A Comparative Mention-Pair Models for Coreference Resolution in Dari Language for Information Extraction

Ghezal Ahmad Jan Zia, Ahmad Zia Sharifi, Fazl Ahmad Amini, and Niaz Mohammad Ramaki

[1] Technical University of Berlin, Berlin Germany,
zia@campus.tu-berlin.de
[2] Nangarhar University, Nangarhar, Afghanistan
sharifi@nu.edu.af
Faculty of Literature, Kabul University
fazlamini@ku.edu.af
Kabul Polytechnique University, Kabul, Afghanistan
ramaki@kpu.edu.af

Abstract. Coreference resolution plays an important role in Information Extraction.This paper covers the investigation of two strategies based on a mention-pair resolver using Decision Tree classifier on structured and unstructured dataset, targeting coreference resolution in Dari language. Strategies are (1) training separate models which is specialized in particular categories (e.g., lexical, syntactic and semantic) and types of mentions (e.g. pronouns, proper nouns) and (2) using a structured dataset on a machine learning library that is designed to classify numerical values. Moreover, these modifications and comparative models describe a contribution of comprehensive factors involved in the resolution of texts. Specifically, we developed the first Dari corpus ('DariCoref') based on OntoNotes and WikiCoref scheme. Both strategies are produced f-score of state-of-the-art.

1 Introduction

Coreference resolution is one of the important tasks for Information Extraction (IE)[4]. It determines whether two expressions in natural language refer to the same entity in the real world. The person name which is a mention in a text, and a pronoun that refer to the same person is an example of coreference resolution. It finds and groups all the mentions in the text according to its referents.

The Message Understanding Conference (MUC) and Automatic Content Extraction (ACE) [13, 10] have used these two terms (a mention or an entity). A mention is a text based reference to an entity. For example, "Ahmad was here" is a named mentions, "The boy was here" is a nominal mentions, or "he was here" a pronominal mentions [5]. In coreference resolution, a noun phrase is called a mention or just called anaphoric noun phrase. In the pair of mentions, the first mention (full mention) is called the antecedent while the second mention is an anaphor.

Dhinaharan Nagamalai et al. (Eds) : SAI, NCO, SOFT, ICAITA, CDKP, CMC, SIGNAL - 2019
pp. 101-114, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90708

Natural language (NL) is inherently ambiguous, even it is not easy for humans to detect all hidden ambiguities existing in NL. In contrast, for a machines with less clever artificial knowledge about the world we live in, it is difficult to detect all ambiguities of the NL. Over the last decades, different techniques (rule-based and statistical) have been applied to coreference resolution and presented reasonably result [17][13]. Apart from English, Arabic, Chinese, Japanese have been addressed based on the features such as syntactic features, semantic feature. Our work is novel in that it is the first work that accomplished the use of mention pairs model on the Dari language. Dari and Pashto are the two official languages spoken in Afghanistan [**?**]. For both spoken languages, the work on coreference resolution is limited. There are no digital resources available to test on any NLP model.

In this paper, we focus on the important features such as semantic, syntactic and lexical features to accomplish the machine learning mention pair models on Dari language. In particular, a set of 10 features are proposed to identify Dari clauses and to group related entities as co-referent. Experimental results show the improvement of the performance for the Dari coreference resolution.

2 Related Research

Two important approaches used to solve coreference resolution. Linguistics-based that rely on linguistic and domain knowledge and the machine learning based that rely on data-driven approaches. In a linguistics-based approach, Hobb's algorithm [8] and (Lee et al.,) [9] are the leading approaches for pronoun and coreference resolution.

The statistical approaches for coreference resolution proposed by Soon et al., [17] and Ng and Cardie, [13] that used semantic and syntactic features. More intelligent research on coreference resolution based on neural network, deep learning, [7, 11, 2] have been done.

Moreover, this is the first research that we incorporated to build a system using machine learning to define all the mentions in real word entities for Dari language. In this research, we address the coreference resolution problem in the context of IE. The perspective of IE on coreference resolution constraint a limited scope on the set of entities to be resolved. We are only interested in resolving those entities to be extracted as part of a specific extraction and ignored the coreference resolution of those names, noun phrases and pronouns that are irrelevant to extraction task in hand.

