A Coreference Resolution Approach using Morphological Features in Arabic

Majdi Beseiso
Department of Computer Science
Al-Balqa' Applied University
Alsalt, Jordan

Abdulkareem Al-Alwani
Department of Computer Science & Engineering
Yanbu University College
Yanbu, Saudi Arabia

***Abstract*—Coreference resolution is considered one of the challenges in natural language processing. It is an important task that includes determining which pronouns are referring to which entities. Most of the earlier approaches for coreference resolution are rule-based or machine learning approaches. However, these types of approaches have many limitations especially with Arabic language. In this paper, a different approach to coreference resolution is presented. The approach uses morphological features and dependency trees instead. It has fivestages, which overcomes the limitations of using annotated datasets for learning or a set of rules. The approach was evaluatedusing our own customized annotated dataset and "AnATAr" dataset. The evaluation show encouraging results with average F1 score of 89%.**

Keywords—Coreference resolution; Anaphora; Alternative Approach; Arabic NLP; morphological features

I. Introduction

Coreference resolution is an important part of natural language processing. It is the process of identifying natural language expressions and determining which of these different entities refer to the same entity [1, 19, 10, 14, 5]. It is significant for the task of detecting events and entities in a text and cluster them [18]. This process helps in many of the NLP applications such as data extraction, text manipulation, and machine translation [1]. Referents are real word objects or entities, which makes coreference resolution an important hard step towards understanding language [5].

This paper focuses on anaphora, and cataphora coreference resolution in Arabic written sentences. Arabic is morphologically rich language and has a distinctive nature, which makes many of the traditional approaches limited [15]. We present a different approach for coreference resolution using deep morphological and syntactical features as well as dependency trees. The approach makes use of the fact that many Arabic words can be morphologically derived from a set of words or roots, to make relations between different words [5]. Dependency trees provide a different type of relations between words depending on the grammatical rules. In this approach, we use both techniques to determine reference relations.Our model has five stages, text preprocessing, pro-forms and noun entities (NEs) extraction, morphological analysis, relating NE and preforms, and output validation. The approach includes many linguistic applications such as morphological analysis, POS tagging, tokenization, and extracting the nouns entities. That is why Different Arabic linguistic tools are used to realize the applications in the different stages. In this paper, we present our own customized and annotated dataset for coreference resolution. It is used for testing along with the "AnATAr" dataset. The evaluation of the system results with average F1 score of 89%.

The paper has seven sections. In section 2, we present some definitions and descriptions related to coreference resolution and after that, in section 3 we review related work. We present the Methodology and the scope of our approach in section 4, which is fully described along with our proposed model in section 5. In section 6, we describe the results of the experiments done and finally in section 7, we conclude and show future work.

II. Basic Concepts and Characteristics of Arabic Coreference Resolution

Anaphora is the process of finding the referent of an anaphor entity that is referring to an entity back in the sentence [14][8]. Cataphora is a similar process where the pro-form precedes the entity, to which it refers [17]. When the anaphor or the cataphor and the entities they are referring to hold the same referent in real world then they are "coreferential" [14][17]. Example 1 is for anaphora coreference and example 2 is for cataphoracoreference.

(1) ذهب **العلماء** إلى المؤتمر **وهم**كانوا فرحين.

The **scientists** went to the conference and **they** were happy.

(2)**هوالعلم**، الذي يغني الشعوب.

It is **science**, which enrich people.

Pronominal anaphora is one of the most frequent types of anaphora coreference resolution that deals with pronouns [10]. Pronoun is a type of pro-form that refers to a noun word or a noun expression. In this approach, we deal with certain types of pro-forms, which are all the independent and the attached pronouns (جميع الضمائر المنفصلة والمتصلة) which falls under the pronominal anaphora.

Typically, coreference resolution is a very hard process even for the English language, as it needs some information about the real world to understand relation between words [18]. There are more challenges that exists when dealing with Arabic language that makes coreference resolution in general and Arabic anaphora in specific a more complicated process. Arabic language sentences can have complex structure [10].

(3) اعتبرناهم كأصدقائنا

We considered them as our friends.

