Journal of Information Systems and Telecommunication
Vol.14, No.4, July-September 2026, 245-267
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

A Hybrid Entity-Centric Approach to Persian Pronoun Resolution with BERT Enhancement

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

[1]. Department of Computer Engineering, NT.C., Islamic Azad University, Tehran, Iran
[2]. Department of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran
[3]. Cyberspace Research Institute, Shahid Beheshti University, Tehran, Iran

Received: 31 Jan 2025/ Revised: 04 Sep 2026/ Accepted: 25 Oct 2026

Abstract

Pronoun resolution, a critical aspect of coreference resolution in natural language processing, is essential for accurate text understanding and generation. This paper introduces a hybrid Persian pronoun resolution model that leverages rule-based and machine-learning approaches. Seven high-precision rule-based modules, tailored explicitly for Persian linguistic nuances, are combined with a random forest classifier enhanced with BERT contextual embeddings and cluster-level features. This approach effectively addresses the challenges of pronoun resolution, surpassing the performance of existing Persian systems on two corpora: the newly developed Mehr corpus (400 documents) and the established RCDAT corpus. Integrating BERT embeddings and cluster-level features significantly contributes to the model's superior performance, achieving a 4.68 F1-score improvement on Mehr and a 4.01 F1-score improvement on RCDAT. The proposed hybrid model demonstrates a practical and interpretable solution, contributing to Persian natural language processing research and applications.

Keywords: Coreference Resolution; Natural Language Processing; Pronoun Resolution; Corpus Development; BERT; Transformer Models.

1- Introduction

Coreference resolution, a fundamental task within natural language processing (NLP), identifies and groups mentions within a text that refers to the same real-world entity [1, 2]. Accurate coreference resolution is essential for various NLP applications, including information extraction, text summarization, machine translation, question answering, and sentiment analysis [3-7]. The coreference resolution task encompasses two distinct subtasks: mention detection, which involves extracting all relevant mentions from a given text, and mention clustering, which groups these mentions into coreference clusters based on their referential relationships [8].

Recent advancements in coreference resolution have been driven by applying deep neural network systems, particularly transformer architectures such as BERT [9]. These models leverage high-dimensional, multi-scale representations to capture nuanced linguistic patterns and dependencies, significantly improving the accuracy of coreference resolution tasks [10-12]. Furthermore, the versatility of transformer architectures has facilitated the development of end-to-end models, such as seq2seq architectures, which handle both mention detection and clustering jointly [8, 13]. This unified approach streamlines the coreference resolution process and has shown promising results in various domains, including the geological domain [14] and multilingual settings [15].

Pronoun resolution, a particularly challenging subtask within coreference resolution [16], is pivotal for achieving coherent text understanding and generation, with significant implications for tasks such as machine translation. However, this task becomes even more complex for under-resourced languages like Persian, where the scarcity of specialized preprocessing tools and annotated corpora hinders the development of effective models. While most coreference and pronoun resolution systems have primarily focused on English and Chinese, researchers working with less-studied languages, such as Persian or Turkish [17], encounter unique challenges due to the lack of tailored linguistic resources. This paper explicitly targets the Persian language, recognizing that models performing well in other languages may not

✉ **Alireza Talebpour**
Talebpour@sbu.ac.ir

generalize effectively to Persian due to structural differences.

Coreference resolution systems are broadly classified into rule-based and machine-learning paradigms [18]. Early rule-based systems relied on hand-crafted rules for mention clustering [19-24]. Among these, the module-based approach, which employs a series of rule-based modules arranged in descending order of accuracy, has been particularly successful [25-27]. However, while rule-based systems demonstrate simplicity and efficacy in extracting entity-level features [28], they often fall short in pronoun resolution, necessitating sophisticated inference from a combination of linguistic and contextual cues. This complex inferential task is often better suited for machine learning models, particularly those leveraging large-scale pre-trained language models for enhanced contextual understanding.

Machine learning systems, leveraging rich semantic features derived from training data, offer a promising avenue for addressing the complexities of pronoun resolution. While early approaches such as mention-pair and mention-ranking models have demonstrated efficacy, their local focus often needs to improve their ability to make accurate mention merge decisions, especially in long-range dependencies [1]. This limitation has spurred the development of entity-based models, which operate at a cluster level and incorporate broader contextual information. However, entity-based models can encounter computational bottlenecks when processing large documents with extensive coreference chains.

