are dependent on others and that refer to items backward in the text, both in the spoken or written text (Halliday and Hassan, 1976). Consider the following example

> *(1.1) Shah Rukh* Khan is off to one of *his* favorite cities- London, with *his* family. Now *he* is looking for another destination, not so much for holidaying though.
> *(The News Islamabad: June 2006)*

> (1.2) Bollywood actress *Bipasha Basu* has been signed for her new film Corporate. *She* is a single working woman, wants to get somewhere in life, on her own terms.
> *(The News Islamabad: June 2006)*

Cohesion in examples 1.1 and 1.2 is introduced due to the terms *he, his*, *her, she* and interpretation of these references depends upon some preceding terms. These referring terms are called anaphors or anaphoric devices (ADs). Halliday and Hassan described anaphora as 'cohesion which points back to some previous items' (Halliday and Hassan, 1976). The 'pointing back' words or phrases are called the anaphors (Halliday and Hassan, 1976) and the entities to which these point are called antecedents and the procedure of determining the antecedents of anaphors and subsequent replacement in some particular discourse is called anaphora resolution. According to Halliday and Hassan when anaphors are replaced by their corresponding antecedents, cohesion no more exists. Personal anaphoric devices (ADs) are the most widely used variety of ADs in Urdu text. These are further classified as first person, second person and third person anaphoric devices. Examples of first person ADs are میری، میرا، میں، ہمارے ہمیں،،مجھے، مجھکو، ہم، ہمارا، ہماری، ہمکو ([mæɪri], [mæɪra], [mæñ], [mʊʤheɪ], [mʊʤhkəʊ], [hʌm], [hʌmɒrɒ], [hʌmɒri], [hʌmkəʊ], [hʌmeɪñ], [hʌmɒreɪ]). Examples of second person ADs are تمہارا، تمہاری، تمکو تم، آپ، آپکا، آپکی، آپکورے،تمہیں،تمھا ([tʊm], [tʊmhɒrɒ], [tʊmhɒri], [tʊmkəʊ], [tʊmheɪñ], [tʊmhɒreɪ], [a:p] [a:pkɒ], [a:pki], [a:pkəʊ]). Examples of third person ADs are وہ، انکا، انکے، انکی، ان، اسکو، اسکے، اسکی، اسکا، اسے، انہیں ([vəʊh], [ʊseɪ], [ʊskɒ], [ʊski], [ʊskeɪ], [ʊskəʊ], [ʊn], [ʊnki], [ʊnkeɪ], [ʊnkɒ], [ʊnheɪñ]).

A lot of work has been done in English for the purpose of anaphora resolution and various

Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 81–90,
Hyderabad, India, January 2008.

algorithms have been devised for this purpose (Aone and Bennette, 1996; Brenan , Friedman and Pollard, 1987; Ge, Hale and Charniak, 1998; Grosz, Aravind and Weinstein, 1995; McCarthy and Lehnert, 1995; Lappins and Leass, 1994; Mitkov, 1998; Soon, Ng and Lim, 1999). Work has also been done in South Asian Languages such as Hindi and Malayalam for the purpose of anaphora resolution (Prasad and Strube, 2000; Sobha, 1998). Prasad and Strube (2000) worked on anaphora resolution in Hindi. Their approach relies on the discourse salience factors and is primarily inspired by the central idea of Centering theory (Grosz, Aravind and Weinstein, 1995). Centering theory has also guided the development of pronoun resolution algorithms, such as the BFP algorithm (Brenan, Friedman and Pollard, 1987) and the S-list algorithm developed by Strube (Strube, 1998). Prasad and Strube (2000) applied these algorithms to the resolution of pronouns in Hindi texts. They showed that the BFP algorithm cannot be successfully implemented for pronoun resolution in Hindi. They argued that better results can be obtained with an algorithm that does not use the Centering notions of the backward-looking center and the centering transitions for the computation of pronominal antecedents, such as the S-list algorithm (Prasad and Strube, 2000). Prasad and Strube used well established approaches for Hindi anaphora resolution. Sobha (1998) used knowledge poor rule based approach for reference resolution in Hindi and Malayalam languages that stands on very limited syntactic information. In Urdu language very little work has been done on discourse level especially in the field of anaphora resolution. Although, most of the anaphoric devices in Urdu and Hindi are same but the style and organization of discourses are bit different that causes the difference in anaphora resolution. Kulsoom et al worked on Urdu anaphora resolution but it appears to be the tip of an iceberg (Kalsoom and Rashida, 1993). Kulsoom et al (1993) only considered the morphological and lexical filters for the resolution of anaphora in Urdu discourses. However, these filters are not sufficient for Urdu anaphora resolution.

The rest of the paper is organized as follows: Section-2 describes the factors that play a vital role in Urdu anaphora resolution. Section-3 presents algorithms, implementation and evaluation for the resolution of personal anaphora; this is followed by the conclusion.

2 Factors that play vital role in Urdu anaphora resolution

Factors that can play a very important role in Urdu anaphora resolution beside morphological and lexical filters are topicalized structures, subject preferences, object preferences, repetitions, section heading and distance. How these factors are helpful in anaphora resolution in English language was worked out by Mitkov (Mitkov, 1998), but their role in Urdu discourse for the resolution of personal pronouns is more cherished. How these factors are helpful in the resolution of anaphoric devices in Urdu is done by Khan et al (Khan, Ali and Aamir, 2006). Ali et al also worked on these factors for the resolution of demonstrative ADs in Urdu discourse (Ali, Khan and Aamir, 2007).

2.1 Morphological and lexical filters

Consider an example in which anaphora is resolved on the basis of morphological filters.