In example 3, a sentence in Arabic was translated into six words in English, which shows how complex Arabic sentences can be. Arabic language has the feature of free word order, which adds to the complexity of coreference resolution. Free word order means that there are almost no restrictions about words order in a sentence. Sometimes the referent is ambiguous specially that pro-forms could exist in a connected form or separated form [10].

(4) ذهبت **سارة** إلى **مريم** و**كانت** سعيدة

Sara went to Mariam and she was happy.

In example 4, there is a connected pro-form in "كانت" and it has ambiguous coreferenceto either Sara or Mariam. In Arabic language, the consistency between the morphological and syntactical features of the pro-form and the related named entities (NEs)needs to be considered. These morphological includes gender, number (singular, dual, plural), and subject or object reference. Finally, one of the main challenges with Arabic language is that there are not enough annotated corpora for coreference resolution [10].

III. Literature Review

Coreference resolution is an important and complex process, which made it the subject of much research work. Most of the research done on coreference resolution showed common processes done by most of the approaches. Which are identifying the search scope, such as the whole document, a sentence, or a set of sentences, and the preprocessing step where the text is segmented, processed and noun phrases or entities are identified. The last step differs where certain tasks are accomplished to do the resolution [19,10].

A survey about anaphora resolution in general and in the Arabic language in specific was presented in "Arabic Anaphora Resolution Using Holy Qur'an Text as Corpus" [10]. The paper presented two types of anaphora resolution. This first type is rule-based approaches. In which, a knowledge base is built to be used in the process. It is easy to implement and does not require much data, but on the other hand, it needs a large set of human formed rules to cover all the needed features for resolution. The statistical approach or the machine learning approach depends on annotated corpora for both training and testing. This approach can have better results when it comes to accuracy, speed and giving a generalized model, but this depends on the annotated data.

In "A Machine Learning Approach to Coreference Resolution of Noun Phrases", they took the path of the machine learning approach [19]. In this approach, Annotated corpus is required to be used as training and testingdata. In addition, they have to determine the feature vector. Which is a set of features used to define the relation between two entities. The next step is to generate training examples, then to build the classifier, and the last step is to generate "coreference chains for test documents" [19]. The accuracy was close to the other approaches. In the types of errors that affect recall, "inadequacy of current surface features" scored 64% of all types. The paper represents a good approach for coreference resolution, but both the features and annotated corpora can restrict the effectiveness of the approach. There has been several trials to overcome these two problems.

In 2012, CoNLL shared task targeted, "modelling of coreference resolution for multiple languages" [18]. The OntoNotes data was the baseline for the modelling, which has different annotation layers, and in the three languages English, Arabic, and Chinese. The paper released by CoNLL 2012 mentioned that the morphology of Arabic language is very complex comparing to English, which has limited morphology and Chinese whichhas very little morphology. The resources available for each language are different and Arabic has the least resources. The shared task presented good data for training and testing coreference resolution in Arabic. CoNLL suggested that a hybrid approach between rules-based approach and machine learning approach to give the highest accuracy.

Chen Chen and Vincent Ng presented a system with a hybrid approach for the CoNLL 2012 in their research [2]. They combined both rule-based approach with statistical approach. They used the lexical information with machine learning to improve the approach. The results showed the effectiveness of the approach. The problem with is hybrid approach is that it showed lower accuracy in Arabic for all the tests that were done. The results on the development set were around 60% for English and Chinese, but for Arabic were around 45%.

In the CoNLL 2012, they stated that Arabic has a complex morphology, and that Arabic has limited resources for comparison. Which lead us to explore Arabic morphological analysis. In the research paper "Arabic Finite-State Morphological Analysis and Generation", they presented a morphological analysis system, which included displaying the root, pattern, and different affixes, mood, voice, etc. [11]. The paper mentioned that Arabic morphology is very challenging as for example Arabic "orthography displays an idiosyncratic mix of deep morphophonological elements" [11]. They presented a system that can recognise all possible written forms of words and even with varying degrees of diacritical marking [11]. In a different research, morphological stemming was used to improve Arabic Mention Detection and Coreference Resolution [5]. The system make use of "finite state segmentation" and relationships between word stems. The usage of stemming features was very effective in Arabic as it increased the accuracy in the testing data.