End-to-end models, which jointly handle all stages of coreference resolution, have recently emerged as powerful contenders, demonstrating remarkable performance in specific settings [29, 30]. Nevertheless, these models grapple with challenges such as scaling to large documents, requiring substantial training data, and addressing biases in their predictions [28]. Moreover, despite notable advancements, pronoun resolution remains formidable due to the intricate interplay of contextual information and external knowledge [31]. The effective extraction and integration of external knowledge solely from training data pose a persistent challenge [32].

Recent research has further illuminated the complexities of pronoun resolution. [33] highlight the importance of anisotropic span embeddings in capturing the varied significance of different mention spans while emphasizing the potential negative impact of higher-order inference on model performance. Additionally, the need for fine-tuning coreference models for diverse linguistic styles, such as those found in clinical narratives, has emerged as a critical area of investigation.

The availability of high-quality annotated corpora is essential for developing effective coreference resolution systems [34]. This study leverages two corpora: the established RCDAT [35] and the newly developed Mehr corpus [36]. We propose a hybrid approach to address pronoun resolution challenges, integrating rule-based modules with a machine-learning module that utilizes hand-extracted features, Glove, and Persian BERT embeddings. The entity-centric design, incorporating cluster-level features, significantly enhances system efficiency.

Our hybrid model, incorporating random forest and MLP neural network classifiers, has demonstrated promising results in Persian pronoun resolution. The random forest classifier has emerged as a particularly effective component due to its capacity to model complex interactions among diverse linguistic features and its inherent interpretability [37]. This effectiveness is further amplified by integrating BERT embeddings into the feature set, allowing the model to leverage rich contextual information and enhance its predictive power. The system's architecture is modular and simple, with all components implemented as sequential processing pipelines. This facilitates a streamlined resolution process, efficiently identifying suitable antecedents for each anaphoric pronoun.

The findings of this study have implications for pronoun resolution, particularly for under-resourced languages like Persian. By addressing the unique challenges faced in this domain and introducing a novel hybrid model enriched with BERT embeddings, our research provides insights and practical solutions that can be leveraged to enhance the accuracy and efficiency of pronoun resolution systems across various languages. Developing the Mehr corpus, a comprehensive Persian coreference resource further contributes to advancing Persian NLP research. The evaluation of our proposed system against existing models on both the Mehr and Uppsala test sets offers evidence of its performance, establishing a new benchmark for Persian pronoun resolution.

The remainder of this article is structured as follows: Section 2 analyzes existing coreference resolution systems, encompassing both English and Persian approaches, focusing on hybrid and transformer-based models. Section 3 delves into the landscape of Persian coreference corpora, offering a critical review and highlighting the unique contributions of our newly developed Mehr corpus. Section 4 details our proposed hybrid model for Persian pronoun resolution, elucidating its architecture, the integration of Persian BERT embeddings, and the role of cluster-level features. Section 5 presents an evaluation of the model, outlining the experimental setup, datasets, and metrics employed and discussing the results obtained on both the RCDAT and Mehr corpora. Finally, Sections 6 and 7 make a discussion and conclude the paper by summarizing the key findings, discussing the implications of our work, and proposing avenues for future research.

2- Related Work

This section provides a comprehensive overview of existing coreference resolution systems, encompassing seminal works in English and recent advancements tailored to the Persian language. We will examine the evolution of methodologies, from early rule-based approaches to cutting-edge deep learning techniques, highlighting key contributions and persistent challenges in the field. Additionally, we will delve into the unique linguistic and resource-related obstacles faced in Persian coreference resolution, setting the stage for our proposed hybrid approach.

2-1- Coreference Resolution Systems in English

A seminal work in rule-based coreference resolution is the system proposed by [27] , which utilizes a series of modules organized in a precision-ranked cascade. This approach prioritizes high-precision rules in earlier modules, gradually transitioning to more recall-oriented rules in later modules. This system's influence is evident in its adoption within Stanford University's preprocessing tools and its adaptation for coreference resolution in various languages, including Indonesian [38].