ملکھی نے چارے کا بڑا سا گھتا اٹھایا اور آگے چل دی، فضل دین نے آگے بڑھ کر ملکھی کا بازو پکڑنا چاہا تھا لیکن نہ جانے کیوں وہ اندر سناٹے کي گونج- سائرہ ہاشمي((2-0)سکڑ کر رہ گیا تھا
[mɒlʌkhi] [neɪ] [ʧɒreɪ] [kɒ] [bə(r)ɒ] [sɒ] [ghʌthɒ] [ʊ:thɒyɒ] [ɒəʊr] [a:geɪ] [ʧʌl] [di]. [fʌzl] [dIn] [neI] [a:geI] [bhə(r)] [kə(r) [mɒlʌkhi] [kɒ] [bɒzʊ] [pɒkə(r)nɒ] [ʧɒhɒ] [thɒ] [leIkIn] [nɒ] [ʤɒneɪ] [kɪyʊñ] [vəʊh] [ʌndə(r)] [sʊkə(r)] [kə(r)] [rəh] [gIyɒ] [thɒ].
Mlukhi took the bundle of grass and moved ahead. Fazal Din had come forward to catch the arm of Mlukhi, but he did not have the courage to do so.

In Urdu, the word وہ ([vəʊh]) refers to both masculine and feminine antecedents. Also, it is used for translation of 'that'. Here the morphological filters are used for anaphoric disambiguation. In the above discourse, terminal of sentence is تھا ([thɒ]) that indicates the third person AD وہ refers to singular and masculine NP i.e. فضل دین ([fʌzl][dɪn]). In this way, ملکھي ([mɒlʌkhi]) will be ruled out to become the antecedent. Similarly, consider the example

انیلہ خط پڑھ کے اپنے ھوش وحواس کھو بیٹھی- کچھ عرصے اسکا علاج ھوتا رھا پھر وہ بہتر ھو گئ- (2-1) (امربیل- بانو قدسیہ)
[əˈnɪlɒ] [xʌt] [pə(r)h] [keɪ] [əˈpneɪ] [həʊʃ] o [hɒvɒs] [khəʊ] [beɪthi] [kʊʧ] [ʌrseɪ] [ʊskɒ]

[ælɒʤ] [həʊta] [rɒhɒ] [phɪr] [vəʊh] [bəhtə(r)] [həʊ] [gɒi].
After reading the letter Aneela lost her senses. She was treated for sometime and then she got better.

Since the terminal of sentence is گئ ([gɒi]), so it means وہ ([vəʊh]) refers to some feminine antecedent that is انیلہ ([əˈnIlɒ]) in the above text. The lexical filters are used to resolve anaphora on the basis of number and gender information. For example

لڑکوں نے پرنسپل صاحب کو درخواست کی "ہمارے امتحانات جلدی کروا دیے جائیں تاکہ بعد میں پراجیکٹ کرنے کیلئے ہمارے پاس زیادہ سے زیادہ وقت ہو-"(2-2)
[lə(r)kəʊ] [neI] [prɪnsɪpʌl] [sɒhɪb] [kəʊ] [də(r)khəʊɒst] [kɪ] [hɒmɒreI] [ɪmti:hɒnɒt] [ʤʌldi] [kərvɒ] [dIyeI] [ʤɒa:eIñ] [tɒkeI] [bɒd] [meIñ] [prɒjəct] [kə(r)neI] [KeIlIyeI] [hɒmɒreI] [pɒs] [zIyɒdɒ] [seI] [zIyɒda] [vʌkt] [həʊ].
Students submitted application to the Principal "our exams should be arranged earlier so that we will have maximum time for our project"

In 2.2, following the number information, the antecedent for ہمارے ([hɒmɒreI]) will be لڑکوں ([lə(r)kəʊ]). So, the remaining candidate پرنسپل صاحب ([prInsIpʌl] [sɒhɪb]) is ruled out on the basis of number mismatch.

Here is another example in which antecedents for third person anaphoric devices are found on the basis of morphological and lexical filters. In the following discourse, antecedent for third person AD وہ ([vəʊh]) is singular, feminine noun phrase فضل بی بی [fʌzəl] [bIbI]) since the terminal of sentence is چاہتی ہے ([ʧɒhti] [hæ]).

فضل بی بی نے جب یہ فیصلہ کر لیا تو اس نے خاندان کے دوسرے افراد کو بھی بتایا کہ وہ بھی پڑھنا چاہتی ہے-یوں تو سارے گاؤں والے اسکی بہت عزت کرتے تھے لیکن اس معاملے میں اسکو اسکی بچیوں کے حوالے سے سمجھانے کی کوشش کی گئ- (2-3)
(خمیازہ- نعمان اخلاق)
[fʌzəl] [bIbI] [neI] [ʤʌb] [yəh] [fæslɒ] [kə(r)] [lIyɒ] [təʊ] [ʊs] [neI] [xɒndɒn] [keI] [dəʊsɒreI] [ʌfrɒd] [kəʊ] [bhi] [bɒtɒyɒ] [kəh] [vəʊh] [pə(r)hnɒ] [ʧɒhti] [hæ]. [yʊñ] [təʊ] [sɒreI] [gɒʊñ] [vɒleI] [ʊski] [bɒhʊt] [Izət] [kə(r)teI] [theI] [lækIn] [Is] [mʊɒmɒleI] [meIñ] [ʊskəʊ] [ʊski] [bʌʧIyʊñ] [keI] [hɒvɒleI] [seI] [sʌmʤɒneI] [ki] [kəʊʃIʃ] [ki] [gɒi].
When Fazal Bibi decided she informed other family members that she also wants to study. Although, she was respectable for the whole village but in this matter she was advised keeping in view her daughters.

2.2 Topicalized structures

In Urdu, topicalized structures are more frequently used. Consider the example

آئ تو آپ کا رد عمل پرب ! مائ فیوڈل لارڈ جب منظر عام صاح کھر (2-4)کیا تھا-
[khə(r)] [sɒhIb]! [mɒi] [fʊdəl] [lɒrd] [ʤʌb] [mʌnzə(r)] [ɒm] [pə(r)] [a:i] [təʊ] [a:p] [kɒ] [rʌdeI] [ʌml] [kIyɒ] [thɒ].
Mr. Kher! When the book "My Feudal" Lord came into the market, what was your reaction.