The traditional approaches showed to a fair extent inability to accurately solve the challenging problem of Arabic coreference resolution. The need of large set of rules for the rule-based approaches or the annotated data and the set of features for the machine learning approaches made these approaches restricted. Research suggested that machine learning approach or in specific, the hybrid approach for coreference resolution should give the highest accuracy. However, in Arabic, the accuracy was still low comparing to other languages. Where research showed the importance of Arabic morphological analysis and how it can effectively improve coreference resolution. This showed that in Arabic, it

is more effective to depend on the usage of morphological and syntactical features for coreference resolution.

IV. Methodology & Scope

Arabic is considered as, highly inflected, agglutinative, and morphologically rich language [5, 15]. These features made Arabic language distinctive from many other languages, leading to the limitations of the traditional approaches in coreference resolution. This proposed model makes use of the nature and complexity of the Arabic language to overcome the limitations of other approaches, by including different morphological analysis techniques along with dependency trees.

According to the University of Duisburg-Essen, Morphology is the study of word forms and a morpheme is the smallest unit that has a meaning [16]. Many Arabic words can be morphologically derived or associated with a list of words or roots. This process is done by removing different prefixes and suffixes attached to the word. Not only, Arabic words can be in different forms, but also many pronouns, prepositions, and conjunctions can be attached to words [5]. In Arabic language, the word root is "the original form of the word before any transformation process", and it has major importance in Arabic language processing [16]. In addition to the different forms of the Arabic word that result from the derivational and inflectional process, most prepositions, conjunctions, pronouns, and possessive forms are attached to words. These orthographic variations and complex morphological structure make Arabic language processing challenging.

كاتب، كتاب ، ك ت ب (5)

In example 5, there are two different words that have the same root. They have the same root of three letters, but their meaning are different. Roots can be used to relate the two words. A stem is one morpheme or more that can accept an affix [16].

In this approach "AlkhalilMorphoSy"was used, which is a morphological analyzer that provides all possible solutions with their morphosyntactic features for a certain set of words [12]. The tool presents a wide range of features such as, vowelization, proclitics and enclitics, nature of the word, stems, roots, and syntactic form. This tool provides effective analysis, which is done over several steps. The tool is built based on the characteristics of the Arabic language, which makes it suitable for our approach. In addition, Alkhali tool was very effective and more accurate in the evaluation against other analyzers [12].

In addition to the use of a morphological analyser, our approach make use of dependency trees to make relation between different words in a sentence. Stanford dependencies describes the representation of grammatical relations between different words in a sentence [13].

Figure 1 is a graphical representation for the Stanford dependencies for the sentence, "Bell, based in Los Angeles, makes and distributes electronic, computer and building products" [13]. This is a clear directed graph to represent the relations between the different words as edge labels. Stanford CoreNLP provides such features in Arabic such as, tokenization, segmentation, part of speech, and dependency trees, which is similar to the figure above [3]. In other words, the Stanford CoreNLP tool is used to provide extra information about the whole sentence.

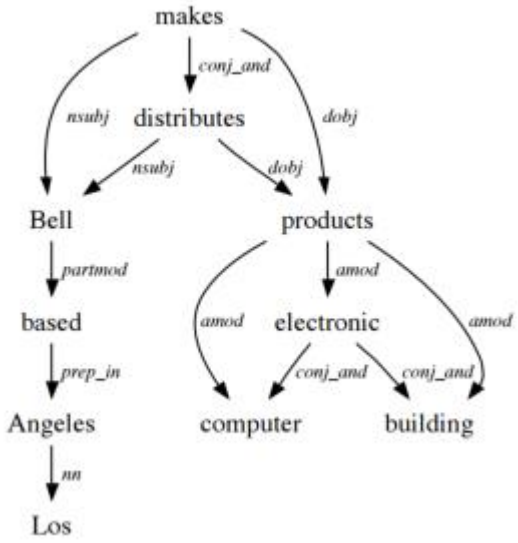

Fig. 1. Standard Stanford dependencies

The approach has a defined scope for the resolution process, which is a sentence with a complete context.

ذهبت سارة إلى مريم فسألتها عن موعد الحفلة(6).

Sara went to Mariam and asked her about the date of the party.