3 Dari Corpus Annotation

In term of resources, Dari is a low-digital resource language. The annotation of Dari text for coreference resolution started with the hope that it will foster research

dedicated to this type of text. Therefore, for supervised learning approaches, we tend to create a balanced corpus in terms of article varities and length towards the newswire domain. We investigated to find the sources that follow the Dari language pure orthographic structure. The DariCoref, a Dari corpus, constructed purely from Azadi Radio news [16] and VoA articles [3], with the objectives of balancing the topics and text size. The DariCoref has been annotated methodically by efforts to embed the state-of-the-art tools.

3.1 Annotation Tools and Format

The corpus annotation with linguistic information requires syntax and semantic knowledge. In particular, the annotation is a complex and time-consuming task that needs a large size dataset to be annotated for the coreference resolution. The majority annotation tools outlined for the English language, and it was challenging to find adaptability for adapting the annotation tools for Dari language, as well as the required interaction for making annotation efficient, including a visual display of markable, a search function for text and more. Since there are many tools available as an open source and run on all major platforms, we examined Dari text with BRAT [19], eHOST [18], WebAnno [20], and MMAX2 tools [12].

MMAX2 is a highly adjustable tool to creating, browsing, visualizing and querying semantic annotations on various levels. It uses token standoff or a standoff file format where one file (the word files) contains a list of the tokens, while the other files (the markable files) contains one or multiple annotation layer [12]. In MMAX2, users at the same time can specify multi-layer coding scheme to cover multiple pointer views to track coreference chain membership.

In MMAX2, we can specify a set of attributes of markable in a schema file where we mainly follow the types of attribute based on OntoNotes and WikiCoref [5, 6]. Therefore, we first create the scheme file as the code is shown in Listing 1.1. Each mention is tagged based on *mention type*, and *coreference type*.

```
<?xml version="1.0" encoding="UTF-8"?>
<annotationscheme>
<attribute id="mention_type" type="nominal_button" name="mentionType">
<value id="mention_type_NE" name="NE"/>
<value id="mention_type_NP" name="NP"/>
<value id="mention_type_PRO" name="PRO"/>
</attribute>
<attribute id="coref_type" type="nominal_button" name="CorefType">
<value id="coref_type_identical" name="IDENT"/>
<value id="coref_type_attributive" name="ATR"/>
<value id="coref_type_attributive" name="COP"/>
```

Fig. 1: Example of DariCoref Annotation in MMAX2

```
</attribute>
</annotationscheme>
```

Listing 1.1: Coref scheme file

Mention Type:

- Named entity (NE): can be a person, organization, location, .., NPs or abbreviations referring to an entity in the real world.
- Noun Phrase (NP): a group of words headed by a noun, or pronoun, when they are not classified as NE.
- Pronominal (PRO): mentions tagged PRO as the following type:
  - personal pronoun,
  - reflexive pronoun,
  - demonstrative pronoun

Coreference Type: MUC and ACE apply *identical* and *attributive* schemes to mentions as a coreferential [13, 10]. In OntoNotes schemes, it is the opposite, which means it differentiates between these two attributes because of its different roles. In addition, the OntoNotes ignored the attributes signalled by copular structures where WikiCoref applied this feature [6]. Therefore, we decided to have the knowledge of all important attributes and applied the following coreference types on the Dari corpus.

- Identical (IDENT): In general, all referential mentions are tagged *identical*
- Attributive (ATR): As *attributive* all mentions in appositive
- Copular (COP): Attributive mentions in copular structures

The *two example* of coreference tagging:

- (Angela Merkel)$_{ATR}$, (Chancellor of Germany)$_{ATR}$
(انگلا مرکل) ,(صدراعظم آلمان)
- Professor Farooqi is (the chancellor of Kabul University)$_{COP}$
(پروفیسور فاروقی)رئیس پوهنتون کابل است

4 Feature Sets and Models

In the first defined strategy, we present a coreference resolution system based on the mention-pair model by [13] and [7]. They used the C4.5 decision tree induction system (Quinlan,1993) to train a classifier that decides whether two noun phrases or not in a document are coreferent.