Early machine-learning approaches to coreference resolution primarily utilized the mention-pair model, which employs a binary classifier to assess the coreference relationship between each mention and potential antecedent [2, 39-42]. While simple and widely adopted, this approach is inherently limited by its local focus, considering only pairwise relationships between mentions. [43, 44] offers a broader perspective by evaluating multiple candidates simultaneously. This model allows for a non-anaphoric mention, i.e., not referring to any previous entity. The mention-ranking architecture has been successfully integrated into numerous deep-learning systems [45, 46] demonstrating its effectiveness in capturing more complex coreference patterns.

To address the limitations of local focus in early mention-pair models, entity-based models [47-49] emerged, leveraging cluster-level features for broader contextual analysis. This led to the development of the cluster-ranking model, which combines entity-based and ranking approaches to score potential antecedents at the cluster level. Simultaneously, hybrid systems have been proposed to exploit the strengths of both rule-based and machine-learning paradigms [50]. These hybrid systems often employ dedicated machine-learning modules for different entity types [50] or use dependency trees for enhanced linguistic analysis [28].

A recent trend in coreference resolution is the adoption of end-to-end neural models, which jointly handle mention detection and clustering, eliminating the need for separate preprocessing and feature engineering steps [46]. These models often employ advanced neural architectures such as bidirectional LSTMs [46] or transformers Lai et al. [51] to capture contextual information and model complex linguistic relationships. For example, [46] introduced an end-to-end mention-ranking model using a bidirectional LSTM to compute span embeddings, while [52] further enhanced this approach by incorporating a biaffine attention mechanism to model pairwise mention interactions.

Building upon the success of end-to-end neural models, recent research has explored novel architectures and training strategies to improve performance further. The integration of pre-trained language models like BERT has significantly boosted the ability of these models to capture contextual information, leading to substantial gains in accuracy [30, 53]. Additionally, approaches like entity equalization [54] and graph-based modeling [55] have been proposed to enhance global coherence and model complex coreference structures. Seq2seq models, in particular, have shown promise in coreference resolution due to their ability to generate coreference [8, 56] directly.

Despite these advances, end-to-end coreference models still face several challenges. One major issue is scaling to large documents with extensive coreference chains, often requiring substantial computational resources and training data. Additionally, these models can need help with higher-order inference and the effective integration of external knowledge [33]. To mitigate these issues, researchers have explored techniques like antecedent pruning [29], dual caching mechanisms [57] for efficient processing of long documents, and innovative architectures that balance the need for both local and global information [33]. Table 1 summarizes the key aspects of the mentioned related works in English coreference resolution.

2-2- Coreference Resolution Systems in Persian

In the context of the Persian language, [58] introduced the first pronoun resolution system designed explicitly for Persian. Their system employed the ranking method to determine the antecedent of pronouns in the PCAC-2008 corpus, which the authors themselves developed. The study [59] elaborated on their system's methodological details and features. The features utilized in their system encompassed pronoun features, antecedent features, and relational features. To evaluate the effectiveness of their approach, the authors compared their features with four models: max entropy, max classifier, perceptron, and c4.5. The results section indicated that the c4.5 classifier achieved the highest accuracy, with an F1 score of 44.7.

Table 1: Summary of Key Contributions in Related Works on Coreference Resolution

Article	Approach/Model	Key Contributions
[27]	Rule-Based	Successful rule-based system using modules with decreasing precision and increasing recall.
[50]	Hybrid (ML + Rule-based)	Hybrid system with dedicated machine learning modules for each entity type; explored the use of dependency trees.
[46]	End-to-End (Mention-Ranking)	Bidirectional LSTM for span representation; joint mention detection and clustering.
[52]	End-to-End (Span-Ranking)	Improved span representation with biaffine attention; joint optimization of mention-ranking and coreference-ranking objectives.
[60]	End-to-End (Higher-Order)	Incorporates higher-order inference for improved accuracy; employs antecedent pruning for efficiency.
[51]	End-to-End (Transformer-based)	Integration of pre-trained transformer models (e.g., BERT) to leverage contextualized language representations.
[61]	Neural Network + Discourse Theory	Combined neural networks with centering theory for a discourse-aware approach.
[54]	Entity Equalization	Introduced entity equalization to consider each mention in the context of its entire coreference chain.
[30, 53]	Contextual Text Embeddings	Utilized contextual text embeddings (e.g., BERT, SpanBERT) to enhance the representation of mentions and improve accuracy.
[62]	Graph Model (GT2G)	Proposed a graph-based model for encoding coreference links, generating complete coreference graphs directly.
[63]	BERT-based (Actor-Critic Method)	Achieved state-of-the-art performance on CoNLL-2012 with an actor-critic reinforcement learning framework for joint mention detection and clustering.
[33]	Anisotropic Span Embeddings	Highlights the importance of anisotropic span embeddings and analyzes the potential negative impact of higher-order inference.
[57]	Dual Cache Mechanism	Introduces a dual cache mechanism to improve efficiency in handling long documents with extensive coreference chains.

