Journal of Artificial Intelligence and Data Mining (JAIDM), x(x): xxx-xxx, xxxx

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Mehr: A Persian Coreference Resolution Corpus

Hassan Haji Mohammadi[1], Alireza Talebpour[2 *], Ahmad Mahmoudi Aznaveh[3] and Samaneh Yazdani[4]

1, 4. Department of Computer Engineering, North Tehran Branch, Islamic Azad University, Tehran, Iran.
2, 3. Department of computer engineering, Shahid Beheshti University, Tehran, Iran.

Article Info

Article History:
Received 12 February 2023
Revised 16 June 2023
Accepted 18 July 2023

DOI:10.22044/JADM.2023.12641.2418

Keywords:
Coreference resolution, Persian coreference corpus, Corpus annotation, Mention, Anaphora resolution, Antecedent.

**Corresponding author:*
Talebpour@sbu.ac.ir (A. Talebpour)

Abstract

Coreference resolution is one of the essential tasks in natural language processing. This task aims to identify all expressions within a text that refer to the same entity in the real world. Coreference resolution is utilized in various fields of natural language processing, including information extraction, machine translation, and question-answering.

This article introduces a new coreference resolution corpus in Persian called the "Mehr corpus." The primary objective of this article is to develop a Persian coreference corpus that addresses some of the shortcomings of the previous Persian corpora while ensuring a high level of agreement among annotators. The corpus annotates coreference relations for noun phrases, named entities, pronouns, and nested named entities. Two baseline pronoun resolution systems have been developed, and the results are reported. The corpus consists of 400 documents and approximately 170,000 tokens. Corpus annotation is performed using the WebAnno pre-processing tool. The experiments demonstrate that based on random forest, the mention-pair model outperforms the other sieve-based baseline models on the Mehr corpus, achieving an F1 score of 59.4%.

1. Introduction

Coreference resolution involves identifying all the mentions in the text that refer to the same real-world entity. This task is utilized in various fields of NLP including machine translation, summarization, sentiment analysis, information extraction, and question-answering [1]. The research direction in this field has shifted from rule-based approaches [2] and early machine-learning systems [3] to state-of-the-art deep-learning systems [4-8]. End-to-end deep learning coreference resolution systems were introduced by Lee *et al*. [9], and this architecture has been used in English and other languages [10, 11] to develop coreference models.

The sixth message understanding conference (MUC-6) [12] introduced the first English corpus for coreference resolution. This corpus consists of 25 documents from Wall Street Journal articles, containing approximately 30,000 tokens. This dataset served as a standard for comparing coreference resolution systems until the introduction of the multilingual ACE 2005 corpus [13]. Over the years, numerous coreference resolution corpora have been developed in English and other languages. Among them, the widely recognized and influential corpus is OntoNotes 5.0 [14]. This corpus is designed for three languages: English, Chinese, and Arabic. OntoNotes 5.0 has emerged as a standard for comparing coreference systems. The following section will present the corpora developed in both English and Persian languages.

This article introduces the Mehr corpus, a Persian coreference resolution corpus. Additionally, two baseline pronoun resolution systems are implemented on this corpus, and the results are reported. Implementing the pronoun resolution system aims to demonstrate the corpus's capability to label various types of pronouns accurately. The Mehr corpus consists of 400 documents,

comprising approximately 170,000 tokens. The corpus documents are divided into two sections: the training section and the test section. Like the MUC6 corpus test section, the Mehr test section contains around 13,000 tokens.

In this corpus, all noun phrases, named entities, pronouns, and nested named entities that participate in a coreference chain are labeled. In addition to the manual labels, specific automatic labels have been included in the corpus. These labels are extracted from the output of the Persian NLP preprocessing tool, which is further explained in the rest of the article. Subsequently, two pronoun resolution baseline models are developed to demonstrate the corpus's practicality.

The primary contribution of this work is the introduction of the Mehr corpus, a pioneering Persian coreference resolution corpus. By filling a significant gap in available resources for the Persian language, the Mehr corpus enables accurate labeling of various pronouns, noun phrases, named entities, and nested named entities participating in coreference chains. Furthermore, incorporating both manual and automatic labels and implementing baseline pronoun resolution systems showcase the corpus's practicality and utility. The comprehensive analysis and experimental results obtained from the Mehr corpus provide valuable insights into the performance of baseline models and facilitate advancements in Persian coreference resolution. Overall, this research significantly contributes to the field by providing a vital resource and paving the way for future studies and improvements in Persian coreference resolution.

Section 2 analyzes the famous English and Persian coreference corpora, while Section 3 describes the presented Mehr corpus. Section 4 presents two baseline pronoun resolution systems, and all experiments and results on the Mehr Corpus are reported. Finally, Section 5 covers the released resources, and Section 6 concludes the paper.

2. Related works

This section reviews related works to this article, focusing on prominent corpora in English and Persian. Firstly, we will examine popular English coreference resolution corpora. Then, the Persian coreference corpora will also be reviewed.

