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 اناکہ، اناکے، اناکی، ان، اسکی، اسکے، اسکیاں، اسکےاں، اناکےاں, انکےاں, انکے, انکی, اسکی, اسکیاں, اسکےاں, اناکیاں, اناکےاں, انکیاں, انکےاں. Examples of second person ADs are تما، تم، تماں، تمہاری، تمو، تمہاری، تمہارے، تمہاریں, تمہارے, تم، تمہاری، تمہارے, تمہاریں, تمہارے. Examples of third person ADs are اناکہ، اناکے، اناکی، ان، اسکی، اسکے، اسکیاں، اسکےاں, اناکےاں, انکےاں, انکی، اسکی، اسکیاں، اسکےاں, اناکیاں، اناکےاں، انکیاں، انکےاں.

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. ©2008 Asian Federation of Natural Language Processing
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 topological 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. ملکی نے چاریہ کا پڑا ساگا گشتی اور آگے جلد، فصل دینے کا پڑا ملکی کا پڑا یکجا نہیں لیکن نہ جانتی کیونہ وہ اندر ہو (سلامی کی کچھ، سامنا پاس ہو 20ے جر کر ۲ گنا نہیں (mOLakh) [neI] [fDei] [ha(r)] [D] [oD] [ghAThO] [D:THO] [bDAR] [a:ge] [fn] [D]. [fHzl] [Dn] [neI] [a:geI] [bha(r)] [b] [r] [mOLakh] [ko] [bD] [gAke(r)] [DO] [fn] [THO] [leIk] [nd] [gDPn] [kYn] [vEh] [Anda(r) [sakr] [b] [r] [eH] [gyp] [THO]. Mulkhi took the bundle of grass and moved ahead. Fazal Din had come forward to catch the arm of Mulkhi, but he did not have the courage to do so.

In Urdu, the word وہ (vOH) 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 ([THA]) that indicates the third person AD وہ (vOH) refers to singular and masculine NP i.e. ملکی (mOLakh) will be ruled out to become the antecedent. Similarly, consider the example ایک لہار وہ نے ہو وہ ملکی ہو وہ ملکی ہو (1) (اربیل، بارنی فیس) [s:m] [xAT] [pa:r] [ke] [a:pe] [hauf] o [hDvE] [khu] [beIn] [kaf] [ar] [UkD].
Since the terminal of sentence is (gDi), so it means (voh) refers to some feminine antecedent that is in the above text. The lexical filters are used to resolve anaphora on the basis of number and gender information. For example

After reading the letter Aneela lost her senses. She was treated for some feminine antecedent that is in the above text. The lexical filters are used to resolve anaphora on the basis of number and gender information. For example

Faint! What should I do with your memories? Everything vanishes with the passage of time but your memories are like unripe grain which needs my blood to flourish.

In 2.4, the word (ap) refers to topicalized structure (kho[r]). Similarly, in discourse 2.5 ADs (tumhri) refer to topicalized structure (kho[r]). 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
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 منتو (muntū) appears repeatedly. So, on the basis of repetition, it will be the potential antecedent for most of personal pronouns e.g., اس (as), اسے (asē), اسکی (askī), اسکو (asko), اسی (asī), اسے (asē), اسے (asē), اسے (asē), اسے (asē) 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

In the above discourse, شعبہ اختر (sho'ab al-ḥakīm) 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.
2.6 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

In describing the relationship between subject and object, the masculine personal anaphoric device is preferred over the feminine. The reason for this preference is not explicitly stated, but it is implied that the masculine form is more appropriate in this context.

Discourse 2.9, consists of frequent use of first person anaphoric devices (commonly referred to as "subject pronouns"). 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. 

Maryal Ali (ماجدہ) is the subject of the main clause so all first person anaphoric devices will refer to Maryal Ali (ماجدہ). Similarly, in case of second person anaphoric devices, object preference will be the highest.

Here, terminals of the sentence are [ناپیچت] and [thd] that are used for personal and masculine NP, but the problem is that [ناپیچت], [ناپیچت], ([ئی]) both are personal, singular and masculine NPs. So the question arises that [ناپیچت] refers to which NP in the preceding sentence?
sentence. Here, the subject preference will be high. So, وہ ([\text{\text{wøh}}]) refers to لارہ کارنوالے.

2.7 NP followed by certain words

Certain NPs in Urdu discourse are followed by words کے متعلق، کی، کا طرف ([\text{\text{kəI}]} [\text{\text{mʊtɒlsək}}], [\text{\text{kəI}]} [\text{\text{bareI}}], [\text{\text{kI}}] [\text{\text{tə(r)ʃ}}]). In such circumstances, these NPs will be given highest priority to become the antecedents. For example,

Jeangir Badar told about his daughter that she has no interest in politics, اسے اسکو سیاست کا کوئی شوق نہیں. (Interview with Jeangir Badar)

\text{یہ} [\text{\text{dɪŋgi:r}}] [\text{\text{bʌdə(r)}}] [\text{\text{neI}}] [\text{\text{apppni}}] [\text{\text{betɪ}}] [\text{\text{ker}}] [\text{\text{bəzər}}] [\text{\text{bʊtydʊ}}] [\text{\text{kəs}}] [\text{\text{ʊzər/ʊskæʊ}}] [\text{\text{sɪdəsɪt}}] [\text{\text{kəʊel}}] [\text{\text{ʃʊk}}] [\text{\text{nɒdɪn}}] [\text{\text{hɒr}}] [\text{\text{ɒsər}}] [\text{\text{ələ}}] [\text{\text{təlɪm}}] [\text{\text{kʊ}] [\text{\text{ɒwɒg}}] [\text{\text{həw}}].

