ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Indian Journal of Science and Technology, *Vol 10(3), DOI: 10.17485/ijst/2017/v10i3/110637, January 2017*

A Computational Model for Resolving Arabic Anaphora using Linguistic Criteria

Abdullatif Abolohom and Nazlia Omar*

FTSM, University Kebangsaan Malaysia, 43600 Bangi, Malaysia,
allatif4@yahoo.com, mo@ftsm.ukm.my

Abstract

Anaphora resolution is seen to be a very challenging and complex problem in the NLP. A majority of the NLP applications used for question answering, information extraction, and text summarisation, need a proper resolution and identification of the anaphora. Despite the fact that several authors have published studies for anaphora resolution in many European languages, including English, very few studies have been published for anaphora resolution in the Arabic language. In our study, we have proposed a novel model for the Arabic pronominal anaphora resolution. Our model contains several steps. In the first step, we have identified the pronouns and removed the non-anaphoric pronouns. In the second step, we have identified a list of the candidates from the context around the anaphora. Lastly, we selected the most probable candidates for every identified anaphoric pronoun. In our study, we have determined the proper rules which can be used for this task. The different linguistic rules depend on the morphological, lexical, heuristic, syntactic, and the positional constraints. We have assessed the performance of our proposed model using the Quran corpus, which was annotated with the pronominal anaphora. Our experimental results indicated that our proposed model was able to yield good results and could also choose the appropriate antecedents with 84.43% accuracy.

Keywords: Anaphora Resolution; Linguistic Rule, Rule Based Method

1. Introduction

In the area of Natural Language Processing (NLP), the anaphora resolution refers to the determination of the earlier entities (i.e., antecedent) to which the specified noun phrase (anaphora) is referring to. Since a very long time, the anaphora resolution is considered as a very challenging but important task for many of the NLP applications, like question-answer, text summarisation, and the machine translations. In general, anaphora refers to a technique which uses the abstract form for referring to a particular form of content and language, in a specific language environment. It is a commonly used phenomenon in NLP and plays an important role in the simplification of the expressions and connecting the words in a context[1]. Despite the fact that the anaphora resolution is actively applied in NLP research, there are very few studies which have used this task in the Arabic language. The most commonly used anaphora in Arabic includes the pronominal anaphora resolution, which takes place if the anaphoric words or phrases are reflexive pronouns, third person pronouns, or the possessive pronouns. The Arabic language has some specific characteristics, like the null pronominal, morphological complexity, and a free ordering, which further complicates the task of pronominal anaphora resolution. In this study, we have focused primarily on pronominal anaphora resolution of the noun phrase antecedents present in the Arabic language, wherein the term pronoun is used to cover the reflexive pronouns, 3rd person pronouns, and possessives. In this study,

*Author for correspondence

we have presented a novel computational model for the Arabic pronominal anaphora resolution. The paper has been structured in the following manner: in Section 2, we have highlighted many of the recent studies that have used the pronominal pronoun resolution. Section 3 describes the corpus of this work and the different linguistic rules that have been applied in this study. Section 3 presents the model developed, while the sections 4 and 5 present the experimental results obtained and the discussion, respectively. Finally, Section 6 presents the conclusions of our study.

2. Related Work

Numerous techniques have been suggested by different authors for applying the anaphora resolution in different languages, like English etc. In general, it can be seen that the computation processes for the anaphora resolution have led to two diverse but, still complementary directions. The two different directions include the knowledge-based (or rule-based) techniques[2-4] and the learning-based methods as described in[5-9]. In these knowledge-based techniques, the anaphora resolution process (or algorithm) contains rules and heuristics, which are based on the linguistic knowledge. On the other hand, the machine learning-based methods required a training data which contained a set of the feature vectors along with a set of the class labels. In their study[10], suggested a rule-based technique for the pronoun reference resolution in the Persian text documents. In their technique, the different rules were exploited for recognising the different pronoun references at every three-sentence interval. This resulted in a 90% performance rate by the system.

