Arabic Anaphora Resolution: A Distributional, Monolingual and Bilingual Approach

Rania Al-Sabbagh* Khaled Elghamry**

re832003@yahoo.com elghamryk@yahoo.com

Faculty of Al-Alsun (Languages), Ain Shams University, Cairo, EGYPT

***Abstract*—**This paper presents an algorithm for Anaphora Resolution (AR) in Arabic. The paper is motivated by the poor performance of current Arabic-English Machine Translation (MT) systems in terms of AR and the fact that AR is an understudied issue in Arabic Natural Language Processing (ANLP). The algorithm suggested follows a distributional, monolingual and bilingual bootstrapping approach to acquire AR-related features that cannot be provided by monolingual resources, using a second language (here English). To the best of the authors' knowledge, this paper is the first attempt for AR in Arabic using a generic corpus, *Al-Ahram Newspaper* (1998-2005). One thing the paper tries to illustrate is the feasibility of using only morphological cues for proper AR.

***Index Terms*—** Arabic Anaphora Resolution, Part-of-Speech Disambiguation, Semantic Features Acquisition, Tokenization, Chunking.

I. Introduction

Anaphora Resolution (AR) is the process of figuring out the antecedent (i.e. referent) of a given anaphor [4] [6]. The paper focuses on encliticized Arabic third person personal pronouns, ها /*hA*/ (her/hers/it/its), ه /*h*/ (him/his/it/its), هم /*hm*/ (masculine: them/their) and هن /*hn*/ (feminine: them/their). In spite of being an area under active research in formal and computational linguistics, AR is understudied in ANLP. The only study of AR in Arabic, to the authors' knowledge, is that of [5], who have studied AR only in Arabic technical manuals, achieving a precision rate of 95.4%. However, their approach has never been tested on unrestricted (i.e. generic) texts.

AR causes problems for Machine Translation (MT) systems dealing with Arabic generic texts such as Sakhr MT system which is available for public use through the website www.ajeeb.com. This is evident in the following Sakhr's translation for an Arabic sentence extracted from *Al-Ahram Newspaper*:

(1) صرحت السيدة قرينة الرئيس بأنها ستكثف عملها بالتعاون ....

Transliteration***: */SrHt Alsydp qrynp Alr}ys b>nhA stkvf EmlhA bAltEAwn/*

Sakhr's Translation: The Mrs. announced the president's wife that *it* will intensify *its* work in cooperation …

Correct Translation: The president's spouse announced that *she* will intensify *her* work in cooperation …

Errors made by MT systems in terms of AR are due to the differences between Arabic and English pronominal systems. The Arabic pronominal system, unlike the English one, does not differentiate linguistically between ±human entities. As a result, both the –HUMAN FEMININE noun الشجرة */Al$jrp/* (the tree) and the +HUMAN FEMININE noun الفتاة */AlftAp/* (the girl) are referred to using the same 3rd person personal pronoun ها */hA/* (she/ her/ hers) as in (2) and (3).

(2) رأيت الشجرة فرويتها****

Transliteration: */r>yt Al$jrp frwythA /*

Literal Translation: I saw *the tree* so I watered *her.*

Gloss: I saw *the tree* so I watered *it*.

(3) رأيت الفتاة فصافحتها

Transliteration: */r>yt AlftAp fSAfHthA /*

Literal Translation: I saw *the girl* so I shacked hands with *her*.

Moreover, Arabic pronouns, unlike the English ones, have the same form for all different grammatical cases; the nominative, the accusative and the genitive. In (4), ها */hA/* (her/hers) is used in the accusative case in its first occurrence and in the genitive case in its second occurrence; yet in both cases it has the same form.

(4) أكرمتها وأحسنت إليها

Transliteration: */>krmth w>Hsnt <lyhA /*

Literal Translation: I gave *her* generously and donated to *her*

Finally, Arabic 3rd person pronouns are sometimes encliticized; the thing that makes them ambiguous. This is evident in المهن */Almhn/* (the professions or their pain), where the last two letters هن */hn/* -being identical to the 3rd person personal feminine plural pronoun هن- are ambiguous; it is not clear whether they are a part of the word – ال */Al/* (the) مهن */mhn/* (professions) - or it is an encliticized pronoun –الم */Alm/* (pain) هن/*hn/* (their).