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

It is the سالمی who is looking beautiful not the سالمی ([\text{\text{sɒlmʊ}}]), since سالمی ([\text{\text{sɒlmʊ}}]) 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 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.
6. Find the repetitions of all noun or noun phrases and increment their corresponding weights for each repetition.
7. 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.

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 topocalized 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 آپاکا کهر صاحب gets high priority to become the antecedent.

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 اسم ([ذکر]) that is resolved to which is assigned weight 12 on the basis of lexical filter and distance preference, so, it is ruled out to become the antecedent since its weight is 1. Similarly for the third person anaphoric device ج، that appears in clause 4, antecedent with highest weight 50 is سانتوہيون which is exist in clause 3. By the same token, for the resolution of the first person anaphoric device میرا، preference has been given to the noun سانتوہيون (Fig-2) that is the subject in the previous clause.

Algorithms fail to correctly resolve the anaphora for discourses as follows

87
Pervaiz Musharaf when expelled Nawaz Government. He issued the charge sheet against him.

In the above discourse, the anaphoric device ([nɒhan]) is resolved correctly to have antecedent ([pə(r)veɪz] [mɒjˈər]) correctly to have antecedent ([nɒhan]) on the basis of distance and subject preference filter but an ([ʂə]) is not resolved correctly to have antecedent ([nɒdz]).

### Table 1: Priority Order for First Person ADs

| Priority Order Weights assigned from top to bottom (Descending Order) | Lexical Information (AD refers to) | 3rd Person, Singular, Masculine, Feminine | 3rd Person, Singular, Masculine, Feminine | 3rd Person, Plural, Masculine, Feminine |
|------------------------------------------------------------------------|-----------------------------------|------------------------------------------|------------------------------------------|------------------------------------------|
| Lexical Filter                                                         | Section heading                   | Section heading                          | Section heading                          | Section heading                          |
| Topicalized Structure                                                  | Noun Phrase followed by certain words | Noun Phrase followed by certain words | Noun Phrase followed by certain words | Noun Phrase followed by certain words |
| Noun Phrase followed by certain words                                  | Distance                           | Distance                                 | Distance                                 | Distance                                 |
| Object                                                                 | Subject                            | Subject                                  | Subject                                  | Subject                                  |
| Repetition                                                             | Repetition                          | Repetition                               | Repetition                               | Repetition                               |

### Table 2: Priority Order for Second Person ADs

| Second Person Anaphoric Devices | Priority Order (Left to Right) |
|---------------------------------|--------------------------------|
| تامہرائ (tum)                    | Topicalized Structure           |
| تامہرائ (tum)                    | Object                          |
| تامہرائ (tum)                    | Topicalized Structure           |
| تامہرائ (tum)                    | Object                          |
| تامہرائ (tum)                    | Topicalized Structure           |
| تامہرائ (tum)                    | Object                          |
| تامہرائ (tum)                    | Topicalized Structure           |
| تامہرائ (tum)                    | Object                          |

### Table 3: Priority Order for Third Person ADs

| First Person Anaphoric | Priority (Left – Right) |
|------------------------|-------------------------|
| سی [mənr]              | Section heading          |
| جی [mədɜ]              | Noun Phrase Followed by Certain words |
| جی [mədɜ]              | Subject                  |
| جی [mədɜ]              | Section heading          |
| جی [mədɜ]              | Noun Phrase Followed by Certain words |
| جی [mədɜ]              | Subject                  |
| جی [mədɜ]              | Section heading          |
| جی [mədɜ]              | Noun Phrase Followed by Certain words |
| جی [mədɜ]              | Subject                  |
| جی [mədɜ]              | Section heading          |
| جی [mədɜ]              | Noun Phrase Followed by Certain words |
| جی [mədɜ]              | Subject                  |
| جی [mədɜ]              | Section heading          |
| جی [mədɜ]              | Noun Phrase Followed by Certain words |
| جی [mədɜ]              | Subject                  |
| جی [mədɜ]              | Section heading          |
| جی [mədɜ]              | Noun Phrase Followed by Certain words |
| جی [mədɜ]              | Subject                  |
| جی [mədɜ]              | Section heading          |
| جی [mədɜ]              | Noun Phrase Followed by Certain words |
| جی [mədɜ]              | Subject                  |
| جی [mədɜ]              | Section heading          |
| جی [mədɜ]              | Noun Phrase Followed by Certain words |
| جی [mədɜ]              | Subject                  |
| جی [mədɜ]              | Section heading          |
| جی [mədɜ]              | Noun Phrase Followed by Certain words |
| جی [mədɜ]              | Subject                  |

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 kinds 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 Wermer, 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. McCarthy, 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*