2.1. English coreference resolution corpora

Over the years, several corpora were created, initially in English and subsequently in other languages. The development of coreference resolution corpora serves two primary goals. The first goal is to develop machine learning models and systems using large amounts of data. The second goal is to utilize these corpora as valuable linguistic resources.

The sixth Message Understanding Conference presented the first comprehensive coreference corpus in English. The MUC-6 and MUC-7 [15] corpora have been suitable references for comparing coreference resolution systems. Following the introduction of the ACE-2005 corpus [13], the MUC corpora are no longer utilized as comparison references for coreference models, and the ACE corpus has taken its place in the development of coreference resolution systems. The MUC corpus is small, and divided into train and test sections. This corpus annotates relatively small noun phrases within the text. The only coreference relation labeled in this corpus is the identity relation. One of the limitations of the MUC corpus is that it does not annotate singleton mentions.

Due to the limited size of the MUC corpus, the ACE corpus was introduced as a replacement. Developed between 2002 and 2008, the ACE corpus encompasses English, Arabic, and Chinese languages. While the ACE corpus aimed to address the flaws of the MUC corpus, it also encountered particular challenges. The first issue is that it only annotates seven named entity types, thereby restricting coreference systems developed on this corpus to those specific entity types. Another problem arises from the need for more specification regarding the train and test sections of the corpus, which makes it challenging to compare the systems developed using ACE.

The shortcomings of the MUC and ACE corpora prompted the development of the OntoNotes 5.0 corpus [14]. While the ACE corpus was introduced to address the limited size of the MUC corpus, it faced limitations in labeling only a limited number of entities. Additionally, both corpora suffered from low consistency in coreference annotations, affecting the inter-annotator agreement [14]. Another challenge revolved around evaluating coreference resolution systems, as different evaluation metrics, scenarios, and train-test splits made it difficult to compare such systems.

The CoNLL-2012 corpus was developed as a benchmark for comparing coreference resolution systems, building upon the OntoNotes 5.0 corpus. This corpus comprises a substantial number of train and test documents. In conjunction with the introduction of this corpus, the authors of the article presented the CoNLL standard, which has served as a benchmark for comparing coreference resolution systems to this day. The CoNLL-2012 corpus consists of approximately one million

tokens for English and Chinese languages and 300k tokens for Arabic. It annotates eighteen named entities and includes only two coreference relationship types: identity and appositive.

In addition to the corpora mentioned above, there have been efforts to develop task-specific coreference resolution corpora. One example is the BUG [16] corpora. However, most of these corpora have a limited token size, and are focused on specific coreference domains, making them unable to serve as replacements for the CoNLL-2012 corpus. Table 1 provides a comparison of English corpora.

2.2. Persian coreference resolution corpora

Producing corpora in Persian presents more significant challenges due to the language's structural properties. However, there have been efforts to develop Persian corpora [17], specifically for coreference resolution in Persian in the recent years. This section provides a detailed examination of these coreference resolution corpora.

The first pronoun resolution corpus in the Persian language is the PCAC-2008 [18]. This corpus comprises 30 articles extracted from BijanKhan's corpus[19]. The authors labeled the antecedent of 2006 pronouns. However, it is worth noting that this corpus only annotates a limited number of pronoun antecedents, making it unsuitable for developing comprehensive coreference resolution systems.

The first coreference resolution corpus in the Persian language is the Lotus corpora [20]. This corpus utilizes 50 significant texts from BijanKhan's corpus to label in-text mentions. In another effort, Mirzae and Safari [21] developed a coreference corpus called PerCoref, which consists of 547 documents and 212,646 tokens. The corpus has been manually annotated with morphological and syntactic information. It includes labeled mentions of pronouns, noun phrases, and verbs in null-subject sentences. The corpus covers various antecedent types, such as anaphoric, cataphoric, indirect, and no-reference mentions. In addition to annotating the identity relation, the corpus also includes labels for inferred, event, quantifier, cross-speech, and person/number suffixes on verb relations. The most recent Persian coreference resolution corpus is the RCDAT corpus, introduced by Rahimi and HosseinNejad [22].

Table 1 . A comparison between English coreference resolution corpora.

Corpus name	Reference	Year	Annotation format	Number of docs	Number of tokens
MUC-6	Grishman *et al*. [12]	1996	SGML	60	25 k
MUC-7	Hirschman *et al*. [15]	1997	SGML	50	40 k
ACE	Doddington *et al*. [13]	2005	ACE	666	303 k
Onto Notes 5.0	Pradhan *et al*. [23]	2007	Conll	2384	1600 k
CoNLL 2012	Pradhan *et al*. [14]	2012	Conll	2384	1600 k
ECB$^+$	Cybolska *et al*. [24]	2014	ECB$^+$	582	-
CIC	Chen *et al*. [25]	2016	Conll	606	195 k
Wikicoref	Ghaddar and Langlais *et al*. [26]	2016	Conll	30	60 k
GUM	Zeldes *et al*. [27]	2017	Conll	25	226 k
Knowref	Emami *et al*. [28]	2018	Winograd	-	-
Preco	Chen *et al*. [29]	2018	Conll	37.6 K	12.33 m
LitBank	Bamman *et al*. [30]	2019	Brat	100	210 k
BUG	Levy *et al*. [16]	2021	BUG	-	108 k

This corpus comprises one million tokens and is formatted in the CoNLL format, a standard format for representing linguistic data. The authors state that the texts in the corpus were extracted from various domains including political, cultural, economic, and sports news sites. The gold labels include coreference, phrase types, named entity tags, and animacy tags. In addition to the manually assigned labels, the corpus contains some automatic labels obtained from NLP preprocessing tools. Furthermore, specific labels obtained from the output of NLP preprocessing tools have been automatically extracted and added to the corpus.