For creating the coreference resolution model, and according to the Dari corpus structure (grammatical and morphological), we describe pairs of mentions with a set of 10 features [4]. They are commonly used in order to test if the antecedent noun phrase m_i is co-referent to the noun phrase m_j. Moreover, this pair of mentions divided into three categories (lexical, syntactic, and semantic) [17]. Lexical features formed information about the number, gender, distance, and all matching based features such as string matching, etc. In syntactic features, it provides information about the grammatical roles of the mentions. Semantic features checks mention referring to the people, organization or location that belongs to the same semantic class [7]. To carry out the resolution procedures, briefly, we describe the features we extract from DariCoref dataset in Table 1.

5 Decision Tree

To accomplish the classification, we selected the Decision Tree Algorithm, a Machine Learning model adapted to determine characteristics of the coreference resolution task [17]. The decision tree is one of the most popular classifiers introduced by Quinlan [15]. Moreover, it is composed of important characteristics, such as feature selection, handling continues and discrete attributes, managing unknown values, and etc.

In the Decision tree structures, leaves represent classifications (value of a class attribute) and branches represent conjunctions of features that significantly contribute to those classifications [7]. In the learning phase for the decision tree, we used the C4.5 algorithm by giving the training examples.

Feature	Description
Distance	the distance between the anaphor and antecedent, measured by the number of sentences that separate the mentions between them.
IsPronoun	This feature is to true if a noun phrase is a pronoun. We compared the noun phrase with possible Dari pronouns such as, (او، ما، تو، شما، وی، ایشان، آنان ، آنها، خود، خویشتن، خویش، این، آن، اینان، اینها)
String Match	If the anaphor or the antecedent strings match between the current mention and another phrase in the previous context.
Demonstrative NP	This feature checks if both noun phrases are demonstrative. (این، آن، اینان، آنان، اینها، آنها)
Number Agreement	Initially, we check a noun phrase in a listing of known singular or plural pronoun in Dari language. Singular Pronoun: (من، ما، تو، شما، او، این، خویش، آنان، آنها، اینها، اینان) and Plural Pronoun: (ایشان، خود، خویشتن،
Semantic compatibility	This feature checks to assess whether the two mentions are semantically compatible.
Gender agreement	For the gender (male, female), we compare a noun phrase with a hand-generated list of male and female pronouns. Further, we obtained the list of common male and female names from Kankor, the National Entrance Exam database for public universities.
IsProperNoun	In English or German languages, the first uppercase character presents the proper noun. For Dari language, there is no uppercase or lowercase characters. We used our own 'HMM-Based Dari Named Entity Tagger', to extract the proper nouns from the text.
Appositive	It check if the anaphor is an appositive of the antecedent candidate. Basically, we do not have access to the syntactic structure, we utilize heuristics (e.g., the existing of a comma between the two entities) to extract this feature.
Alias	whether the anaphor is an alias of the antecedent or vice versa.

Table 1: Features employed in our mention pair model [7]

In addition, features are the main subject for the coreference resolution in the decision tree. Therefore, our system relies on the three categories (lexical, syntactic and semantic) features associated with the mentions. Thus, the tree structure builds by sorting them down the tree from the root to some leaf node, in the resulting, the decision tree indicates whether two noun phrases are coreferent or not (0 or 1). For example, the starting point is the root node, if the training examples belong to the same class, further division is not necessary and this node becomes a leaf node referring to that class of examples. Moreover, if the training examples have the same feature values rather than equal class value, the possibility of the division is rare. Otherwise, the best division of training examples is chosen based on feature sets and from the feature set, a feature that gives highest information gain. Therefore, for building such a tree, the C4.5 algorithm examines the differences in entropy that defines randomness of the data by choosing a feature when generating sub-lists [15, 7].

$$Entropy = \sum -p(x).log_2 p(x) \tag{1}$$

The entropy is just a metric which measures the impurity in a collection of the training dataset. Therefore, a measure *information gain* can be defined for the effectiveness an an attribute for classifying the training data.

$$IG(A, S) = E(S) - \sum p(t)E(t) \tag{2}$$

*IG(A,S) = Entorpy(S) - (Weighted Avg) * Entropy(each feature)*

We believe this strategy can be applied to all dataset with similar structures and scenario.