Further more[11], suggested a computational model for the pronominal anaphora resolution present in the Turkish language. Their suggested model was based on the Hobbs' Naïve Algorithm that exploited the surface syntax of the sentences present in a particular text. Another study[12] suggested a rule-based algorithm for resolving personal anaphora used in the Urdu text, that was seen to resolve the anaphoric devices and showed a success rate of 86%[13] suggested a rule-based technique for solving the personal pronoun resolutions present in the Pashto language. They tested their algorithm manually using around 210 personal pronouns which were resolved successfully out of 269 pronouns, thus, showing 78% accuracy. They carried out their test on the text derived from the published stories and novels. Also, a model was proposed by[14], for the anaphora resolution in the Malay text documents. It was seen that their model contained three elements, i.e., anaphora resolution process, syntactic knowledge process, and semantic-world knowledge process. All the three elements could be defined depending on the observable facts present in the Malay language. The problem of pronominal anaphora resolution in the Chinese language, including the zero pronouns, was addressed by[15]. They applied the syntactic rule-based pronoun resolution algorithm, i.e., the Hobbs algorithm, on the gold standard hand analyses of the Penn Chinese Treebank. They noted that their algorithm could achieve 77.6% accuracy for the apparent third-person pronouns and 73.3% accuracy for the zero pronoun resolution.

In their study[16] proposed a new rule-based technique for identifying the impersonal occurrence of the “ه” pronoun. Their system was a first step in the automated classification of the pronouns in the Arabic language, and they achieved very encouraging results (i.e., F-measure= 66.66%). However, their results, specifically the recall values (54.54%), showed that their rules were not very reliable indicators of the pronoun status. Though several anaphora resolution techniques have been suggested, the research, in this area, is still underway. However, the anaphora resolution in the Arabic language is severely lacking and there are very few studies that address the problem. In this study, we aim to address the issue of the anaphora resolution in the Arabic by applying a rule-based method.

3. Data Set and Linguistic Rules Used for Anaphora Resolution

The availability of annotated corpora with coreferential links is vital for anaphora resolution systems. However, there is few study carried out in the anaphoric corpus annotation for Arabic[17]. In this paper, we made use of a corpus from the Qurapro, annotated with antecedent references of anaphoras. The Qurpro is characterized by its relatively considerable number of anaphors labelled with antecedent referential information. In Table 1, we have shown the key statistics of the Qurpro corpus.

Table 1. Key statistics of the qurpro corpus.

Measure	Count
token	127795
pronoun	24679
3rd person pronouns	11544
sentences	6236

For determining the proper precursor of an anaphora and for disambiguating between them, many linguistic rules are employed which prefer certain candidates over others. The basis of the selection of such linguistic criteria is the linguistic perspective which takes into account anaphora referents and attributes of the Arabic language. Linguistically, the classification of the linguistic rules can be classified as: lexical, morphological, syntactic, and positional constraints. Candidates are assigned a score with a value-1, 0, 1, or 2 for each rule. These rules are described in Table 2.

Table 2. Dissipations of the Linguistic Rules and their Respective Scores

Linguistic rules	Description
Definiteness	A score of 1 is given if an NP is definite and of 0 if not.
Lexical Reiteration (LR)	A score of 2 is given if the NP repeated two or more times in which the pronoun occurs.
Recency	A score of +1 is assigned to the recency NP to the anaphora and0 if not.
Genitive object emphasis(GOE)	A score of 1 is given to NPs that are realized in the position of the genitive object
Section Heading (SH)	A score of 1 is assigned to the NPs which occur in the section heading where the pronoun appears; else, 0 is assigned.
Non-Prepositional Preference	A score of 1 is given to NPs that is not part of a prepositional phrase and 0 if not.
Patterns of sentences	2 if the candidate belongs to the considered patterns of sentences and 0 otherwise
Referential Distance(RD)	A score '+2' is assigned to NPs in the previous sentence or two sentences and further than those are given 0.
Term Preference (TP)	A score of +1 is assigned to NPs identified as representing terms in the genre of the text.
grammatical function. (GFNP)	Scores of +1 are given to an NP that has the same syntactic structure as the corresponding anaphora and 0 if not.
Nominal Predicate case (NPC)	Scores of +1 are given to the NPs in the nominal predicates of this type of sentences and0 if not.
First Noun Phrases(FNP)	A score of +1 is issued to the first NP of each sentence and 0 if not).
Closet Subject Nominative Case Preference	A score of 1 is allotted to an NP in case it is a subject in the same sentence as the anaphora.

4. Proposed model

We have developed a model to resolve Arabic anaphora resolution. The computational model comprises of several main modules that gradually operate the sets of pairs of anaphors and the candidate antecedents. The main modules are pre-processing module, anaphora identification, Identification of candidates, and Anaphora resolving module. The following sections discuss each of these modules.