Each line of the RCDAT corpus contains the following information: document name, sentence number, token, pos tag (16 labels), named entity (12 gold labels), token stem or itself, token without prefix and suffix, named entity (3 gold labels), coreference chain number (gold), animation, phrase type, and pos tags (100 labels).

3.Mehr corpus

The MEHR corpus consists of 400 documents obtained from the MEHR News Agency website at different intervals.[1] The corpus is divided into train and test sections. Table 2 presents statistical information about the Mehr corpus including the number of documents, sentences, and tokens for the train/dev and test sections. The test section contains approximately 13,000 tokens, similar to the test section of the MUC corpus. The Mehr corpus is publicly available[2].

The Mehr corpus is developed following the CoNLL standard, which was introduced in reference [14]. The CoNLL standard is commonly used to refer to the TSV (Tab-Separated Values) format in the field of Natural Language Processing (NLP). The origin of this format is various NLP tasks that have published their format in this way. In this format, each token is displayed on a separate line, and each column represents a specific annotation. The authors of the Mehr corpus utilize the WebAnno pre-processing tool [31] for annotating the corpus. Upon annotating the document, the WebAnno tool automatically generates the CoNLL file as its output. Figure 1 provides an example of the annotation steps performed using the WebAnno tool.

Table 2. Statistical information of Mehr's corpus.

	Document counts	Sentence counts	Token counts
Train and development	357	3189	158047
Test	43	286	13214
Total	400	3475	171261

In a CoNLL document, mentions within a coreference chain are assigned a shared index label. This means that all mentions within the same chain receive the same index. For instance, in the sentence "Ali came to school. He brought a book with himself," the mentions of **Ali**, **he**, and **himself** form a coreference chain with a common index. Any mention referring to "Ali" would be added to this coreference chain throughout the rest of the document. The Mehr corpus does not annotate noun phrases that do not participate in coreference chains, known as singletons. However, the corpus does annotate singletons for specific types of Persian pronouns including personal, demonstrative, and reflexive pronouns (unlike the RCDAT corpus). The pronouns whose antecedent is not in the text or have no reference are specified in the Mehr corpus. Figure 2 shows the frame of the Mehr conll file.

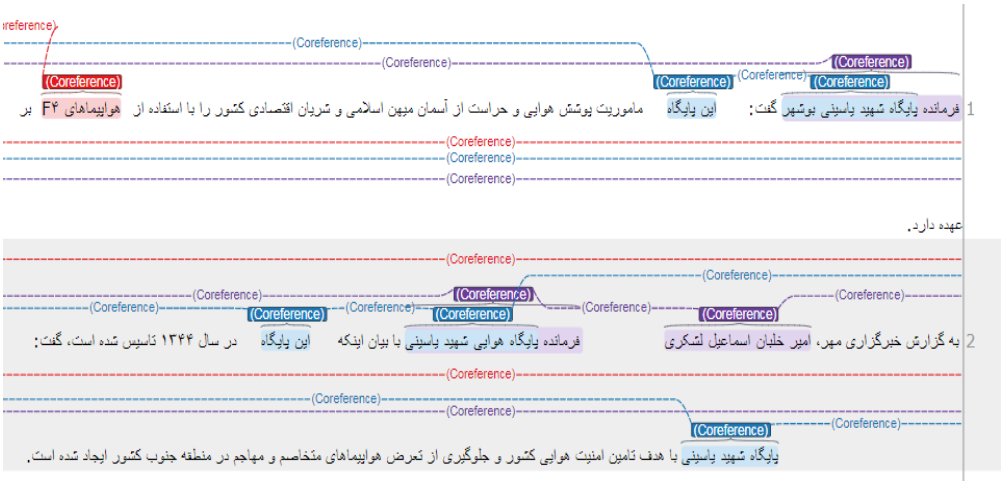

Figure 1. A frame of WebAnno annotation tool.

3.1. Mehr corpus annotation layer

The Mehr corpus consists of two types of labels: gold and automatic. This section examines both types of labels in detail.

3.1.1 Gold labels

The Mehr corpus includes annotations for identity coreference resolution relations involving pronouns, noun phrases, named entities, and nested named entities. In Persian, some pronouns have their antecedents removed from the text, with the antecedent typically being the speaker of the phrase. For instance, in the sentence "We should be proud of Iran," the antecedent of the pronoun "we" has been omitted. Unlike the RCDAT corpus, the Mehr corpus treats this type of pronoun as a singleton, indicating that it does not have a corresponding antecedent within the text.