5.1 Test Data and Error Analysis

The training and testing documents for several NLP tasks evaluated the performance of the system by the parameters defined by MUC *(precision, recall and f-measure)* [1].

For the training of our model, we split the dataset (DariCoref) with the size of 20K tokens into training (80%) and testing set (20%) to measure the model performance. The performance of the model according to the training and testing data shows low accuracy. This is because of the poor generalization of the tree structure and it is called overfitting. Therefore, it appears to optimize the generalization capabilities and removing parts of the tree that do not provide power to classify instances.

5.2 Pruning

To improve the performance of the resolution, it needs to do the pruning technique associated with decision trees. There are several pruning techniques, we used the Pessimistic Error Pruning (PEP) that based on that we avoid some sub-tree (rule) [15]. The tree which produced the misclassification rates on its training data are overly optimistic and if it utilized for pruning, produce overly large trees. To obtain a more realistic estimating generalization errors, if $N(t) =$ *the total number of training examples at node* t, and in consequence, $e(t) =$ *represents number of examples not classified to the majority class at node* t, then $r(t) = \frac{e(t)}{N(t)}$ is an estimate of the misclassification error rate in a single node t. The rate with the continuity correction is given as:

$$r'(t) = \frac{e(t) + 0.5}{N(t)} \tag{3}$$

For the whole sub-tree, T_t the misclassification error rate can be calculated as, $r(T_t) = \frac{\sum e(i)}{\sum N(i)}$ where i covers the leaves of the sub-tree. Thus the corrected misclassification rate after continuity correction will be

$$r'(T_t) = \frac{\sum(e(i) + 0.5)}{\sum N(i)} = \frac{\sum e(i) + N_t * 0.5}{\sum N(i)} \tag{4}$$

where N_t is the number of leaves. Moreover, concerning in (3) and (4), $N(t) = \sum N(i)$, as they refer to the same situated of examples; normally, the rate might be clarified to numbers of misclassification: $n'(t) = e(t) + 0.5$ for a node and $n'(T_t) = \sum e(i) + N_t * 0.5$ for a sub-tree. Normally, a tree is pruned and a node t becomes a leaf if the $n'(t) \leq n'(T_t)$ holds. However, this might even now infrequently occur, resulting in a quite optimistic pruning. For that reason, Quinlan suggests weaker condition: $n'(t) \leq n'(T_t) + SE(n'(T_t)$, where,

$$SE(n'(T_t)) = \sqrt{\frac{n'(T_t) * (N(t) - n'(T_t))}{N(t)}} \tag{5}$$

is the standard error for the sub-tree (T_t) [15].

Furthermore, there is another simple way that defined by [7] with testing different rules. To overcome with overfitting, we could prevent the algorithm commonly by stopping condition for a node. It ought a chance to prevent, if all the instances depend to the same class or all the attribute values are the same and maybe some more restrictive condition. By illustrating major rules it will be not difficult to discover which features are utilized the highest accurate rules and which were generally trivial.

Rule :
DIST = 0
IPRO = 0
DEM = 1
SEMCL = 1
APP = 1
->class 1 [95.7%]

This rule denotes that the nouns are co-referent (class 1), if the antecedent is not a pronoun, ($IPRO = 0$), both are demonstrative nouns ($DEM = 1$), both belong to the same semantic class ($SEMCL = 1$), 'j' is appositive to 'i' ($APP = 1$). Similar to our preliminary experiments, the above rules shows that the lexical features and semantic features generate a good result. The test results in Table 2 presents the best precision and best recall at the same time comparing to other models.

Dataset	Precision	Recall	F-measure
DariCoref	89.7	61.2	73.1
WikiCoref	87.9	59.3	70.6
MUC-6	83.1	52.3	64.1
Soon et.at.	67.3	58.6	62.6

Table 2: Results on the entire DariCoref, WikiCoref, using MUC metrics.