In example 6, it can be Sara or Mariam who asked the question which means that the context not complete and this is out of the scope of this approach.

ذهبت سارة إلى مريم فسألتها عن موعد الحفلة فقالت مريم لم أكن أعلم من (7) قبل.

Sara went to Mariam and asked her about the date of the party then Mariam said I did not know that before.

In example 7, the context was complete as there is a reference to Mariam in the same sentence. That makes coreference resolution possible and detectable by our approach.

We present an approach that makes use of the nature and "morphology richness" of the Arabic language, which can be considered word-based features. In addition, we include sentence-based features using the Stanford CoreNLP tool. Our approach does not require a wide range of rules neither a large annotated data set, and still provides an effective solution for Arabic coreference resolution.

V. Proposed Model

The approach consists of five different stages. They are designed in order to make the best use of Arabic words forms, and sentences structure. The stages have different scope, goal,

and complexity in order to reach the maximum accuracy and performance.

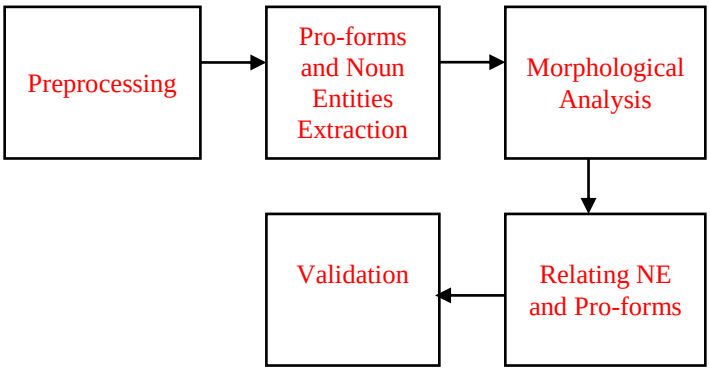

Fig. 2. The four stages of the approch

These five stages are shown in figure 2 as a pipeline of different natural language processing tasks.

A. Preprocessing

In this stage, all sentences will be processedand words to prepare them for the next steps. This stage includes different tasks aiming at making the input text in the correct format for the next stages. The first task is sentence splitting which is done using dots in the sentences. Morphological analysis is the second stage and it is applied to all entities using "AlkhalilMorphoSy" with its different features. For the third task, we use Stanford POS Tagger, which includes a module supporting Arabic to perform POS tagging on the input text. The last step is filtering the output of the second step "solutions of Alkhalil" based on the results of the third step or the Stanford POS tagger. "Alkhalil" is a context free analyzer. Which means it does not consider the context of the word it is processing. On the other hand, POS tagger consider the context. When merging the output of the two systems we can have a contextual morphological and syntactical solution. This filtering is applied by comparing the POS-tag, retrieved from Stanford system with Alkhalil solutions and then choosing the solutions, which are compatible with the Stanford tag.

B. Pro-forms and Noun Entities Extraction

The second stage that comes after preprocessing is the extraction of the needed entities, which are pro-forms and noun entities. Both of them have different characteristics that is why each has a different approach for extraction.

In the second stage, we start with extracting pro-forms. The approach uses the output of "Alkhalil" to distinguish between different types of pro-forms, connected and separated pro-form. Connected pro-forms are attached to Arabic words, such as the suffix "ه" in the word "سألته". Separated pro-forms are notattached to word such as "هو""He". This is can be done with the help of a predefined list of Arabic pro-forms.

The second step in this stage is to find the set of all possible related noun entities (NEs). In the approach, we apply Named-entity recognition (NER) to extract the noun entities in the following steps. First, we prepare a "gazetter" which is a list or a corpus of different entities' names such as names of persons, locations, and organizations. Then we compare the text with the gazetter and do an initial NE tagging. Last step is to use a set of regular expressions to extract possible NE

C. Morphological Analysis

In the next stage is finding the morphological and syntactical features for pro-forms. We can simply do that using a lookup-table approach since the number of possible pro-forms is limited in Arabic language. Then, we find the morphological features for each possible NE, using the output of the morphological analyzer and the output of POS-tagger. The next step is to filter NE set by removing all inconsistent NEs in terms of morphological and syntactical features. We will call the output of this stage the PCNE (the possible consistent NE).