[1] https://www.mehrnews.com/

[2] https://github.com/hasanhasanhaji/Mehr_corpus

9)	*	-	-	-	-	-	-	عادل	admin	0	0
(9	*	-	-	-	-	-	-	دهدشتی	admin	0	1
9)	*	-	-	-	-	-	-	معاون	admin	0	2
-	*	-	-	-	-	-	-	اسبق	admin	0	3
-	*	-	-	-	-	-	-	سیاسی	admin	0	4
-	*	-	-	-	-	-	-	و	admin	0	5
-	*	-	-	-	-	-	-	اجتماعی	admin	0	6
-	*	-	-	-	-	-	-	فرماندار	admin	0	7
(10)|(9	*	-	-	-	-	-	-	دشتستان	admin	0	8
-	*	-	-	-	-	-	-	به	admin	0	9
-	*	-	-	-	-	-	-	تحولات	admin	0	10
-	*	-	-	-	-	-	-	مثبت	admin	0	11
2)	*	-	-	-	-	-	-	دولت	admin	0	12
-	*	-	-	-	-	-	-	تدبیر	admin	0	13
-	*	-	-	-	-	-	-	و	admin	0	14
(2	*	-	-	-	-	-	-	امید	admin	0	15
-	*	-	-	-	-	-	-	در	admin	0	16
6)	*	-	-	-	-	-	-	استان	admin	0	17
(6	*	-	-	-	-	-	-	بوشهر	admin	0	18
-	*	-	-	-	-	-	-	اشاره	admin	0	19
-	*	-	-	-	-	-	-	کرد	admin	0	20
-	*	-	-	-	-	-	-	و	admin	0	21
-	*	-	-	-	-	-	-	گفت	admin	0	22

Figure 2. A frame of the Mehr conll file.

Non-referential pronouns are also labeled as singletons in the Mehr corpus. For example, the pronoun "it" in the sentence "it is an important issue" would be labeled as a singleton. Annotating non-referential pronouns as singletons enable training systems to identify and handle non-referential pronouns effectively.

In corpus creation, one important aspect to consider is the inter-annotator agreement, which measures the level of agreement between annotators. In the field of corpus production, the Kappa and MUC metrics are widely recognized. In the case of the Mehr corpus, two annotators were involved in the labeling process. The inter-annotator agreement, as measured by the Kappa metric, is approximately 91% for mention detection and about 93% for coreference resolution annotation.

3.1.2 Automatic labels

Automatic labels for the Mehr corpus are generated by Persian NLP pre-processing tools. The accuracy of these labels is dependent on the accuracy of the preprocessing tools utilized. This article employs the Hazm pre-processing tools[3], the comprehensive platform of intelligent content analysis[4], and the FarsiYar NLP tools[5]. Table 3 provides an overview of the accuracy of different NLP tools used in this study. It is worth noting that the entire corpus, comprising 171261 tokens, has undergone automatic morphological annotation including animacy, syntactic role, lemma, and part-of-speech tags.

Table 3. Statistical characteristics of Mehr's corpus.

Persian tools	Accuracy
Pos tagger	97.1
Shallow parser (Chunker)	89.9
Named-entity tagger	89
Pronoun recognition	96
Animacy recognition	91
Constituency parser	85.2

In addition to this, the pre-processing tool has automatically labeled a total of 44,960 noun phrases. Moreover, the corpus contains 11,175 automatically annotated named entities. Furthermore, human annotators have carefully labeled 19,833 in-text mentions, ranging from single tokens to multi-token expressions. This comprehensive annotation process enhances the richness and depth of the corpus, making it a valuable resource for studying and developing coreference resolution systems in Persian.

4. Experiments

In this section, two pronoun resolution systems are developed to demonstrate the practicality of the Mehr corpus. The first baseline system utilizes a mention-pair machine learning approach, where features are specifically designed for pronoun resolution. The second baseline system is built upon the Stanford sieve-based system, with sieves tailored to the Persian language. Before implementing the two baselines, the raw text undergoes preprocessing using NLP pre-

[3] https://www.sobhe.ir/hazm/

[4] http://catotalservices.com/UIHome/Index.aspx

5 https://text-mining.ir/

processing tools. The section proceeds by describing the two baseline systems in detail, followed by the presentation of the obtained results.

4.1. Mention-pair baseline

The pronoun resolution system in this approach is built upon a set of manually designed features. The feature set employed in this model encompasses pronouns, antecedents, and relational features. The initial step involves identifying mentions within the text, which are crucial for developing an effective coreference resolution or pronoun resolution system. In this research work, mentions include various types such as pronouns, noun phrases, named entities, and nested named entities.

During the training phase, a positive example is generated by pairing a pronoun with its closest actual antecedent. On the other hand, a negative example is formed by pairing the pronoun with all the mentions occurring between the pronoun and its true antecedent. The feature vector, as outlined in Table 4 , plays a crucial role in constructing the pronoun resolution system. It comprises the primary features contributing to the system's development and performance.

Once the feature vector is created, the next step is to select the best classifier for the dataset. The random forest classifier has been chosen as the best classifier through 10-fold cross-validation experiments. During the development process, the optimal hyper-parameters for this model are estimated.