Such differences make Arabic AR a non-trivial task that is further complicated by the large number of the preprocessing tasks required for a decent AR algorithm. A

*Rania Al-Sabbagh is with the Faculty of Al-Alsun (Languages), Ain Shams University, Cairo, Egypt. Phone: 202-2876992. E-mail: re832003@yahoo.com.

**Khaled Elghamry is with the Faculty of Al-Alsun (Languages), Ain Shams University, Cairo, Egypt. E-mail: elghamryk@yahoo.com.

***Buckwalter's Transliteration scheme (Buckwalter 2002, Diab et al. 2004). URL: http://www.qamus.org/transliteration.htm

**** Examples are adapted from Al-Hashemy (2004)

real-world AR system vitally relies on the efficiency of the pre-processing tools, such as POS taggers, morphological analyzers, NP chunkers, parsers and so on [6]. Sometimes, such NLP tools might not be as accurate as demanded or they might not even be available.

The rest of the paper falls in three parts. The first handles Arabic AR preprocessing phases, discussing their difficulties and the suggested solutions together with their performance rates. The second part discusses the AR algorithm, AR-related features and the algorithm's performance. Finally, the conclusion discusses further procedures intended by the authors for a better AR algorithm.

II. AR Preprocessing Phases

AR preprocessing handles a number of issues including tokenization, Part-of-Speech (POS) tagging and disambiguation, semantic features acquisition and nominal chunking. AR in Arabic is still an open area of research. Therefore, the authors have intended to reinvent necessary NLP tools so as to know whether AR in Arabic requires special processing for such tools.

A. Tokenization:

The authors have developed a corpus-/dictionary-based tokenizer (i.e. a light stemmer that strips off proclitics and enclitics). The dictionary used for the tokenizer is that of Buckwalter [2]. According to a manual evaluation of a random sample, the tokenizer has achieved a recall rate of 0.997 and a precision rate of 0.95. Therefore, its performance rate according to the F-measure is **97.3**% (token-based).

As aforementioned, enclitic Arabic third person personal pronouns can be ambiguous. Such a problem can be considered as a subtask of tokenization and also a subtask of POS disambiguation. For example, منـه */mnh/* has been ambiguously analyzed as one-word female name (Menah) and also as a preposition procliticized to a third person personal pronoun (from him). Using Buckwalter's most frequent sense (i.e. the first analysis in the list of Buckwalter's analyses) disambiguates such a word, giving it the second analysis. The ambiguous لهـا */lhA/* is analyzed as one-word past verb (he played) and a preposition procliticized to a third person personal pronoun (for her). The tokenizer correctly disambiguates it, analyzing it as a preposition and a pronoun.

In order to further enhance pronoun's disambiguation, a heuristic has been used. The heuristic relies on a simple grammatical rule in the Arabic language saying that only one definite marker can exist within the same word boundary [3]. In Arabic, both pronouns and the definite article */Al/* are markers of definiteness. Therefore, the same word cannot be procliticized to */Al/* and encliticized to a pronoun at the same time. That is why if a word is analyzed by the tokenizer and Buckwalter's most frequent sense as:

<PRON>هم <DET>ال

The analysis of the second part هم */hm/* is converted from a PRONOUN into a NOUN meaning (frustration).

The use of the tokenizer; of the most frequent sense in Buckwalter's lexicons (i.e. the first analysis in the lexicon) and the use of the definiteness heuristic have achieved an F-measure performance rate of **~ 97%** manually measured on a random sample.

B. POS Tagging and Disambiguation

Buckwalter's Arabic Morphological Analyzer (AraMorph) [2] is the system used to provide POS tags and morphological analysis. Buckwalter's Analyzer gives all possible analyses of a given word. Due to the rich morphology of the Arabic language a word might be given up to 17 analyses. Since the antecedent of a given anaphor is typically a noun, nouns need to be disambiguated. Moreover, an appropriate POS disambiguation module leads to a better nominal chunking.