D. Finding the Related NE and Pro-forms

In this stage, dependency trees are usedto relate NEs and pro-forms. First, we find the dependency tree using Stanford Core NLP. If there is a path in the dependency tree between the pro-form and some PCNE, we choose the one with the shortest path in dependency tree graph. Otherwise, we find the nearest NE in term of number of words between the pro-form and each PCNE.

E. Output Validation

The output of the fourth stage can be considered as the final output showing the coreference between different entities. An extra step is performed to reach a more accurate result, which is validation stage. We run different tests on the related entities to validate the relation between them. For example, we check if the pro-form and the noun entity have the same gender, and number. By the end of this stage, we would have completed our model and reached the final output by relating noun entities to pro-forms.

F. Example

In the following example, a text input of an Arabic sentence that goes through the different stages of the approach starting with preprocessing until defining the related pro-forms and noun entitiesis shown.

(8) محمد هو الطالب الأفضل في الجامعة

Mohamed is the best student in the university.

The tables below show the output of different stages using example 8. Table 1 is the output after applying the first step.After applying the second stage, there is only one pro-form, which is "هو", or "he" and two NEs are found. Which is shown in Table I. After completing all the stages, in the last step, there is a path in the dependency tree between "هو" and "الطالب" or "he" and "student". Therefore, the result is, "الطالب" and "هو" are related.

TABLE I. FIRST STAGE OUTPUT

[1] اللاحق Suffix	[2] الحالة الإعرابية POS Tags	[3] الجذر Root	[4] الوزن Pattern	[5] نوع الكلمة Type	[6] الجذع Stem	[7] السابق Prefix	[8] الكلمة المشكولة Voweled Word
[9] #	[10]#	[11]#	[12]#	اسم علم[13]	محمد[14]	[15]#	مُحَمَّد[16]
[17]#	[18]#	[19]#	[20]#	ضَمِير الغائِب – للمُفرَد المُذَكَّر [21]	هو[22]	[23]#	هُوَ[24]
[25]#	مفرد مذكر [26] مرفوع في حالة التعريف	طلب[27]	فَاعِلُ[28]	اسم فاعل[29]	طالب[30]	ال: [31] التعريف	الطَّالِبُ[32]
[33]#	مفرد مذكر [34] مرفوع في حالة التعريف	فضل[35]	أَفْعَلُ[36]	اسم تفضيل[37]	أفضل[38]	ال: [39] التعريف	الْأَفْضَلُ[40]
[41]#	[42]#	[43]#	[44]#	حرف جر[45]	في[46]	[47]#	فِي[48]
ة: تاء التأنيث [49]	مفرد مؤنث [50] منصوب في حالة التعريف	جمع[51]	فَاعِلَةٌ[52]	اسم جامد[53]	جامعة[54]	ال: [55] التعريف	الْجَامِعَةُ[56]

TABLE II. SECOND STAGE OUTPUT

[57]#	مفرد مذكر مرفوع في حالة التعريف[58]	طلب[59]	فَاعِلُ[60]	اسم فاعل[61]	طالب[62]	ال: التعريف[63]	الطَّالِبُ[64]
ة: تاء التأنيث[65]	مفرد مؤنث منصوب في حالة التعريف[66]	جمع[67]	فَاعِلَةٌ[68]	اسم جامد[69]	جامعة[70]	ال: التعريف[71]	الْجَامِعَةُ[72]

VI. RESULTS & DISCUSSION

In this approach, we mainly used two annotated corpora for development and testing. For this part, we built our own dataset using different types of emails. In addition, we used "AnATAr" corpus, which consists of 70 different texts of Tunisian books [17].

A. Customized E-Mail Corpus

We built a data set corpus consist of business communication e-mails and e-mails of social Activities. All the emails are in Arabic language. The Arabic Corpus has approximate 900 emails. Whichare classified according to different domains. Figure 3 graphically shows the categorization of emails among different domains.