The hyper parameters grid search, as provided in Table 5 , is performed to find the best combination. The best-obtained hyper parameters are 20 for the number of estimators, 1000 for the maximum depth, and "Gini" for the criterion. Table 6 presents the results of the 10-fold cross-validation performed on the training data, without creating chains.

The next step is to generate the output pronoun resolution chain after selecting the best classifier and hyper parameters using the training data. The pronoun resolution model aims to identify the best antecedent among the previous candidates for each pronoun in the test section. The model scans the text from right to left, pairing each pronoun with the preceding antecedent candidate. A feature vector is then created for each pronoun-antecedent candidate pair. This feature vector is subsequently fed into the random forest classifier to calculate the probabilistic output. Using the best-first clustering approach, the candidate with the highest probability output is determined as the pronoun's antecedent.

Table 4. Features used in the mention-pair model.

row	Description
	Pronoun features
1	Is it a personal pronoun?
2	Is it a demonstrative pronoun?
3	Is it a reflexive pronoun?
4	Is it a third-person pronoun?
5	Is it a speech pronoun?
6	Pos label of the first token on the left side of the pronoun
7	Pos label of the second token on the left side of the pronoun
8	Pos label of the third token on the left side of the pronoun
9	Pos label of the first token on the right side of the pronoun
10	Pos label of the second token on the right side of the pronoun
11	Pos label of the third token on the right side of the pronoun
12	Is pronoun subject?
13	Is pronoun object?
14	Pronoun number
	Antecedent features
15	How many tokens does the antecedent have?
16	Is the antecedent a pronoun?
17	Is the antecedent a demonstrative phrase?
18	Pos label of the first token on the left side of the antecedent
19	Pos label of the second token on the left side of the antecedent
20	Pos label of the third token on the left side of the antecedent
21	Pos label of the first token on the right side of the antecedent
22	Pos label of the second token on the right side of the antecedent
23	Pos label of the third token on the right side of the antecedent
24	Antecedent number
25	Is Antecedent subject?
26	Is Antecedent object?
	Relational features
27	Sentence distance
28	Token distance
29	Number agreement
30	Subject agreement
31	Object agreement
32	String match
33	Is the distance between the pronoun and antecedent candidate less than three?
34	Are the pronoun and the antecedent candidate in the same sentence?

Table 5. Random forest grid-search hyper-parameter values.

Hyper-parameter	Values
Max depth	None,10,20,30,40,50,60,70,80,90,100
Number of estimators	100,200,500,1000
Criterion	Gini, Entropy

Table 6. 10-fold cross-validation results using the best hyper parameters.

precision	recall	F1
85.65	64.46	73.52

The development section establishes a probability threshold for the pronoun's antecedent. If the antecedent candidates of a pronoun do not surpass this probability threshold, the pronoun is

considered non-anaphoric, meaning it does not have a corresponding antecedent.

4.2. Multi-pass sieve baseline

The architecture of this system is derived from the Stanford sieve-based model [32]. The architecture of this system follows the concept of using sieves, which are arranged in a sequential order based on precision. The system incrementally builds coreference partial clusters, where each mention is assigned a solution within the current sieve or transferred to subsequent sieves. Unlike the mention-pair model, features in this architecture are created at the cluster level, offering a more holistic approach.

The architecture is highly modular and comprises eight Persian sieves designed explicitly for the Persian language. The overall structure of the model, including mention selection within a given sieve, antecedent selection for a given mention, and feature sharing, follows the same principles as the Stanford sieve-based system. The Persian sieves used in this model are as follows:

- Sieve 1- Speaker sieve. In the Speaker sieve, all pronouns within a quote are linked to the speaker of that quote. This sieve relies on a set of handcrafted rules to identify the speaker of the sentence. For example, in the sentence "The president said: I will explain about this issue," the pronoun "I" is connected to the speaker, the president. Since the Mehr corpus text is non-conversational, specific handwritten rules are applied to determine the speaker, mainly focusing on identifying the verbs used in quotations, such as "said." These rules aid in accurately assigning the appropriate antecedent to the pronoun based on the speaker within the quote.
- Sieve 2- Exact string match. The Exact string match sieve creates a new cluster when two mentions have an exact string match. If two mentions in the text have identical strings, they are grouped together in a new cluster. For instance, in the sentence "Iran has many natural resources. Oil and gas are among the most important natural resources of Iran," the two occurrences of the noun phrase "Iran" form a new cluster since their strings are an exact match.
- Sieve 3- String head match .In this sieve, two mentions are placed in the same cluster if their string's heads are identical and satisfy some conditions. The sameness of the head string of the two mentions is not a sufficient condition for merging them. For example, two noun phrases, "Shahid Beheshti University" and "MIT University," have the same head, but they are not coreference. The two conditions mentioned are as follows. The first condition is that the head of both mentions is the same. For example, under this sieve, two noun phrases, "Iran court" and "Iran national court," are placed in the same cluster.
- Sieve 4– Proper name match: In this sieve, if both mentions are proper names and satisfy a series of hand-crafted rules, form a new cluster. For example, two mentions, "Rouhani" and "Hasan Rouhani," form a new cluster.
- Sieve 5– Location match: In this sieve, if two mentions are named entities and both have the “LOC" label, and one is a substring of the other, forming a new cluster. For example, two mentions, "Gorgan" and "Gorgan city," are merged under this sieve.
- Sieve 6– Title match: In many Persian texts, proper nouns come immediately or before their titles. For example, in the sentence "Ebrahim Raisi participated in this meeting as the president of Iran." The specific name "Ebrahim Raisi" and his title "President of Iran" are mentioned in the text. Under this sieve, proper names are placed in a cluster with their title by a series of rules.
- Sieve 7– Demonstrative phrase match: In the corpus text, there may be a demonstrative mention that refers to a previous mention in the text. For example, in the text, the noun phrase "car exhibition" may appear first, followed by the noun phrase "this exhibition." Under this sieve, these two mentions form a new cluster.
- Sieve 8– Pronoun resolution sieve: Previous sieves have incrementally formed coreference chains. In this sieve, the pronoun is connected to one of the previously created chains based on cluster-level features. After applying the pronoun to a previous cluster, the cluster's representative is considered the actual antecedent of the pronoun. Table 7 shows the features used in the sieve-based model for pronoun resolution. The features of a cluster are obtained as a union of the features of the mentions in that cluster.