4.1 Pre-processing

The pre-processing module comprises: POS tagging, NP chunking, and the identification of grammatical relation. The tasks are concisely described below.

4.1.1 Part-of-Speech Tagging

Part of Speech (POS) tagging is the procedure of allocating an opposite part of speech or word category to every word in the corpus. For our purposes, The Arabic Statistical POS Tagger (ASPOST) was utilised[18,19]. Its precision is 95.8%. ASPOST is capable of being trained on multiple Arabic corpora. The system included different approaches for smoothing and treatment of unidentified words.

4.1.2 NP Chucker

The second module of the pre-processing phase contains noun phrase identification. In Arabic, a noun phrase is a set of words which work together, starting with a noun, proper noun, conditional particle or pronoun. It may or may not be accompanied by a group of modifiers. The noun phrase chunker receives the input, forsakes the sequences for the POS-tagged tokens, and determines the NP boundaries. A set of the static NP patterns analyses the noun phrases present in the text. We have identified, approximately 200 NP patterns for determining the Arabic NPs. In Table 3, we have shown a small sample of the patterns. An 8.35% accuracy was seen for the noun phrase chunker[20]. The Arabic noun phrase chunker pertains to a rule-based chunker, whereas the NP patterns were ascertained on the basis of prior linguistic knowledge.

Table 3. A Sample of Arabic NPS patterns.

NP pattern	Example	English translation
REL V PRON CONJ V PRON DET N	الذين ءامن وا و عمل وا ال صلحت	Those who have believed and do good
PRON CONJ N PRON	انت و جوز ك	you and your spouse
DET N DET ADJ	ال مسجد ال حرام	Al-Masjid al-Haram (the holly mosque in Makkah)
COND V N PN	من يتعد حدود الله	and whoever exceeds the limits of Allah

4.1.3 *Grammatical Relations Extraction*

The Grammatical Relation (GR) refers to a linguistic relation that has been established using grammar, wherein the linguistic relation is a connection between the linguistic constituents or forms. In this stage, every sentence available in the search space underwent analysis and the grammatical relation (like subject and object) of every NP in the sentences was determined. The GR is an important feature in identifying the proper antecedent. Furthermore, a rule-based module was also applied for identifying the GRs. The GR extractor accuracy was seen to be 89.60%[21].

4.2 Anaphora Identification

The aims of this phase are to identification of pronouns and the elimination of the non-anaphora ones by removing the pleonastic. The identification of anaphora is carried out by referring to their grammatical parts of-speech. Third personal pronouns, reflexive, and possessive pronouns were marked based on to their occurrence within a sentence. The elimination of pleonastic pronouns consists on removing the pronouns that correspond to the specific model: adjective ال+انه من (It is + adjective) as for the example انه من المهم (It is important to).

4.3 Identification of Candidates

The probable candidates were chosen as antecedents of anaphora. Generally, all NPs before anaphora are considered as probable candidates for antecedents. In this phase, the candidates are filtered based on the gender and number agreements and a search scope.

4.3.1 *Gender and Number Filter*

The major aim of the gender and number filter was decreasing the size of the probable NP candidates. Here, we eliminated the candidate antecedents present in the candidate set which were incompatible to the anaphora, morphologically, prior to the actual resolution stage. Arabic anaphora and its antecedent are in agreement (with gender and number) except for the two cases. Plural of non-human entities and Plural of inanimate entities (male or female) which admit a singular female anaphora.

4.3.2 *Distance Filtering*

The Search Scope was used for determining the anaphora antecedent in the sentence boundary. Generally, the search scope is set for the preceding N sentences. The boundary limit depends on the type of anaphora that has to be identified. We observed that a majority of the NPs candidates were present in a 17-sentence window before pronouns, as shown in Figure 1. The X-axis indicates the distance present between the antecedents and pronouns in a sentence, while the Y-axis indicates the antecedent number which occurred in this distance. In Figure 1, we have shown the antecedents in the Qurapro corpus, which are present prior to the pronouns. It is seen that a majority of the antecedents are present in the same sentence (0) or in 3 earlier sentences, which is seen to rapidly decay.