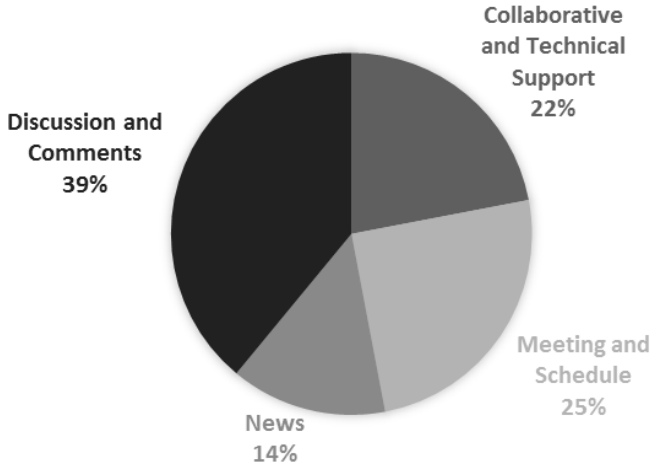

Fig. 3. Email Catergorization for Specific Corpus

B. AnATAr" Arabic Corpora

These Corpora were annotated using "AnATAr" tool, consist of "a technical manual, newspaper articles, texts of Tunisian books used for basic education" [17]. Anaphoric relations are annotated in the corpora where some of the pronouns where not included as they were cataphoric.

C. Results

The pervious corpora were used in the testing process. Precision, Recall and F Measure are used to evaluate the performance of our method. Assuming that the total number of pronouns in a text has given "R" results and the number of pronouns and referents extracted by the proposed algorithm are "X" of which "N" are correct, then precision is "N" divided by "X" and recall is "N" divided by "R" where F1 measure is calculated according to the following equation, $F1 = (2 * R * P)/(R + P)$. Table 3 shows the results that we got using our approach.

TABLE III. RESULTS

Corpus	R	X	N	Precision	Recall	F1
Customized	1053	1038	951	0.916	0.903	0.909
AnATAr	1148	1190	1018	0.855	0.886	0.87

The results show that the approach effectively and accurately was able to extract pro-forms and NEs while detecting the coreference relations between them. Another observation is that using morphological features along with dependency trees is a successful approach for coreference resolution. This approach was able to achieve high accuracy without the need to define a set of rules or the usage of large amount of annotated data for training. We can also note that the results for our customized set is more accurate than the "AnATAr" set. The main reason for the difference in the results is due to the difference in annotation scheme. In

example (9), the connected pronoun "ت" is not annotated in the"AnATAr" corpus where our approach recognize it as a pronoun and find the coreference relation for it which can affect the accuracy using this set.

(9) فاستلقيت على فراشي.

Then I slept on my bed.

Unfortunately, we could not compare our results to others for multiple reasons. First, annotated corpora and tools for Arabic language coreference resolution are very hard to obtain. Second, the available resources do not have the same scope and same evaluation methods for anaphora [1]. Third, the entities extracted and considered for coreference resolution such as type of pronouns are different from one approach to another. These reasons make comparing the results to other approaches very hard.

Both ruled based and machine learning approaches showed limitations with Arabic language conference resolution. The first type requires a large set of rules and the second needs annotated data, which add to the limitations of the approaches. The model proposed obtained all the results without the need of both a large set of rules or annotated data, which overcomes a great limitation of traditional approaches. Even a hybrid approach for Arabic conference resolution, which was suggested by 2012, CoNLL shared task targeted showed many limitations regarding Arabic language coreference resolution [2, 18]. The approach had average results of 60% where in Arabic it dropped to 45%, which means it did not calculate half of the relations right. We cannot compare the numbers directly, but our model does not require the resources that such approach needs and it shows positive results with average F1 score of 89%.

We observed multiple error sources. The complexity of the Arabic language was big challenge for the approach. For example, sometimes some parts of the words were identified wrongly as connected pronouns. Especially that in Arabic most of the connected pronouns are just one letter, which can be easily mistaken as part of any word. Another problem would be the ambiguity of some sentences. The scope of the approach is sentences with complete context, but this cannot easily be identified. An example of ambiguity, the word "أكل" which can be a verb or noun with the same letters, but the diacritics are different.