Table 7. Features related to the pronoun sieve.

Feature name	Proposed system
Number	1- Static list for pronouns 2- NER labels 3- POS labels
Gender	1- Static lexicon 2- Static pronouns
Animacy	1- Static list for pronouns 2- NER labels 3- Static list for titles 4- Static list for Persian names 5- Entity label from Persian Corpus [33]
Distance	Less than 4 sentence

4.3. Results

This section compares the results obtained from the two baseline models with the Persian end-to-end system proposed by Rahimi and HosseinNejad. [22]. In the evaluation section, the test part of the Mehr corpus is utilized to assess the performance of the pronoun resolution systems. The test section of the Mehr corpus consists of approximately 13,214 tokens, which is comparable in size to the test section of the MUC-6 corpus. Table 8 presents the test results of the pronoun resolution systems, providing an overview of their performance.

The results clearly indicate that the mention-pair model exhibits higher efficiency compared to the rule-based system in the pronoun resolution task. One of the key factors contributing to this outcome is the accurate modelling of the relationships between features in pronoun resolution. The mention-pair model leverages the capabilities of the random forest classifier, which proves to be effective in modelling un-conjoined features. The reliable performance of the mention-pair model further underscores the dependability and suitability of the Mehr corpus for pronoun resolution tasks. These results highlight the efficacy of the mention-pair model and the robustness of the Mehr corpus as valuable resources for advancing pronoun resolution research in the field of Persian natural language processing.

The architecture of the baseline machine learning system presented in this paper is a Mention-pair model, which resembles the system developed by Rahimi and HosseinNejad. However, the baseline system proposed in this paper achieves higher accuracy by employing features that are specifically designed to capture the characteristics of pronouns, antecedents, and their relationships.

The feature set used in the machine learning model is rich and well-suited for the task at hand. By incorporating these informative features into the model, the overall efficiency and effectiveness of the system are enhanced. Utilizing a comprehensive and appropriate feature set contributes significantly to the performance of machine learning models.

Table 8. Reported results of the presented system on Mehr corpus.

System	Precision	Recall	F1
Sieve-based model	36.4	35.6	36
Mention - pair	60	58.82	**59.4**
System presented by [22]	59.5	54.96	57.14

5. Release of Resources

The corpus and pre-processing files mentioned in the paper are freely accessible and can be downloaded from the following GitHub repository: https://github.com/hasanhasanhaji/Mehr_corpus. The repository provides a convenient platform for researchers and practitioners to access the Mehr corpus and associated pre-processing tools. By making the corpus openly available, the authors promote transparency and facilitate further research and development in the field of pronoun resolution and coreference resolution. Interested individuals can visit the provided link to access the corpus and pre-processing files for their own analysis and experimentation.

6. Conclusion and Future Works

This research presents the Mehr corpus, a coreference resolution corpus in Persian. The corpus consists of 400 documents divided into train and test sections. It annotates various types of mentions including noun phrases, named entities, pronouns, and nested named entities. Unlike previous Persian coreference resolution corpora, the Mehr corpus also labels singleton pronouns. To demonstrate the usability of the corpus, two pronoun resolution systems were developed. The first system is a mention-pair machine-learning system, while the second system is a sieve-based model. The machine learning system utilizes random forest as the classifier and achieves an F1 score of approximately 60 points, indicating its effectiveness in resolving coreference. The Mehr corpus serves as a valuable resource for studying coreference resolution in Persian, and the

implemented systems demonstrate the practical applications of the corpus in developing accurate pronoun resolution models.

In contrast to previous Persian pronoun resolution systems that only provided statistical results, this article goes beyond by presenting the final coreference chains. The achieved 60% F1 score in the baseline Persian end-to-end pronoun resolution system is a notable efficiency, especially considering the inherent challenges in pronoun resolution. This baseline system can serve as a benchmark for future studies, allowing researchers to compare their results against this established performance. The availability of the Mehr corpus enables the training and development of more advanced models for coreference and pronoun resolution in Persian. The researchers can leverage this corpus to create higher-performing systems, leading to further advancements in the field of Persian language processing.