By analyzing our results on the training data, and according to the feature sets described before, we created another test data (unseen) 5K tokens, based on DariCoref corpus and deployed this data on the model as trained before. The model searches for the corresponding phrase pair in the test data to define whether the model predicts correctly or not. As shown in Table 3, the pairs of m_i and m_j is defined the contribution for each feature.

Coreferent	— Not Coreferent
760	19240
210	4780

Table 3: Confusion Matrix - illustrate the number of coreferent or not coreferent entities

Our observations led us to the conclusion, that after implementing the pruning techniques, the model leads to higher precision and recall. The best score is achieved by the (lexical and semantic) feature based system as demonstrated on (class 1), by including the syntactic features reduced the accuracy for the resolution. Similar, semantic features improve the recall but reduce the precision. Furthermore, to precisely analyze the errors and have a closer look at the model outputs or to examine the actual noun phrase pairs, in some steps the model classified incorrectly, for example, if a noun phrase close enough, and agree in number, it is still feasible that the model generate not co-referent mentions. Such as in this sentence: واسیلی، نماینده روسیه در سازمان ملل گفت که امنیت افغانستان.... *(Vasili, Russian's envoy in the United Nation said that security of Afghanistan ...)*. In this example, واسیلی *(Vasili)* and نماینده روسیه در سازمان ملل *(Russian's envoy to the United Nation)* is an attributive but our classifier connected the antecedent $'m_i'$ *(Vasili)* to a pronoun *(he)*: (واسیلی، او) from the next sentences. Because, in the model we defined mention pairs, where in some sentence we have more than two mentions that refer to each other. Therefore, with the help of rules defined in [7, 13] and the pruning technics, we reached the following conclusions:

- Most important feature Distance and string-match.

- The demonstrative pronoun is not important.
- In some rules, the appositive feature worked well. As the dataset was based on the newswire domain and it works perfectly.

6 Decision Tree Classifier using Scikit-learn

Our next strategy is to use the same model, DectionTreeClassifier based on Scikit-learn, specialized free software machine learning library that is designed to classify numerical and scientific libraries such as NumPy, and SciPy [14]. It's an extremely intuitive way to classify or label objects and the binary splitting makes it more efficient in a well-constructed tree.

For designing the model, first, we changed the format of our dataset (DariCoref) to a structured dataset of 1K words as shown in Table 4. It contains both antecedent and anaphora and their co-referent relation specified by target value *(Set#)*. In the *Anaphor & Antecedent* columns, we added all the sample of nouns, noun phrases and pronouns. In the *NE* column, we defined its type, such as named entities, pronouns, etc.

Anaphor & Antecedent	NE	Coreference	Sets#
محمد اشرف غنی	ne	Coref	Set_0
رئیس جمهور افغانستان	ne	Coref	Set_0
اعضای صلح هلمند	ne	Coref	Set_1
آنان	pro	Coref	Set_1
...	...	...	Set_n

Table 4: Structured Dari dataset.

In addition, the DecisionTreeClassifier in Scikit-learn classifies based on numerical value and our dataset is in categorical format. We used the LabelEncoder to encode them to numerical values. Furthermore, we divided the dataset into training (80%) and testing (20%). Below Table 5 shows the *precision, recall* and *f1-score* on each sets. We plot the mention pairs next to each other to display their referent as in Figures 2, 3. The shapes of referents and the distribution of data confirm that our algorithm learns correctly.

7 Conclusion and Future Work

In this paper, we introduced a solution to one of the major and complex IE tasks. Moreover, Dari language is lack of annotated corpus for the task of coreference resolution. It will be a resource and more demanding for researchers and linguistics who are investigating in NLP. Furthermore, we have discussed the development process