VII. Conclusion & Future Work

In this paper, we presented an alternative approach to coreference resolution in Arabic language using morphological features and dependency trees. The approach consist of five stages text preprocessing, pro-forms and noun entities (NEs) extraction, morphological analysis, relating NE and preforms, and output validation. For testing and evaluation, we designed a customized Arabic annotated corpus using different types of emails for coreference resolution and we used the "AnATAr" dataset. The results indicated the effectiveness of the approach.

In the future, we plan to expand the scope of the approach to include multiple sentences instead of just one sentence, which means we need to alter the structure of the model to be able to handle the new scope. We plan to explore new ways to improve our results, by for example, combining machine learning in our model. Machine learning can be used for improving the process of morphological analysis by learning new rules for the process. In addition, it can give indication about which morphological features or morphemes have more importance in the process of coreference resolution.

References

[1] A. Soraluze, O. Arregi, X. Arregi, and A. D. de Ilarraza, "Coreference chenResolution for Morphologically Rich Languages," Procesamiento de Lenguaje Natural, vol. 55, pp. 23–30, Sep. 2015.

[2] C. Chen and V. Ng, "Combining the best of two worlds: a hybrid approach to multilingual coreference resolution," CoNLL '12 Joint Conference on EMNLP and CoNLL - Shared Task, pp. 56–63, Jul. 2012.

[3] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60, 2014.

[4] I. A. Al-Sughaiyer and I. A. Al-Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," Journal of the American Society for Information Science and Technology, vol. 55, no. 3, pp. 189–213, 2004.

[5] I. Zitouni, J. Sorensen, X. Luo, and R. Florian, "The impact of morphological stemming on arabic mention detection and coreference resolution," Association for Computational Linguistics, pp. 63–70, 2005.

[6] J. Bajard, L. Didier and P. Kornerup, "An RNS Montgomery Modular Multiplication Algorithm," IEEE Trans. Computers, vol. 47, no. 7, pp. 766-776, July 1998.

[7] J. O. Williams, Narrow-Band Analyzer, Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

[8] J. G. Carbonell and R. D. Brown, "Anaphora resolution," pp. 101–96, Aug. 1988.

[9] J. Wolf and K. Pattipati, "A File Assignment Problem Model for Extended Local Area Network Environments," Proc. 10th Int'l conf. Distributed Computing Systems, pp. 221-230, 1990.

[10] K. M., A. Farghaly, and A. Aly, "Arabic Anaphora resolution: Corpus of the holy Qur'an annotated with Anaphoric information," International Journal of Computer Applications, vol. 124, no. 15, pp. 35–43, Aug. 2015.

[11] K. R. Beesley, "Arabic finite-state morphological analysis and generation," COLING '96 Proceedings of the 16th conference on Computational linguistics, vol. 1, pp. 89–94, May 1996.

[12] M. Boudchiche, A. Mazroui, A. Lakhouaja, A. Boudlal, and Mohamed Ould Abdallahi Ould Bebah, "AlKhalil Morpho Sys 2: A robust arabic morpho-syntactic analyzer," Journal of King Saud University - Computer and Information Sciences, May 2016.

[13] M.-C. de Marneffe and C. D. Manning, "Stanford typed dependencies manual," Sep. 2008.

[14] R. Al-sabbagh, "Arabic anaphora resolution using the web as corpus," in Proceedings of the seventh conference on language engineering, Cairo, Egypt, 2007.

[15] R. Roth, O. Rambow, N. Habash, M. Diab, and C. Rudin, "Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking," HLT-Short '08 Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, pp. 117–120, Jun. 2008.

[16] "Reviews: Morphology and syntax," in University of Duisburg-Essen. [Online]. Available: https://www.uni-due.de/SHE/REV_MorphologySyntax.htm.

[17] S. Hammami, L. Belguith, B. Hamadou, and Abdelmajid, "Arabic Anaphora resolution: Corpora Annotation with Coreferential links," International Arab Journal of Information Technology (IAJIT), vol. 6, no. 5, pp. 481–490, Dec. 2009.

[18] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang, "CoNLL-2012 Shared Task: Modeling multilingual

unrestricted coreference in OntoNotes," Proceedings of EMNLP and CoNLL-2012: Shared Task, pp. 1–40, 2012.

[19] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A machine learning approach to Coreference resolution of noun phrases," Computational Linguistics, vol. 27, no. 4, pp. 521–544, Dec. 2001.