فاطمہ! تمھاری یادوں کا کیا کروں؟ گھونسلے بوٹ اڑ کر کہیں نہ کہیں چلے جاتے ھیں- لیکن تمھارے عطا کردہ بوٹ تو صبح و شام خون جگر کا چوگا مانگتے ھیں- (2-5)(امربیل- بانو قدسیہ)
[fɒtImɒ]! [tʊmhɒri] [yɒdəʊñ] [kɒ] [kIyɒ] [kə(r)ʊñ]? [ghəʊsəleI] [bʊt] [ʊr] [kə(r)] [kɒhi:ñ] [nɒ] [kɒhi:ñ] [ʧɒleI] [ʤɒteI] [hæñ] [leIki:n] [tʊmhɒreI] [a:tɒ] [kə(r)dɒ] [bʊt] [təʊ] [sʊbh] o [ʃɒm] [xʊneI] [ʤIgə(r)] [kɒ] [ʧəʊgɒ] [mɒñgteI] [hæñ]
Fatima! What should I do with your memories? Every thing vanishes with the passage of time but your memories are like unripe grain which needs my blood to flourish.

In 2.4, the word آپ ([a:p]) refers to topicalized structure بصاح کھر ([khə(r)] [sɒhIb]). Similarly, in discourse 2.5 ADs تمھاری ([tʊmhɒri]) and تمھارے ([tʊmhɒreI]) refer to فاطمہ ([fɒtImɒ]). It must be noted that whenever topicalized structures appear in the Urdu discourses these become preferred antecedents for second person anaphoric devices.

2.3 Count of occurrences

It can be the case that in a particular discourse if a certain NP appears more frequently then it will be the potential antecedent for pronouns appearing in that text. For example, consider the following discourse

منٹو سے اخفا برتا گیا ، اسکی کئ وجوحات ہیں-منٹو ایک غیر جانبدار منٹو کے معاصر ادیب اس کے رویے اور تیز کلامی سے ادیب تھا- منٹو کی وجہ سےوہ ناپسندیدہ تھا- کھلم کھلا شراب نوشی نالاں تھے- نظر میں کی لوگوں ثقہ اسے پرپےدر پے فحاشی کے مقدمات نے - (2-6)بنا دیا تھا ملعون
(سعادت حسن منٹو کی کہانی- انیس ناگی)
[mʌntʊ] [seI] [ʌxfɒ] [bə(r)tɒ] [gIyɒ]. [Iski] [kɒi] [vɒʤʊhɒt] [hæñ]. [mʌntu:] [ək] [ghəIr] [ʤɒnIbdɒr] [əˈdi:b] [thɒ] [mʌntʊ] [keI] [mɒa:sIr] [əˈdi:b] [ʊskeI] [rɒvIyeI] [pəʊr] [teIz] [kɒlɒmi] [seI] [nɒlɒñ] [theI]. [khʊlʌm] [khʊlɒ] [ʃɒrɒb] [nəʊʃi] [ki] [vʌʤhɒ] [seI] [vəʊh] [nɒpɒsʌndIdɒ] [thɒ]. [mʌntʊ] [pə(r)] [pæ] [də(r)] [pæ] [fɒhɒʃi] [keI] [mu:kʌdmɒt]

[neɪ] [ʊseɪ] [sʌkɒ] [ləʊgəʊn] [ki] [nʌzə(r)] [meɪñ] [mʌlʊ:n] [bɒnɒ] [dɪyɒ] [thɒ].

Anger was shown to Muntoo. It has several reasons. Muntoo was an un-biased writer. Due to his aggressive attitude, his fellows were always angry with him. He was not liked because he used to drink openly. Due to continuous court cases regarding obscenity, he was not liked by gentlemen community.

Here the proper noun منٹو ([mʌntʊ]) appears repeatedly. So, on the basis of repetition, it will be the potential antecedent for most of personal pronouns e.g. وہ ([vəʊh]), اسے ([ʊseɪ]) and اسکے ([ʊskeɪ]) appearing in the above text.

2.4 Section headings

Section headings get high preference to become antecedents for most of personal pronouns in Urdu discourses. Consider the following example

شعیب اختر

شعیب اختر کرکٹ بورڈ کیلئے وہ انوکھا لاڈلا بن چکے ہیں جو گیند بیٹ سے کھیلنے کی بجائے چاند کی تمنا کرتے ہیں- وہ واحد باؤلر ہیں جنہوں نے اتنی کرکٹ نہیں کھیلی جتنا ان فٹ ہو کر آرام کیا ہے- وہ شہرت اور مقبولیت کے لحاظ سے نہایت خوش قسمت کھلاڑی ہیں، جسکی واحد خوبی یہ ہے کہ وہ دنیا کے تیز ترین باؤلر ہیں- جسکے نخرے عمران سے بھی زیادہ اٹھائے جاتے ہیں- (2-7) ("فیملی میگزین"- جون 2006)

[ʃʊæb] [ʌxtə(r)]

[ʃʊæb] [ʌxtə(r)] [krɪkɪt] [bəʊrd] [keɪlɪyeɪ] [vəʊh] [anəʊkhɒ][lɒdlɒ] [bʌn] [ʧʊkeɪ] [hæñ] [ʤəʊ] [geɪnd] [bæt] [seɪ] [kheɪlneɪ] [ki] [bɒʤæɪ] [ʧɒnd] [ki] [tɒmʌnnɒ] [kə(r)teɪ] [hæñ]. [vəʊh] [vɒhid] [bəʊlə(r)] [hæñ] [ʤɪnhəʊñ] [neɪ] [ɪtni] [kɪrkət] [nɒhɪñ] [kheɪli] [ʤitna:] [ʌnfɪt] [həʊ] [keɪ] [a:rɒm] [kɪyɒ] [hæ]. [vəʊh] [ʃəʊhrʌt] [ɒəʊr] [mʌkbʊlɪət] [keɪ] [lɪhɒz] [seɪ] [nɪhɒyʌt] [xʊʃ] [kɪsmʌt] [khɪlɒri] [hæñ] [ʤɪski] [vɒhɪd] [khʊbi] [yəh] [hæ] [kəh] [vəʊh] [dʊnyɒ] [keɪ] [teɪz] [tɒri:n] [bəʊlə(r)] [hæñ]. [ʤɪskeɪ] [nʌkhreɪ] [ɪmrɒn] [seɪ] [bhi] [zɪyɒdɒ] [ʊthɒeɪ] [jɒteɪ] [hæñ].