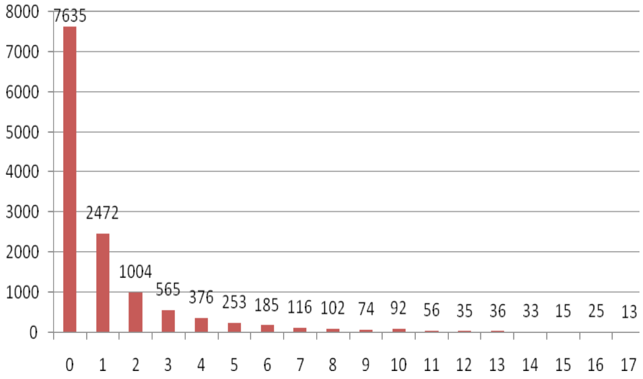

Figure 1. Distribution of antecedent distances of a pronoun.

4.4 Anaphora Resolving

Here, is the main part of the anaphora resolution model. Following the determination of the anaphors and removing and filtering the candidate list, we chose the most suitable antecedents for each anaphora from the listed most likely candidates. For each candidate, we employed a set of linguistic rules. All likely candidates were allotted a score value for each preference rule. We determined the scores for each rule and joined them to the preference score for each precursor. All candidates were graded and the one with the uppermost aggregate point score was suggested as the precursor. If several candidates showed similar point scores, the one nearest the anaphor was selected.

5. Results and Discussion

The assessment were carried out in two key segments: The first segment provides a gender and number filtering assessment, examining the number of pairs which are retained or filtered with regards to their morphological compatibility. The second segment is evaluating the proposed model. The performance of our model was evaluated using the following accuracy metric:

Accuracy = Number of correctly resolve anaphora / Number of all anaphoras

6. Experiments Results

The morphological filter aimed to decrease the candidate set size by eliminating the morphologically incompatible pairs. In our experiments, we observed that the resulting candidate set size (after the application of the morphological filter) was reduced to 17777 from 81174 pairs, which was an 88.1% reduction. In Table 4, we have shown the Precision and recall after applying this morphological filter.

Table 4. The result of morphological filter.

Precesion	Recall
% 85.21	90.20%

Table 5 summarizes the experimental results of the proposed Arabic anaphora resolution model. The results indicated that the proposed model obtained satisfactory results and could also select the right antecedents with 84.43% accuracy.

Table 5. Precision the result of the proposed anaphora resolution model.

Number of Anaphora	Correctly resolved	Incorrectly resolved	Out of scope	Percentage correct resolved
9235	7345	1409	481	84.43

Each rule has contribution to the model. To determine the contribution of each rule, we investigate how much the model loses on exclusion of a rule. Rule contributions are shown in Table 6. As the tables shows, the most important rules that form the “hard core” of the model, are Recency, Definiteness, Grammatical Function (GFNP), Frequency Indicator (LR), and Nominal Predicate Case (NPC). These rules gave realistically high contributions to the model; and thus, enhanced the model’s performance. The 5 least significant vital rules were Non-Prepositional Preference, Patterns of Sentences (PS), Referential Distance (RD), First Noun Phrases (FNP), and Closet Subject. These rules only influenced the model slightly; therefore, a small reduction in model performance. Genitive Object Emphasis (GOE), Section Heading (SH) Preference, and Term Preference (TP) rules, all influenced the model insignificantly. This model could be improved further by employing several machine learning models, such as logistic regression, to routinely ascertain the required score to correctly weight each rule.

Table 6. The contribution of each rule.

Rules	Contribution of each rules
Genitive objct mphasis (GOE)	0.00
Frequency indicator(LR)	1.87
Recency	11.13
Definiteness	4.48
Section Headig Preference (SH)	0.00
Non-Prepositional Preference	1.63
Term Preference (TP)	0.00
Referential Distance(RD)	0.29
Patterns of sentences(PS)	1.19
Grammaticl function (GFNP)	3.52
Nominal Predicate case(NPC)	5.12
First Noun Phrases(FNP)	0.16
Closet Subject Nominative Case Preference	0.29

7. Conclusion

In this paper, we present a model based on rule based method for the Arabic anaphora resolution. It consists of two main steps. A first preliminary step involves three NLP tasks POS tagging, NP chunking, and grammatical relation identification. The second step consists on identifying all the pronouns in the text and eliminating non-anaphoric pronouns such as pleonastic ones. The third step aims to browse the surrounding context for each anaphora in order to identify all possible candidates and to filter them using the gender and number agreements and search space. The final step uses to find out the most likely antecedent for anaphora. The contribution of each rules to the model are investigated. The results indicate that several rules, such as Frequency Indicator (LR), Nearest NP, Term Preference (TP), and Nominal Predicate case (NPC) are more helpful for Arabic anaphora resolution. The results illustrate that the proposed model achieved good results.