The procedures of POS disambiguation include:

(1) Using the most frequent sense in Buckwalter's lexicons [2] which correctly disambiguates **84.6%** of the ambiguous words (type-based precision) according to a manually evaluated random sample,

(2) Using our tokenizer that reduces the percentage of the ambiguous words to **34%**,

And **(3)** Using a number of heuristics that rely on Arabic function words: prepositions {من */mn/* (from), على */ELY/* (on), بـ */b/* (by) … etc.}, demonstratives {هذا */h*A/* (this), تلك */tlk/* (that) … etc.}, relative pronouns {الذي */Al*y/* (who/which) … etc.}, negative {لا */lA/* (no), لن */ln/* (no) … etc.}, and interrogative particles {لماذا */lmA*A/* (why), كيف */kyf/* (how) … etc.}. The first two are typically followed by a noun, whereas the last three are typically followed by a verb. For instance, Buckwalter's most frequent sense and the tokenizer have given the following analysis to the ambiguous شعر */$Er/* that can be either a verb (feel) or a noun (poetry):

من <PREP>شعر <VERB>

The heuristics used correctly analyze شـعر */$Er/* in the preceding context as a noun, meaning *poetry*. These heuristics have achieved a **69.3%** precision rate measured only according to ambiguous words.

Integrating all three procedures has resulted in a precision rate of **~ 94%** measured according to a random sample that is manually evaluated.

C. Named Entities

A named-entity base is compiled by the authors using the World Wide Web (WWW). Google search engine is used to surf the WWW for lists of Arabic names. As a result, a list of **5,000** names, male and female, is compiled and manually filtered.

D. Semantic Features Acquisition

For all previous preprocessing tasks, only monolingual evidences (i.e. cues extracted from the Arabic language) have been used. However, for semantic features acquisition and nominal chunking, the authors have integrated *monolingual and bilingual bootstrapping techniques*.

Bilingual bootstrapping extracts required semantic information (i.e. gender, number and rationality) from the available English generic corpora and resources. This idea is motivated by the fact that such necessary AR-related semantic features are not provided by Arabic morphological analyzers. According to Buckwalter's analyses, only **32.8%** of the nouns

are marked for number, **35.5%** are marked for gender and **0%** is marked for rationality.

Bilingual bootstrapping adds such semantic features to the Arabic noun translations of the English nouns that precede or follow specific English cues. Bilingual bootstrapping falls in the following steps:

(1) Some English cues, namely *who, which, a, an, each, every, some, any* and *all,* are used to extract words from generic English corpora. *Which* and *who* are markers for rationality; they refer to –HUMAN and +HUMAN semantic features respectively. *A, an, each* and *every* are typically followed by SINGULAR nouns, whereas *some, any* and *all* are usually followed by PLURAL nouns.

(2) The resulting English words are submitted to Golden Al-Wafi English-Arabic MT system [1],

(3) Number and gender semantic features are added to the Arabic noun translations of the English nouns.

(4) The output list of the Arabic nouns which are tagged for number, gender and rationality is manually filtered, resulting in a noun-base of **50,000** types tagged for semantic features.

Monolingual evidence depends on Arabic cues (i.e. Arabic demonstratives and relative pronouns) to extract such semantic features. For instance, the demonstrative هذا */h*A/* (this) is usually followed by a SINGULAR, MASCULLINE noun; as in هذا الولد */h*A Alwld/* (this boy). The relative pronoun التـي */AltY/* (who) is usually preceded by a SINGULAR, FEMININE noun as in الفتاة التي */AlftAp Alty/* (the girl who). Monolingual evidence has increased the size of the noun-base to **60,000** nouns tagged for number, gender and rationality. The list has also been manually filtered.

As a result of the final noun-base resulting from the bilingual/ monolingual bootstrapping algorithm, **90%** of the corpus nouns are tagged for number, **93.3%** are tagged for gender and **44%** are tagged for rationality.

E. Nominal Chunking

Nominal chunking is motivated by the fact the anaphor (i.e. the pronoun) might refer to an entire chunk rather than a single word as in:

(5) صرح الرئيس الامريكي انه سيتدخل

Transliteration: */SrH Alr}ds AlAmryky Anh sytdxl/*

Translation: *The American president* declared that *he* would interfere

The antecedent is the complete nominal chunk الرئيس الامريكي/ */Alr}ds AlAmryky* / (the American president).

Bilingual bootstrapping is also used to extract nominal chunks from English generic corpora and resources. **First**, a simple parser for English has been developed to extract English nominal chunks. Second, the resulting chunks are submitted Golden Al-Wafi English-Arabic MT system, resulting in **60,854** possible chunks. Out of this list **96.4%** has been validated as existing chunks in Arabic generic corpora.