Shoaib Akhter

Shoaib Akhter has become a burden over the cricket board. He is the only bowler in Pakistani cricket team who has not played much cricket rather always took rest because of being unfit. He is lucky to become popular only because he is the fastest bowler in the world. He is given more importance even compared to Imran.

In the above discourse, شعیب اختر ([ʃʊæb] [ʌxtə(r)]) is section heading, so it will be the preferred antecedent for most of anaphoric devices appearing in the discourse and all other NPs will be ruled out to become the potential antecedents.

2.5 Distance

Distance plays an important role in finding the antecedents. For each anaphoric device such as ([ʊs], انکی اکن، اسکی، ان، ااس ، اسے، اسکو، اسک [ʊseI], [ʊskəʊ] [ʊskɒ] [ʊskɪ] [ʊn], [ʊnkɒ], [ʊnkɪ]), preference is given to the nearest object present in the same or immediate previous sentence. Consider following discourse

طلوع آفتاب سے تھوڑی دیر بعد ایک جھیل کے قریب پہنچ کر انور علی نے اپنے ساتھیوں کو رکنے کا حکم دیا اور اس نے لیگرانڈ کو گھوڑے سے اتار کرزمین پر لٹا دیا- بعض سپاھیوں نے تھیلوں سے باسی روٹیاں نکالیں اور ساتھیوں میں تقسیم کیں اور وہ جھیل کے کنارے بیٹھہ گئے- انور علی کا ایک ساتھی جراحی کا تجربہ رکھتا تھا- اس نے پٹی کھول کر لیگرانڈ کے زخم کا معائینہ کرنے کے بعد انور علی سے کہا "اگر آپ اجازت دیں تو میں گولی نکال کے زخم داغ دیتا ہوں-" اس نے لیگرانڈ کی نبض پرکھنے کے بعد کہا- "اگر انکا بخار اتنا تیز نہ ہوتاتو میرا کام آسان ہوتا-" (2-8) (اور تلوار ٹوٹ گئ- نسیم حجازی)

[tɒlʊeɪ] [əˈftɒb] [seɪ] [thəʊri] [deIr] [bɒd] [ɒeIk] [ʤhi:l] [keɪ] [kɒrIb] [pəʊñʧ] [kə(r)] [ʌnvə(r)] [ʌli] [neI] [ɒpneɪ] [sɒthIyəʊñ] [kəʊ] [rʊkneɪ] [kɒ] [hʊkum] [dIyɒ] [ɒəʊr] [ʊsneɪ] [ləgrɒñd] [kəʊ] [gəʊreɪ] [seɪ] [ʊtɒr] [kə(r)] [zɒmIñ] [pə(r)] [lɪtɒ] [dɪyɒ]. [bɒz] [sɪpɒyəʊñ] [neɪ] [theɪləʊñ] [seɪ] [bɒsi] [rəʊtiyɒñ] [nɪkɒli:ñ] [ɒəʊr] [sɒthIəʊñ] [meɪñ] [tʌksIm] [ki:ñ] [ɒəʊr] [vəʊh] [ʤhi:l] [keI] [kɒrIb] [bæIth] [gɒyeɪ] [əˈnvə(r)] [ɒli] [kɒ] [ək] [sɒthi] [ʤɒrɒhi] [kɒ] [tɒʤɒrbɒ] [rʌrktɒ] [thɒ] [ʊsneɪ] [pʌti] [khəʊl] [kə(r)] [ləgrɒñd] [keɪ] [zʌxʌm] [kɒ] [mʊ:ænɒ] [kə(r)neI] [keɪ] [bɒd] [ʌnvə(r)] [ʌli] [seɪ] [kɒhɒ] "[agə(r)] [a:p] [Iʤɒzʌt] [deɪñ] [təʊ] [mæñ] [gəʊli] [nIkɒl] [kə(r)] [zɒxʌm] [dɒg] [deItɒ] [hʊñ]. [ʊsneɪ] [ləgrɒñd] [ki] [nʌbz] [pɒrʌkhneɪ] [keɪ] [bɒd] [kɒhɒ] "[agə(r)] [ɪnka] [bʊ:xɒr] [ɪtnɒ] [teɪz] [nɒ] [həʊtɒ] [təʊ] [meɪrɒ] [kɒm] [a:sɒn] [həʊtɒ]"

A little after the sun rise, when they reached the lake Anwer Ali ordered his colleagues to stop and laid Legrand on the ground taking him from horseback. Some soldiers took the dried bread from bags and distributed them among other soldiers and sat on the bank of the lake. One friend of Anwer Ali had the experience of surgery. He asked Anwer Ali after inspecting the wounds of Legrand, " if you permit me , I can do the surgery after taking out the bullet from his body". The friend further added, "Had his fever not this much the job would have been easier".

In discourse 2.8, the preferred antecedents for اس نے ([ʊsneI]) are lying in the same or in the immediate previous sentence. Similarly, in (2.3), the antecedents for third person ADs اس نے، اسے، اسکو، اسکی ([ʊsneI], [ʊseI], [ʊskəʊ], [ʊskɪ]) are resolved on the basis of distance.