Pronoun resolution in Persian indeed presents several challenges. One of these challenges is the presence of pleonastic pronouns, where the pronoun "it" refers to an ambiguous antecedent. Resolving such cases requires understanding the context and disambiguating the potential referents. Another challenge in Persian pronoun resolution is the mismatch of gender and number between the pronoun and its referent in the text. The Persian language uses plural pronouns for singular antecedents in certain situations as a form of respect.

Additionally, a singular pronoun may be used to refer to a plural inanimate antecedent. Resolving these mismatches requires handling the linguistic nuances and variations in the Persian language. Furthermore, Persian poses linguistic limitations compared to English, which adds to the complexity of developing pronoun resolution systems. The differences in grammar, syntax, and linguistic structure require tailored approaches and models specific to Persian. Another significant challenge in developing Persian systems is the weakness and slowness of pre-processing tools. Efficient and accurate pre-processing tools are crucial for obtaining reliable linguistic features and annotations. Overcoming the limitations of these tools and improving their performance can enhance the overall quality of Persian pronoun resolution systems.

Indeed, future works can focus on expanding the Mehr corpus to address the mentioned challenges and include additional linguistic phenomena such as pleonastic pronouns, gender and number mismatches, and pro-drop pronouns. By enriching the corpus with these specific cases, it can serve as a valuable resource for training and evaluating more advanced pronoun resolution systems in Persian. Additionally, developing the first Persian deep learning zero anaphora system would be a significant advancement. Pro-drop pronouns, where the pronoun is omitted and implied from the context, pose a particular challenge in pronoun resolution. Leveraging deep learning techniques can potentially capture the intricate patterns and dependencies required for accurately resolving pro-drop pronouns in Persian.

References

[1] J. Antunes, R. D. Lins, R. Lima, H. Oliveira, M. Riss, and S. J. L. Simske, "Automatic cohesive summarization with pronominal anaphora resolution," *Computer Speech & Language*, vol. 52, pp. 141-164, 2018.

[2] V. K. P. Artari, R. Mahendra, M. A. Jiwanggi, A. Anggraito, and I. Budi, "A Multi-Pass Sieve Coreference Resolution for Indonesian," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021, pp. 79-85.

[3] S. Martschat and M. Strube, "Latent structures for coreference resolution," *Transactions of the Association for Computational Linguistics"*, vol. 3, pp. 405-418, 2015.

[4] H. Chai and M. Strube, "Incorporating Centering Theory into Neural Coreference Resolution," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 2996-3002.

[5] T. M. Lai, T. Bui, and D. S. Kim, "End-to-end neural coreference resolution revisited: A simple yet effective baseline," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8147-8151: IEEE.

[6] L. Miculicich and J. Henderson, "Graph Refinement for Coreference Resolution," *arXiv preprint arXiv*:.16574, 2022.

[7] V. Žitkus, R. Butkienė, and R. Butleris, "Linguistically aware evaluation of coreference resolution from the perspective of higher-level applications," *Natural Language Engineering*, pp. 1-30, 2023.

[8] B. Bohnet, C. Alberti, and M. Collins, "Coreference Resolution through a seq2seq Transition-Based System," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 212-226, 2023.

[9] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," *arXiv preprint arXiv*:.05365, 2017.

[10] M. Klemen and S. Žitnik, "Neural coreference resolution for Slovene language," *Computer Science*

and Information Systems, vol. 19, no. 2, pp. 495-521, 2022.

[11] Ş. Demir, "Neural Coreference Resolution for Turkish," *Journal of Intelligent Systems: Theory and Applications*, vol. 6, no. 1, pp. 85-95, 2023.

[12] R. Grishman and B. M. Sundheim, "Message understanding conference-6: A brief history," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[13] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel, and R. M. Weischedel, "The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation," in *Lrec*, 2004, vol. 2, p. 1: Lisbon.

[14] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes," in *Joint Conference on EMNLP and CoNLL-Shared Task*, 2012, pp. 1-40: Association for Computational Linguistics.

[15] L. Hirschman, "MUC-7 coreference task definition, version 3.0," *Proceedings of MUC-7,*, 1997.

[16] S. Levy, K. Lazar, and G. Stanovsky, "Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation," *arXiv preprint arXiv:.03858*, 2021.

[17] M. Asgari-Bidhendi, B. Janfada, O. Roshani Talab, and B. Minaei-Bidgoli, "Parsner-social: A corpus for named entity recognition in persian social media texts," *Journal of AI and Data Mining*, vol. 9, no. 2, pp. 181-192, 2021.

[18] N. S. Moosavi and G. Ghassem-Sani, "A ranking approach to Persian pronoun resolution," *Advances in Computational Linguistics"*. Research in Computing Science, vol. 41, pp. 169-180, 2009.