Monolingual evidence for nominal chunking relies on unambiguous nominals (i.e. nouns and adjectives) that are extracted from the corpus and on a set of cues, based mainly on definiteness markers. The heuristics can get nominal chunks such as:

(6) الوضع الأقتصادي الراهن

Transliteration: */AlwDE Al>qtSAdy AlrAhn/*

Translation: The current economic situation

(7) تفعيل دور الأمم المتحدة

Transliteration: */tfEYl dwr Al>mm/*

Translation: Implementing of the United Nations' role

(8)موضوع ثقافي اجتماعي رائع

Transliteration: */mwDwE vqAfY AjtmAy rA}E/*

Translation: A gorgeous cultural social topic

According to a manual evaluation to a random sample, the nominal chunking module has a precision rate of **~ 85%**.

III. AR ALGORITHM

The AR algorithm makes use of all previous information extracted through preprocessing phases. The algorithm depends on three AR-related features: *recency, collocational information (i.e. distributional evidence)* and *agreement*.

Recency means that the closest candidate antecedent that agrees in gender, number and rationality to the anaphor is the correct one. For instance in (10), there are two candidate antecedents for the pronoun ها */hA/* (a 3rd person feminine pronoun), namely, السيدة */Alsydp/* (the lady, FEM) and خطة */xTp/* (a plan, FEM), however, the correct antecedent is خطة */xTp/* (a plan) which is the closest to the pronoun.

(9) اعلنت السيدة الاولى عن خطة العمل التي سيتم من خلالها ....

Transliteration: */AElnt Alsydp AlAwlY En xTp AlEml Alty sytm mn xlalhA/*

Literal Translation/ Gloss: The first lady has declared the working plan through which ….

Collocational evidence depends on finding out the collocational relation between candidate antecedents and the pronoun's carrier. Chi-square (X^2) and Relative Frequency (RF) are the two association measures used for the collocational evidence. Collocational evidence is a good feature for such examples as (11):

(10) لـم يـتمكن هتلـر مـن إبـادة اليهـود الـذين رأت جماعـة مـنهم أن تهـاجر إلـى فلسطين لتصطدم بأهلها ..

Transliteration: */lm ytmkn htlr mn <bAdp Alyhwd Al*yn r>t jmAEp mnhm >n thAjr <lY fylsTyn ljSTdm b>hlhA/*

Translation: Hitler couldn't exterminate the Jews; some of whom have immigrated to Palestine to face its citizens

In (11), the bigram between the pronoun's carrier أهل */>hl/* (citizens) and the candidate antecedent فلسطين */flsTyn/* (Palestine, SG, FEM) has a higher X^2 and RF than the bigram between the same pronoun's carrier and the candidate antecedent جماعة */jmAp/* (a group, SG, FEM).

Finally, the algorithm stipulates that the antecedent agrees in gender and number with the anaphor. That is why العمل */AlEml/* (the work: SG, MASC) in (3) is excluded from the set of candidate anaphors being a MASCULINE word, whereas the pronoun ها */hA/* requires a FEMININE antecedent. Rationality feature is a good feature as far as the 3rd person personal pronoun is concerned, since it refers only to +HUMAN entities. For instance, in (12) both القضاة */AlqDAp/* (the judges) and مطالب */mTAlb/* (demands) are PLURAL, however, only the first is +HUMAN; and therefore, only the first is a possible candidate.

(11) حقق القضاة مطالبهم
Transliteration: */Hqq AlqDAp mTAlbhm/*
Translation: The judges have achieved their demands

Therefore, the algorithm chooses the candidate with the highest collocational evidence provided that it agrees in gender and number with the anaphor and it is the closest candidate to the anaphor (i.e. the pronoun).

IV. RESULTS AND DISCUSSION

Evaluation has been performed in two different manners. First, a gold standard set of 5,000 tokens manually annotated for AR is being used, and second, random samples (~ 3, 000 tokens) of the algorithm's output have been manually evaluated. A mean performance has been measured out of these two evaluations. The corpus used is that of Al-Ahram Newspaper (1998-2005), approximately 20, 000, 000 tokens and 971, 000 types.

One main problem of the distributional approach to AR is *data sparseness*. When the algorithm works on individual words candidates, it achieves a precision rate of **84%**. However, when it works on nominal chunks, precision declines to **~ 71%**, since an individual word such as الأزمة */Al>zmp/* (the crisis) is more frequent than the complete nominal chunk الأزمة الأقتصادية الطاحنة */Al>zmp Al>qtSAdyp AlTAHnp/* (The severe economic crisis). As a result of data sparseness, in many cases both the X^2 and the RF are not an indicative AR-feature.