In their study, [64] introduced a novel pronoun resolution corpus tailored explicitly for the Persian language. Building upon the previous work [58], the authors enhanced the system's efficiency by incorporating a new feature set into their newly developed corpus, and including these additional features led to a remarkable 20% increase in performance. Furthermore, [65] improved the results by applying a series of heuristics to solve the problem of an unbalanced dataset generated in the system [58] and adding semantic information. To evaluate the effectiveness of their approach, the authors examined various classifiers, including maximum entropy, perceptron, SVM, random forest, KNN, and C4.5. The best-reported results were the decision tree, about 75 F1 on the PCAC-2008 corpus.

[66] proposed a rule-based system explicitly designed for Persian pronoun resolution. The system consists of two main components: pre-processing and antecedent selection. By employing five simple rules, the system determines the antecedent of pronouns, considering up to three sentences before the pronoun occurrence. The researchers conducted their experiments using a corpus of five Persian blogs containing 20 pages. A comparative analysis between their system and the one presented in [58] revealed that it achieved higher accuracy and recall rates, indicating improved pronoun resolution performance.

[67] introduced the first coreference resolution system for Persian. The authors developed the Lotus corpus and used 17 features to construct their coreference system. The system was tested on 20% of the Lotus corpus in the evaluation phase using various classifiers, including a decision tree, support vector machine (SVM), and neural network (NN). Among the classifiers, the neural network approach demonstrated the best performance, achieving an accuracy of 39.4 F1 metric.

The previous systems developed for coreference resolution in Persian were primarily based on statistical approaches and utilized limited corpora for their experiments. However, these systems could not form output chains and did not employ widely recognized evaluation metrics such as CoNLL for performance assessment. The research direction in Persian coreference resolution took a significant turn with the development of the RCDAT corpus [35] , which consists of one million tokens. The corpus developers reported 60 conll points by developing the mention-pair model using the SVM algorithm on the RCDAT corpus.

Another notable system that evaluated its results on the RCDAT corpus is the system introduced in [39]. This article employed a fully connected neural network as the coreference classifier. A graph-based algorithm was utilized to construct coreference chains, leading to a significant improvement of 2.2 CoNLL points compared to the previous system. This article employed a fully connected neural network as the coreference classifier. A graph-based algorithm was utilized to construct coreference chains, significantly improving 2.2 CoNLL points compared to the previous system. In their subsequent paper [40], the authors further enhanced the results by maintaining the general framework of the model while modifying the feature vector and incorporating a semantic feature introduced in [68]. The feature set in this article consisted of three components: heuristic features, word vectors, and semantic features. Coreference chains were formed using a hierarchical clustering algorithm to generate the output. Notably, they reported a remarkable result of 64.54 CoNLL points on the RCDAT corpus, showcasing the effectiveness of their approach and the advancements made in Persian coreference resolution.

3- Coreference Corpora in Persian

The availability of high-quality annotated corpora is a prerequisite for developing and evaluating effective coreference resolution systems. In this study, we leverage two distinct Persian corpora: the established RCDAT corpus and the newly developed Mehr corpus. The RCDAT corpus comprises 1,599 documents and is a valuable resource for evaluating coreference systems in Persian. However, it exhibits limitations due to the absence of annotations for specific linguistic phenomena, such as non-referential pronouns and nested named entities, which are crucial for comprehensive pronoun resolution.

We have developed the Mehr corpus to address these limitations and provide a more comprehensive resource for Persian coreference research. In the following parts of this section, we comprehensively examine the available corpora in Persian. Furthermore, we provide detailed insights into the development process and specifics of the Mehr corpus, shedding light on its construction methodology and characteristics.