2.6 Subject and object preference

In Urdu, especially for the resolution of personal ADs (first person, second person and third person), subject and object preference plays a very important role. Consider the example

انور علی نےخط کا مضمون پڑھنا شرع کیا۔ مراد علی نےلکھا تھا -"بھائ جان اسلام علیکم۔ میں سرحد کی دفاعی چوکیوں کے معائنے کیلیے گیا ھوا تھا، اسلیےآپ اور بھابی جان کے خطوط کا جواب نہ دے سکا۔ مجھے ایک مہینے کی چھٹی مل گئ ھے لیکن میں گھر آنے سے پہلے چچا اکبر خان کے پاس جانا چاھتاھوں (2-9) (اور تلوار ٹوٹ گئ- نسیم حجازی)

[ʌnwə(r)] [ʌli] [neI] [xʌt] [kɒ] [mʌzmʊn] [pə(r)nɒ] [ʃʊrʊ] [kIyɒ]. [mʊrɒd] [ʌli] [neI] [lIkhɒ] [thɒ] "[bhɒi] [ʤɒn]! [əˈsɒleImuleIkʊm] [mæñ] [sə(r)həd] [ki] [dIfɒyi] [ʧəʊkIyʊñ] [keI] [mʊæneI] [keIlIyeI] [gIyɒ] [hʊvɒ] [thɒ] [IslIyeI] [a:p] [ɒəʊr] [bhɒbi] [ʤɒn] [keI] [xɒtʊ̃t] [kɒ] [ʤɒvɒb] [nɒh] [deI] [sʌkɒ]. [mu:ʤheI] [æk] [mɒhIneI] [ki] [ʧʊti] [mIl] [gɒi] [hæ] [lækIn] [mæñ] [ghə(r)] [a:neI] [seI] [peIhleI] [ʧʌʧa:] [əˈkə(r)] [xɒn] [keI] [pɒs] [jɒnɒ] [ʧɒhtɒ] [hʊñ]

Anwer Ali started reading the letter. Murad Ali had written "my brother! Regards, I have gone for the inspection of defense posts. Therefore, I was unable to send reply to yours and your Mrs. letters. I have got leave for one month. However, before coming home I want to visit uncle Akber Khan"

Discourse 2.9, consists of frequent use of first person anaphoric devices مجھے، میں ([mu:ʤheI], [mæñ]). Discourse 2.9 is in the form of direct speech. In such type of discourse, for resolution of first person anaphoric devices highest, preference will be given to subject of the main clause i.e. the clause just before the reported speech starts. مراد علی ([mʊrɒd] [ʌli]) is the subject of the main clause so all first person anaphoric devices will refer to مراد علی ([mʊrɒd] [ʌli]). Similarly, in case of second person anaphoric devices, object preference will be the highest.

پریا نے راج سے کہا – "تم کیو ں رو رھے ھو- میں تمھا رے ساتھہ ھوں- ھمیشہ تمھارا ساتھہ نبھاؤں گی-"(2-10)

[prɪya] [neɪ] [rɒʤ] [seɪ] [kɒhɒ] "[tʊm] [kIyʊñ] [rəʊ] [rɒheI] [həʊ] [mæñ] [tʊmhɒreI] [sɒth] [hʊñ] [hɒmeIʃɒ] [tʊmhɒrɒ] [sɒth] [nɪbhɒʊñ] [gi]".

Priya said to Raj,"Why are you weeping, I am with you and will always be with you"

عمرو نے عمارہ کو کہا-"عورت کو کہنا تمہیں / تمکو نجاشی کا تیل لگائے-جو دوسرا کوئ بھی نہیں لگا سکتا-" (نقوش- رسول نمبر) (2-11`)

[ʊmru:] [neɪ] [ʌmɒrɒ] [kəʊ] [kɒhɒ] "[a:əʊrʌt] [kəʊ] [kəhnɒ] [tʊmhæñ]/ tʊumkəʊ] [nɒʤɒʃi] [kɒ] [teɪl] [lɒgɒeɪ] [ʤəʊ] [dʊ:sɒrɒ] [kəʊi] [nɒhɪñ] [lɒgɒ][sʌktɒ]"

Umroo asked Ammara "Ask the woman to massage you with the oil of Najashi that is not possible by any other".

جج ملزم سے-"تم نے بہت صفائ اور ھوشیاری سے جرم کیا ھے" ملزم جوابا جج سے- "شکریہ جناب آپ پہلےآدمی ھیںجنھوں نے میرے فن کی تعریف کی-" (2-12)

[ʤʌʤ] [mʊlzɪm] [seɪ] "[tʊm] [neɪ] [bɒhʊt] [sɒfɒyi] [ɒəʊr] [həʊʃɪɒri] [seɪ] [ʤʊrm] [kIyɒ] [hæ]. [mʊlzɪm] [ʤavabʌn] [ʤʌʤ] [seɪ] "[ʃʊkrɪɒ] [ʤɒnɒb] [a:p] [pəhleɪ] [a:dmi] [hæñ] [ʤInhəʊñ] [neɪn] [mæreɪ] [fʌn] [ki] [tɒrɪf] [ki]"

Judge said to the accused, "you did the crime very professionally and cleverly". Accused replied "thanks sir, you are the first person who praised my expertise".

Again, 2.10, 2.11 and 2.12 are in the form of direct speech. In all above discourses, second person ADs آپ ،تمھارا ،تمھیں ،تمکو ،تو ،تم ([tʊm], [tʊ], [tʊmkəʊ], [tʊmhæñ], [tʊmhɒrɒ], [tʊmhɒri],[a:p]) have direct objects such as جج ،ملزم ،عمارہ ([ʌmɒrɒ] [mʊlzIm], [ʤʌʤ]) of the main clause as their potential antecedents.

Here is an example in which for the resolution of third person anaphoric device وہ ([vəʊh]), potential antecedents are found using subject preference filter.