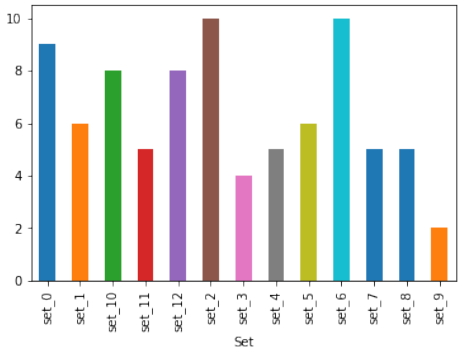

Fig. 2: Taget values

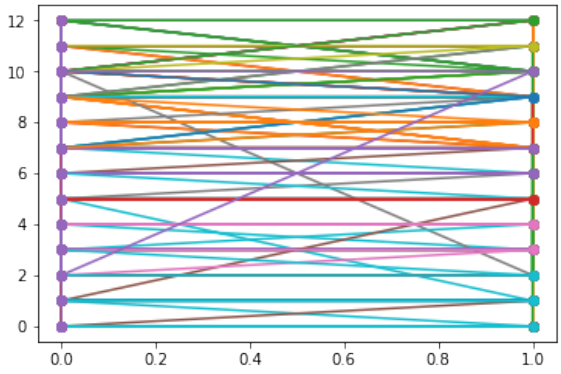

Fig. 3: The relation of Antecedent and Anaphora

	precision	recall	f1-score	support
0	0.53	1.00	0.69	9
1	0.57	0.67	0.62	6
2	0.45	0.62	0.53	8
3	0.40	0.80	0.53	5
4	0.75	0.86	0.80	7
5	1.00	0.80	0.89	10
6	1.00	0.25	0.40	4
7	0.67	0.40	0.50	5
8	1.00	0.67	0.80	6
9	1.00	0.10	0.18	10
10	1.00	1.00	1.00	5
11	1.00	1.00	1.00	5
12	1.00	1.00	1.00	2
avg/total	0.79	0.68	0.66	82

Table 5: Accuracy Rate on Structured Dari dataset.

of coreference resolution model based on two effective methods (Decision Tree algorithm) using structured and unstructured datasets. In our model, we followed the machine learning procedures that it relies basically on the following steps: Preprocessing modules, feature sets, learning algorithm and training examples. Therefore, each component interrelates with all others and significantly important for coreference resolution. If any component fails, it will lead to a loss of performance.

During the development of features, we devised our own techniques for the evaluation of features, to make it capable of Dari text. To the best of our knowledge, this is the first attempt incorporating lexical and semantic feature into coreference resolution on the Dari. The results of our model are promising that is comparable

in performance to best relevant models based on decision tree or non-learning. Moreover, we have proved that error pruning is applicable to the classification of mentions. For improving the reliability of the dataset and model, it will be good to test it with another model such as mention ranking model.

Bibliography

[1] Nancy Chinchor. Overview of muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*, 1998.

[2] Kevin Clark and Christopher D Manning. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*, 2016.

[3] VOA Dari. Voice of America, Dari. `https://www.darivoa.com/z/3018`, 2018. [Online; accessed 20-Sep-2018].

[4] Jenny Rose Finkel and Christopher D Manning. Enforcing transitivity in coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 45–48. Association for Computational Linguistics, 2008.

[5] Abbas Ghaddar and Philippe Langlais. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *LREC*, 2016.

[6] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: The 90\% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, 2006.

[7] Sanghoon Kwak and Takahiro Aoyama. Coreference resolution with decision tree. Technical report, Technical report CS224N, Final Project Stanford University, Spring, 2008.

[8] Shalom Lappin and Herbert J Leass. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561, 1994.

[9] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task*, pages 28–34. Association for Computational Linguistics, 2011.

[10] Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 135. Association for Computational Linguistics, 2004.

[11] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[12] Christoph Müller and Michael Strube. Multi-level annotation of linguistic data with mmax2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214, 2006.

[13] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 104–111. Association for Computational Linguistics, 2002.

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[15] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[16] Azadi Radio. Azadi Radio, Dari. `https://pa.azadiradio.com/z/2120`, 2018. [Online; accessed 20-Sep-2018].

[17] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544, 2001.

[18] Brett R South, Shuying Shen, Jianwei Leng, Tyler B Forbush, Scott L DuVall, and Wendy W Chapman. A prototype tool set to support machine-assisted annotation. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 130–139. Association for Computational Linguistics, 2012.

[19] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

[20] Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, 2013.