8. References

1. Houfeng W, On Anaphora Resolution within Chinese Text, Applied Linguistics. 2004; 52(4):113–19.
2. Holen G. Automatic Anaphora Resolution for Norwegian, Springer, Lecture Notes in Computer Science. Berlin, Germany; 2007. 4410.
3. Dutta K, Prakash N, Kaushik S. Resolving Pronominal Anaphora in Hindi using Hobbs' Algorithm, Web Journal of Formal Computation and Cognitive Linguistics. 2008; 1(10).
4. Ali R, Khan M, Rabbi I. Strong Personal Anaphora Resolution in Pashto Discourse. Proceedings of EEE ICET 3rd International Conference on Emerging Technologies, Islamabad, Pakistan; 2007.
5. Soon W, Ng H, Lim D. A Machine Learning Approach to Coreference Resolution of Noun Phrase, Computational Linguistics. 2011; 27.
6. Ng V. Cardie C. Improving Machine Learning Approaches to Coreference Resolution. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002.
7. Jauregi AZ, Sierra B, Uriarte O, Ceberio K, Illarraza A, Goenaga I. A Combination of Classifiers for the Pronominal Anaphora Resolution in Basque, Springer, Lecture Notes in Computer Science, Germany. 2010; 6419:253–60.
8. Abdul-Mageed M. Automatic Detection of Arabic Non-Anaphoric Pronouns for Improving Anaphora Resolution, Journal ACM Transactions on Asian Language Information Processing. 2011; 10:1–11.
9. Le M Tran, Nguyen T, Ha Q. Co-Reference Resolution in Vietnamese Documents Based on Support Vector Machines, International Journal of Asian Language Processing IALP. 2011; 89–92.
10. Fallahi F, Shamsfard M. Recognizing Anaphora Reference in Persian Sentences, International Journal of Computer Science. 2011; 8:324–29.
11. Tüfekçi P, Kılıçaslan Y. A Computational Model for Resolving Pronominal Anaphora in Turkish using Hobbs Naïve Algorithm, International Journal of Computer, Information Science and Engineerin. 2005; 1(5):1416–20.
12. Khan M, Ali M, Khan M. Treatment of Pronominal Anaphoric Devices in Urdu Discourse. Proceedings of IEEE ICET 2nd International Conference on Emerging Technologies, UET Lahore, Pakistan; 2006.
13. Ali R, Kha M, Ahmad R, Rabbi I. Rule based Personal References Resolution in Pashto Discourse for Better Machine Translation. Proceedings of IEEE ICEE 2nd International Conference on Electrical Engineering; 2008. P. 1–6.
14. Noor N, Aziz M, Noah S, Hamzah M. Anaphora Resolution of Malay Text: Issues and Proposed Solution Model. International Conference on Asian Language Processing IALP; 2011. P. 174–77.
15. Converse S. Resolving Pronominal References in Chinese with the Hobbs Algorithm. Proceedings of SIGHAN workshop on Chinese language processing; 2005. p. 116–22.
16. Hammami S, Sallemi R, Belguith L. A Bayesian Classifier for the Identification of Non-Referential Pronouns in Arabic. Proceedings of the 7th International Conference on Informatics and Systems; 2010 March 28-30. p. 1–6.
17. Hammami S, Belguith L, Hamadou A. Arabic Anaphora Resolution: Corpora Annotation with Coreferential Links, The International Arab Journal of Information Technology. 2009; 6:481–89.
18. Albared M, Omar N, Aziz M, Nazri M. Automatic Part of Speech Tagging for Arabic: an Experiment using Bigram Hidden Markov Model. Proceedings of the 5th International Conference, Rough Set and Knowledge Technology, Beijing, China. 2010 October 15-17.
19. Albared M, Omar N, Aziz M. Improving Arabic Part-of-Speech Tagging through Morphological Analysis. Proceedings of the Intelligent Information and Database Systems, Daegu, Korea. 2011 April 20-22.
20. Mohammed M, Omar N. Rule based Shallow Parser for Arabic Language, Journal of Computer Science. 2011; 7:1505–14.
21. Hammadi O, Aziz M. Grammatical Relation Extraction in Arabic Language, Journal of Computer Science. 2012; 8:891–98.