لارڈ کارنوالس کو فیصلہ کن جنگ کیلئے ٹیپو سلطان کی تیاریوں کا علم تھا- وہ یہ جانتا تھا کہ موجودہ حالات میں جنگ کا طول دینا نقصان دہ ھو سکتا ھے- وہ جنگ کے آنے والے حالات کےبارے سوچتا تو پریشان ھونے لگتا- (2-13)(اور تلوار ٹوٹ گئ- نسیم حجازی)

[lɒrd] [kɒrnIvɒlIs] [kəʊ] [fæslɒ] [kʊn] [ʤʌg] [keIlIyeI] [ti:pʊ̃] [sʊltɒn] [ki] [tIyɒrIyʊ:ñ] [kɒ] [Ilm] [thɒ]. [vəʊh] [yəh] [ʤɒntɒ] [thɒ] [kəh] [məʊdʒʊdɒ] [hɒlɒt] [meIñ] [ʤʌg][kɒ] [tʊ̃:l] [deInɒ] [nʊksɒn] [dəh] [həʊ] [sʌktɒ] [hæ]. [vəʊh] [ʤʌñg] [keI] [a:neI] [vɒleI] [hɒlɒt] [keI] [bɒreI] [səʊʃtɒ] [təʊ] [prIʃɒn] [həʊneI] [lʌgtɒ].

Lord Kernevalis was aware of the preparations of Tipu Sultan about the final war. He knew that it will be quite dangerous to lengthen the war and he was worried to think about the results of the war.

Here, terminals of the sentence are لگتا، تھا ([lʌgtɒ], [thɒ]) that are used for personal singular and masculine NP, but the problem is that لارڈ کارنوالس ، ٹیپو([lɒrd] [kɒrnIvɒlIs], ([tIpʊ]) both are personal, singular and masculine NPs. So the question arises that وہ ([vəʊh]) refers to which NP in the preceding

sentence. Here, the subject preference will be high. So, وہ ([vəʊh]) refers to لارڈ کارنوالس.

2.7 NP followed by certain words

Certain NPs in Urdu discourse are followed by words کےمتعلق، کےبارے، کی طرف ([keI][mʊtalək]], [keI][bareI], [kI][tə(r)f]). In such circumstances, these NPs will be given highest priority to become the antecedents. For example,

جہانگیربدرنے اپنی بیٹی کے بارے بتایا-کہ اسے/اسکو سیاست کا کوئ شو ق نہیں ،ہاں اسے اعلی تعیم کا شوق ھے-اس نے ماسٹرز کرنے کا ارادہ کررکھا ھے-(2-14)

(Interview with Jehangir Badar)

[ʤɔhɔñgi:r] [bʌdə(r)] [neɪ] [ʌppni] [beɪti] [keɪ] [bɔreɪ] [btɔyɔ]] [kəh] [ʊseɪ/ʊskəʊ] [sɪyɔsʌt] [kɔ] [kəʊei] [ʃəʊk] [nɔhIñ] [hɔñ] [ʊseɪ] [a:lɔ] [ta:lɪm] [kɔ] [ʃəʊq] [hæ]. [ʊsneɪ] [ma:stə(r)z] [kə(r)neI] [ka:] [æra:dha:] [kə(r)] [rʌkhɔ] [hæ].

Jihangir Badder told about his daughter that she has no interest in politics, However, she is interested in higher education. She has the intention to do her masters degree.

ماں سلمی کی طرف دیکھہ دیکھہ کے قربان ھو رہی تھی کیو نکہ وہ دکھہ رہی تھی-(2-15)بہت بھلی

[mɑ:ñ] [sʌlmɔ] [ki] [tə(r)f] [deIkh] [deIkh] [keI] [kʊrbɔn] [həʊ] [rahi] [thi] [kIyʊkəh] [vəʊh] [bɔhʊt] [bhʌli] [dhIkh] [rɔhi] [thi].

The Mother was looking towards Salma very lovingly since she seemed very beautiful.

It is the سلمی who is looking beautiful not the ماں ([mɑ:ñ]), since سلمی ([sʌlmɔ]) is followed by certain class of words.

3 Implementations and evaluations

An informal algorithm for the resolution of first person anaphoric devices is as follows:

1. Examine the next clause in the discourse. If no clause exists then finish.
2. If the current clause consists of first person anaphoric devices then go to step-3 else go to step-1.
3. Access the previous clause.
4. If the current clause consists of section headings, noun phrase followed by certain words then assign weight to these filters else assign priority to noun or noun phrase appearing as a subject of the clause.
5. If no subject exists then go to step-3.

Similarly, an informal algorithm for the resolution of second person ADs is as follows:

1. Examine the next clause in the discourse. If no clause exists then finish.
2. If the current clause consists of second person anaphoric devices then go to step-3 else go to step-1.
3. Access the previous clause.
4. If the current clause consists of topicalized structures then assign weight to these filters else assign priority to noun or noun phrase appearing as an object of the clause.
5. If no object exists then go to step-3.

In the same way an informal algorithm for the resolution of third person ADs is as follows:

1. Examine the next clause in the discourse and if no clause exists then go to step-9.
2. If the current clause consists of third person anaphoric devices then go to step-3 else go to step 1.
3. Access the previous clause.
4. Apply the lexical and morphological filters to assign the weight to nouns or noun phrases that follow the morphological and lexical filters.
5. If current clause consists of section headings or topicalized structures or noun phrase preceded / followed by certain class of words then assign the weight of these filters.
6. If current clause consists of noun or noun phrase as subject and objects (direct, indirect) then assign the weight value for these filters.
7. If the current clause does not consists noun or noun phrase as subject, object or contains no section headings, topicalized structures and noun phrase preceded by certain words then go to step- 3.
8. Find the repetitions of all noun or noun phrases and increment their corresponding weights for each repetition.
9. Record the results and Finish