[19] M. Bijankhan, "The role of the corpus in writing a grammar: An introduction to a software," *Iranian Journal of Linguistics*," vol. 19, no. 2, pp. 48-67, 2004.

[20] M. Nazaridoust, B. M. Bidgoli, and S. Nazaridoust, "Co-reference Resolution in Farsi Corpora," in *In Advance Trends in Soft Computing: Proceedings of WCSC 2013, Cham*, 2013, pp. 155-162: Springer International Publishing.

[21] A. Mirzaei and P. Safari, "Persian Discourse Treebank and coreference corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[22] Z. Rahimi and S. HosseinNejad, "Corpus based coreference resolution for Farsi text " (in eng), *Signal and Data Processing Research* vol. 17, no. 1, pp. 79-98, 2020.

[23] S. S. Pradhan, E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "Ontonotes: A unified relational semantic representation," in *International Conference on Semantic Computing (ICSC 2007)*, 2007, pp. 517-526: IEEE.

[24] A. Cybulska and P. Vossen, "Guidelines for ECB+ annotation of events and their coreference," in Technical Report: Technical Report NWR-2014-1, VU University Amsterdam, 2014.

[25] Y.-H. Chen and J. D. Choi, "Character identification on multiparty conversation: Identifying mentions of characters in tv shows," in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, pp. 90-100.

[26] A. Ghaddar and P. Langlais, "Wikicoref: An english coreference-annotated corpus of wikipedia articles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 136-142.

[27] A. Zeldes, "The GUM corpus: Creating multilayer resources in the classroom," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 581-612, 2017.

[28] A. Emami, P. Trichelair, A. Trischler, K. Suleman, H. Schulz, and J. C. K. Cheung, "The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution," *arXiv preprint arXiv:.01747*, 2018.

[29] H. Chen, Z. Fan, H. Lu, A. L. Yuille, and S. Rong, "PreCo: A large-scale dataset in preschool vocabulary for coreference resolution," *arXiv preprint arXiv:.09807*, 2018.

[30] D. Bamman, O. Lewke, and A. J. a. p. a. Mansoor, "An annotated dataset of coreference in English literature," *arXiv preprint arXiv:.01140*, 2019.

[31] S. M. Yimam, I. Gurevych, R. E. de Castilho, and C. 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*, 2013, pp. 1-6.

[32] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky, "Deterministic coreference resolution based on entity-centric, precision-ranked rules," *Computational Linguistics*, vol. 39, no. 4, pp. 885-916, 2013.

[33] M. S. Rasooli, M. Kouhestani, and A. Moloodi, "Development of a Persian syntactic dependency treebank," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 306-314.

مهر : پیکره هم مرجع در زبان فارسی

حسن حاجی محمدی[1]، علیرضا طالب پور[2*]، احمد محمودی ازناوه[3] و سمانه یزدانی[4]

[1،4] دانشکده مهندسی کامپیوتر، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران.

[2و3] دانشکده مهندسی کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران.

ارسال ۲۰۲۳/۰۲/۱۲؛ بازنگری ۲۰۲۳/۰۶/۱۶؛ پذیرش ۲۰۲۳/۰۷/۱۸

چکیده:

شناسایی عبارات هم مرجع، یکی از وظایف اساسی در حوزه پردازش زبان طبیعی است. هدف این وظیفه، شناسایی تمام عبارات موجود در یک متن است که به یک موجودیت واقعی در جهان اشاره دارند. شناسایی عبارات هم مرجع در زمینه‌های مختلف پردازش زبان طبیعی، از جمله استخراج اطلاعات و ترجمه ماشینی استفاده می‌شود. در این مقاله، یک پیکره جدید به نام "مجموعه داده مهر" برای عمل شناسایی عبارات هم مرجع در زبان فارسی معرفی می‌شود. هدف اصلی این مقاله، پیکره فارسی برای این عمل است به طوری که ضعف‌های مجموعه‌های قبلی را پوشش داده و هم‌چنین اطمینان از توافق بالا بین برچسب‌گذارها را فراهم آورد. این پیکره، ارتباطات هم مرجع را برای گروه‌های اسمی، موجودیت‌های اسمی، ضمیرها و گروه‌های اسمی تودرتو برچسب‌گذاری کرده است. در این مقاله، دو سیستم پایه برای شناسایی مرجع ضمایر توسعه داده شده و نتایج حاصل از آن‌ها گزارش شده است. مجموعه داده مهر شامل ۴۰۰ سند و حدود ۱۷۰٬۰۰۰ توکن است. برچسب‌گذاری مجموعه داده با استفاده از ابزار پیش‌پردازش WebAnno انجام شده است. نتایج آزمایش‌ها نشان می‌دهند که مدل جنگل تصادفی نامیده-جفت بر روی پیکره مهر عملکرد بهتری نسبت به مدل پایه مبتنی بر روش غربال دارد. کارایی این مدل ۵۹.۴٪ (بر حسب معیار F1) گزارش شده است.

کلمات کلیدی: شناسایی عبارات هم مرجع، جنگل تصادفی، پیکره هم مرجع فارسی، حاشیه زنی پیکره، نامیده، مرجع