Sometimes, all AR-related features do not work at all as in (13), where both التقدم */Altqdm/* (progress) and الانسان */AlAnsAn/* (Man) agree in gender and number with the anaphor ه */h/* (i.e. both are SINGULAR and MASCULINE), and الانسان */AlAnsAn/* (Man) is the closest candidate to the anaphor, yet the correct one is التقدم */Altqdm/* (the progress).

(12) التقدم التكنولوجي الذي يتسارع الانسان في الوصول اليه
Transliteration:*/Altqdm Altknwlwjy Al*y ytsArE AlAnsAn fy AlwSwl Alyh/*
Translation: The technological progress to which Man is eager to reach

Such cases are frequent when the pronoun's carrier is a function word, a preposition, for instance. This is because function words are very frequent and therefore they can occur frequently with all candidates, and thus the X^2 and the RF are not indicative.

Moreover, the X^2 and the RF can give the wrong candidate as in (14). The pronoun's carrier أوان */>wAn/* (time) is more frequent with الصورة */AlSwrp/* (the picture: SG, FEM) than with النصيحة */AlnSyHp/* (the advice: SG, FEM). Besides, the former is the closest candidate to the pronoun ها */hA/*.

(13) ثبت ان النصيحة علي الصورة التي تقدم بها ليس هذا اوانها
Transliteration: */vbt An AlnSyHp ElY AlSwrp Alty tqdm bhA lys h*A AwAnhA/*
Translation: It is proved that the advice in the way it is presented is not in its appropriate time

As a result, the authors are currently developing and studying the use of a better technique to measure collocational evidence and the use of more AR-related features as discussed below.

V. CONCLUSION AND FUTURE WORK

This paper presented an algorithm for AR in Arabic, with a precision rate varying from **71%** to **84%** according to the used input units, whether individual words or nominal chunks. Many preprocessing tasks are required for an AR algorithm. The authors have developed their own tools so as to know whether AR requires special processing, especially it (i.e. AR) is still an open area of research in Arabic. It has been proved that AR does not require special treatment for preprocessing; its preprocessing tools can be used for any other NLP application.

One thing that the paper showed is that only morphologically-/distributionally-evident features can achieve a reasonable performance for an AR algorithm in comparison to the algorithm of Mitkov et al. (1998), taking into consideration the huge set of features used by Mitkov et al. and the fact that they have applied their algorithm on restricted corpora (i.e. technical manuals).

Currently, the authors are developing their algorithm in the following ways:

First, currently available ANLP tools for POS disambiguation and nominal chunking are being tested to see whether they will achieve better results than the tools developed by the authors.

Second, a better technique to get collocational information is being used so as to avoid data sparseness; instead of measuring the X^2 and the RF of the complete nominal chunk, such association measures will be used to get the collocational information between the pronoun's carrier and the nominal chunk's head word.

Finally, two morphologically-evident features are to be added to the algorithm. The first is definiteness. Preliminary results show that the correct antecedent is more likely to be definite. The second feature that will be added is that of coreferential chains. Preliminary results show that within -20 window size, pronouns of the same person, number and gender tend to form a coreferential chain.

Moreover, evaluation needs to be further developed. First, a larger set of golden standard data is to be formed. Second, a more comprehensive framework of evaluation needs to be implemented: for example, comparisons to baseline models will be a good evaluation metric.

REFERENCES

[1] ATA Software Technology Ltd. (2002). Golden Al-Wafi Translator Software. Version 1.12.

[2] Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. LDC. Catalog number LDC2002L49, ISBN 1-58563-257-0.

[3] Hasan, A. (1999). *AlnHw AlWAfy.* Vols. 1 and 2. Cairo: dAr AlmEArf.

[4] Jurafsky, D. & Martin, J. (2000). *Speech and Language Processing; An Introduction to Natural Language Processing, Computational linguistics and Speech Recognition.* New Jersey: Prentice Hall Ltd.

[5] Mitkov, R. (1998). Robust Pronoun Resolution with Limited Knowledge. *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, Montreal, Canada, 869-875.

[6] Mitkov, R. (2001). Outstanding Issues in Anaphora Resolution. *Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing,* Mexico City, 110-125.