Algorithms are implemented in Visual C++. Implemented algorithm gets the input that is constructed manually. For this purpose each discourse is divided into clauses and is stored as Unicode text file for input to anaphora resolution

program. For better understanding, consider the example of discourse 2.8 and its division into clauses.

```
clause(sub(انور علی,sng,msc),dob(ساتھیوں,plu,msc),vb(sng,msc)).
clause(sub(اس),dob(لیگرانڈ,sng,msc),vb(sng,msc)).
clause(sub(سپاھیوں,plu,msc),dob(روٹیاں,fem,plu),vb(plu,msc)).
clause(sub(nill),dob(ساتھیوں,plu,msc),vb(plu,msc)).
clause(sub(وہ),dob(جھیل,sng,fem), vb(plu,msc)).
clause(sub((ساتھی,sng,msc),dob(جراحی,sng,msc),vb(sng,msc)).
clause(sub(اس),dob(لیگرانڈ,sng,msc),vb(sng,msc)).
clause(sub(ساتھی,sng,msc),dob(انور علی,msc,sng),vb(sgn,msc)).
clause(sub(آپ),dob(nill),vb(plu,msc)).
clause(sub(میں),dob(زخم,msc,sng),vb(sng,msc)).
clause(sub(اس),dob(لیگرانڈ,sng,msc),vb(sng,msc)).
clause(sub(انکا),dob(بخار,msc,sng),vb(sng,msc)).
clause(sub(nill),dob(میرا),vb(sng,msc)).
```

Fig 1

Table-1, Table-2 and Table-3 show the order of weights assigned to various filters for the resolution of first person, second person and third person anaphoric devices. The implemented algorithm aims to determine the efficiency in terms of accuracy and reliability of the proposed order of factors. For this purpose various experiments were conducted over various text genres. To evaluate the success rate of every experiment, *precision is* calculated as defined below. The average length of each discourse in sentences was 4-6.

Precision = *Number of correctly resolved anaphors / Number of anaphors attempted to be resolved*

The results of the three experiments are as follows

Experiment#	Precision
1	78%
2	80%
3	80%

Table-1 shows that in case of first person anaphoric devices the priority has been assigned on the basis of section heading, noun phrase followed by certain words and then subject. It means that if no section heading or noun phrase followed by certain words are present then the subject in the main or previous clause will be the potential antecedent for first person anaphoric devices. Similarly, Table-2 for second person anaphoric devices, exhibits that weights will be assigned in descending order (left – right). It means that the leftmost filter that is topicalized structure will get the highest weight for second person ADs. Consider the following output (Fig-2) produced by anaphora resolution program, for the resolution of second person anaphoric device آپکا in the discourse 2.4, topicalized structure کھر صاحب gets high priority to become the antecedent.

Clause 1, SUB (آپکا) RESTO (2 کھر صاحب)

Fig 2

Again, in case of third person anaphoric devices weights as shown in Table-3 have been assigned in descending order (top - bottom). It means the weight of section heading filter will be larger in value than that of subject filter. Consider a noun or noun phrase which is section heading as well as a repeated noun and also lexical filter applies on it. For this noun or noun phrase all the weights will be summed up. A noun with highest weight will be given preference to become the antecedent for third person anaphoric device. This is demonstrated by the following output generated for discourse (2.8) by our anaphora resolution system. This discourse contains total 13 clauses from 0 – 12. Clause 1 contains third person anaphoric device اس ([ʊs]) that is resolved to انور علی which is assigned weight 12 on the basis of lexical filter and distance preference, so, ساتھیوں is ruled out to become the antecedent since its weight is 1. Similarly for the third person anaphoric device وہ, that appears in clause 4, antecedent with highest weight 50 is ساتھیوں. By the same token, for the resolution of the first person anaphoric device میرا, preference has been given to the noun ساتھی (Fig-2) that is the subject in the previous clause.

```
clause  1 , SUB ( اس )     RESTO  ( 12 انور علی )( 1 ساتھیوں )
clause  4 , SUB ( وہ )     RESTO  ( 50 ساتھیوں )( 12 انور علی )( 11 سپاھیوں )( 11 لیگرانڈ )( 11 روٹیاں )
clause  6 , SUB ( اس )     RESTO  ( انور علی )( 12 ساتھیوں )( 7 لیگرانڈ )( 7 جھیل )( 7 ساتھی )
clause  8 , SUB ( آپ )     RESTO  (5 انور علی )( 3 ساتھیوں )( 3 لیگرانڈ )( 3 ساتھی )( 1 روٹیاں )( 1 سپاھیوں )
clause  9 , SUB ( میں )    RESTO  ( 14 ساتھی )( 6 لیگرانڈ )( 2 جھیل )( 1 سپاھیوں )( 0 روٹیاں )( 0 ساتھیوں )
clause  10 , SUB ( اس )    RESTO  ( 31 ساتھی )( 30 لیگرانڈ )( 12 ساتھیوں )( 11 زخم )( 7 جھیل )( 2 رسپاھیوں )
clause  11 , SUB ( انکا )  RESTO  ( 49 لیگرانڈ )( 30 ساتھیوں )( 13 ساتھی )( 8 سپاھیوں )( 7 روٹیاں )
clause  12 , DOB ( میرا )  RESTO  ( 14ساتھی )( 10 لیگرانڈ )( 7 انور علی )( 2 جھیل )( 2 زخم )( 2 بخار )
```

Fig 3

Algorithms fail to correctly resolve the anaphora for discourses as follows

پرویز مشرف نے نواز حکومت برخواست کی تو انہوں نے ان کے خلاف چارج شیٹ جاری کی-(3-15)

[pə(r)veɪz] [mʊʃʌrf] [neI] [nɒvɒz] [hɒkʊmət] [bə(r)kxʊɒst] [ki] [təʊ] [ʊnhəʊñ] [neI] [ʊnkeI] [xIlɒf] [ʧɒrʤ] [ʃi:t] [ʤɒri] [ki].

Pervaiz Musharaf when expelled Nawaz Government. He issued the charge sheet against him.

In the above discourse, the anaphoric device انہوں ([ʊnhəʊñ]) is resolved correctly to have antecedent پرویز مشرف ([pə(r)veɪz] [mʊʃʌrf]) on the basis of distance and subject preference filter but ان (ʊn]) is not resolved correctly to have antecedent نواز ([nɒvɒz]).

Third Person ADs	وہ [vəʊh]	اس، اسکا، اسکی، اسکے،اسکو،اسے [ʊs], [ʊskɒ], [ʊski], [ʊskeI], [ʊskəʊ], [ʊseI]	ان، انکا، انکی، انکے،انکو، انہیں [ʊn], [ʊnkɒ], [ʊnki], ʊnkeI], [ʊnkəʊ],[ʊnhæñ]
Lexical Information (AD refers to)	3rd Person, Singular, Plural, Masculine, Feminine	3rd Person, Singular, Masculine, Feminine	3rd Person, Plural, Masculine, Feminine
Priority Order **Weights assigned from top to bottom (Descending Order)**	Lexical Filter	Lexical Filter	Lexical Filter
	Section Heading	Section Heading	Section Heading
	Topicalized Structure	Topicalized Structure	Topicalized Structure
	Noun Phrase followed by certain words	Noun Phrase followed by certain words	Noun Phrase followed by certain words
	Subject	Distance	Distance
	Object	Subject	Subject
	Repetition	Object	Object
		Repetition	Repetition

Table 1: Priority Order for First Person ADs

Second Person Anaphoric Devices	**Priority Order (Left to Right)**	
تم، تو [tʊ], [tʊm]	Topicalized Structure	Object
تمہیں، تمکو [tʊmhæñ], [tʊmkəʊ]	Topicalized Structure	Object
تمہاری [tʊmhɒri]	Topicalized Structure	Object
تمہارا [tʊmhɒrɒ]	Topicalized Structure	Object
تمہارے [tʊmhɒreI]	Topicalized Structure	Object
آپ [a:p]	Topicalized Structure	Object
آپکو [a:pkəʊ]	Topicalized Structure	Object
آپکی [a:pki]	Topicalized Structure	Object
آپکا [a:pkɒ]	Topicalized Structure	Object
آپکے [a:pkeI]	Topicalized Structure	Object

Table 2: Priority Order for Second Person ADs

First Person Anaphoric	**Priority (Left – Right)**		
میں [mæñ]	Section heading	Noun Phrase Followed by Certain words	Subject
مجھے [mʊʤæ]	Section heading	Noun Phrase Followed by Certain words	Subject
مجھکو [mʊʤkəʊ]	Section heading	Noun Phrase Followed by Certain words	Subject
میرا [mærɒ]	Section heading	Noun Phrase Followed by Certain words	Subject
میری [mæri]	Section heading	Noun Phrase Followed by Certain words	Subject
میرے [mæreI]	Section heading	Noun Phrase Followed by Certain words	Subject
ہم [hʌm]	Section heading	Noun Phrase Followed by Certain words	Subject
ہمیں [hʌmæñ]	Section heading	Noun Phrase Followed by Certain words	Subject
ہمکو [hʌmkəʊ]	Section heading	Noun Phrase Followed by Certain words	Subject
ہمارا [hɒmɒrɒ]	Section heading	Noun Phrase Followed by Certain words	Subject
ہماری [hɒmɒri]	Section heading	Noun Phrase Followed by Certain words	Subject
ہمارے [hɒmɒreI]	Section heading	Noun Phrase Followed by Certain words	Subject

Table 3: Priority Order for Third Person ADs

4 Conclusion

One central question addressed in this paper is to determine the optimal order of the factors to find the preferred antecedents for the personal ADs in Urdu text. Rule based algorithms for the resolution of personal anaphoric devices are presented which are capable of resolving these anaphoric devices with 78-80% success rate in all kind of text genres. This success rate can be increased with improvement in certain rules especially for third person anaphoric devices.

References

M.N. Ali, M.A. Khan, and M. Aamir. 2007. Computational Treatment of Demonstrative Pronouns in Urdu. *In Proceedings of International Conference on Language and Technology CLT07*, 25-31. Bara Gali Summer Campus, Pakistan.

C. Aone , S. Bennett. 1996. Applying Machine Learning to Anaphora Resolution. In Wermter, S., Riloff, E., Scheler, G. (Eds) *Connectionist, Statistical and Symbolic approaches to learning for NLP*, 302-314. Springer, Berlin.

S. Brenan, M. Friedman and C. Pollard. 1987. A *25th Annual Meeting of the ACL*, 155-162. Stanford, Ca, USA,.

N. Ge, J. Hale and E. Chaniak. 1998. A Statistical Approach to Anaphora Resolution. *Proceeding of the workshop on very large Corpora,* 161-171. Montreal, Canada.

B. Grosz, J. Aravind and S. Weinstein. 1995. Centering a framework for modelling local coherence of discourse. *Computational Linguistics,* 21 (2), 203-225.

M. Halliday, R. Hassan. 1976. Cohesion in English. Longman, London.

B. Kalsoom, B. Rashida. 1993. Urdu Anaphora Resolution in Monologue. *M.Sc. Computer Sc. Thesis*, *Department of Computer Science University of Peshawar,* NWFP, Pakistan.

M.A. Khan , M.N. Ali and M. Aamir. 2006. Treatment of Pronominal Anaphora in Urdu Discourse. *In Proceedings, IEEE, ICET Conference on Emerging Technologies,* 543-548. Peshawar, Pakistan.

S. Lappins, H. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4), 535-561.

J. MaCarthy, W. Lehnert. 1995. Using decision trees for coreferences resolution. *Proceedings of the 14th International Conference on AI,* 1050-1055. Montreal, Canada.

R. Mitkov. 1998. Robust Pronoun Resolution with Limited Knowledge. *Proceedings of 17th International conference on Computational Linguistics*, 869-875. Montreal, Canada..

R. Prasad, M. Strube. 2000. Discourse Salience and Pronoun Resolution in Hindi. *Penn Working Papers in Linguistics*, Vol. 6.3, 189-208.

L. Sobha. 1998. Anaphora Resolution in Malayalam and Hindi. *Doctorial dissertation submitted to Mahatma Gandhi University*, Kottayam, Kerala.

W.M. Soon, H.T. Ng, and C.Y. Lim. 1999. Corpus based learning for noun phrase coreference resolution. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and in very large Corpora,* 285-291. University of Maryland, USA.

M. Strube. 1998. Never look back: An alternative to Centering. In *Proceedings of the 17th Int. Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* 1251–1257. Montreal, Quebec, Canada.