3-1- Early Corpora in Persian

The PCAC-2008 corpus [58], a pioneering effort in Persian pronoun resolution, consists of 31 texts extracted from BijanKhan's corpus [69]. However, it is limited in scope, focusing solely on the closest antecedent annotation for a subset of pronouns (2006) and needing more diversity in text genres and topics, potentially hindering its generalizability. In contrast, the Lotus corpus [67] represents the first comprehensive Persian coreference resolution corpus. However, its annotation is limited to coreferent noun phrases within 50 lengthy texts from BijanKhan's corpus, raising concerns about its ability to capture the full spectrum of coreference phenomena in Persian.

3-2- PerCoref Corpus

The PerCoref corpus [70] , comprising 212,646 tokens across 547 documents, represents a notable effort to compile a diverse dataset for Persian coreference resolution. However, its construction methodology, which involves random collection of documents from the web, raises concerns regarding the consistency and potential biases in its annotation. This methodological limitation may hinder the generalizability of findings derived from experiments conducted on this corpus. Nevertheless, PerCoref encompasses a wide range of coreference phenomena, including three reference types (direct, indirect, and none) and six coreference relations (Identity, person/number suffix on the verb, Inferred, Event, Quantifier, and Cross-speech), offering a potentially valuable resource for exploring the complexities of Persian coreference. Table 2 provides detailed statistics regarding the composition of this corpus, and Table 3 presents the statistical distribution of these coreference types within the corpus.

Table 2: PerCoref statistics

Number of documents	547
Number of tokens	212646
Number of sentences	6688
Number of paragraphs	4449
Average paragraph per document	8.13
Average sentence per document	1.5
Average token per document	31.79

Table 3: PerCoref corpus coreference relations

Coreference relations	***Count***	
Identity	11812	58.4
person/number suffix on verb	3021	14.9
Inferred	2005	9.9
Event	1509	7.4
Quantifier	968	4.7

Coreference relations	*Count*	
Cross-speech	901	4.4

3-3- RCDAT Corpus

The RCDAT corpus [35] , comprising one million words annotated in the CoNLL format, is a substantial resource for Persian coreference resolution. It includes gold standard annotations for coreference chains, named entities, and animacy and covers a various political, sports, and cultural texts from various news sources. However, its annotation scheme lacks coverage of important linguistic phenomena pertinent to pronoun resolution, such as non-referential pronouns and nested named entities. This limitation constrains its utility for comprehensive pronoun resolution tasks, particularly those that require a deeper understanding of referential ambiguities and complex syntactic structures. Table 4 provides detailed statistical information about the RCDAT corpus.

Table 4: RCDAT's statistical information

	count
Number of documents	1599
Number of tokens	1052637
Number of labeled tokens	408138
Number of entities	86960
Number of pronouns	2687
Number of proper nouns	3412
Number of animate entities	4266

Each line of the RCDAT corpus provides detailed information about individual tokens, including the document name, sentence number, token itself, part-of-speech tags (16 tags), named entity tags (both 13 and 3 granularities), the token's root form, original token form, coreference chain index (gold standard), coreference chain number (gold standard), animacy, phrase type, and a more fine-grained set of part-of-speech tags (100 tags).

3-4- Mehr Corpus

To address the limitations of existing Persian corpora, the authors developed the Mehr corpus [36], a comprehensive resource designed explicitly for Persian coreference resolution. This corpus comprises 400 texts randomly extracted from the Mehr news agency, ensuring diverse topics and linguistic styles. Annotation was meticulously performed by two expert annotators using WebAnno [71] , a collaborative annotation tool that facilitates high-quality and consistent annotations. WebAnno also automatically generates the CoNLL file for the labeled chains, streamlining the data preparation process for subsequent model training and evaluation. The Mehr corpus is divided into distinct training, development, and testing sets, following established best practices for model development and evaluation. The test set, comprising approximately 13,000 tokens, is specifically designed to align with the size of the MUC7 corpus test set [72], enabling direct comparison and standardized evaluation against other coreference resolution systems.
Fig1 visually represents the annotation process within the WebAnno environment, illustrating the user interface and annotation workflow. Table 5 presents a comprehensive summary of the statistical characteristics of the Mehr corpus, including the number of tokens, sentences, and documents in each split. This detailed information allows a thorough understanding of the corpus composition and its suitability for various coreference resolution tasks.

Table 5: Statistical characteristics of Mehr's corpus

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

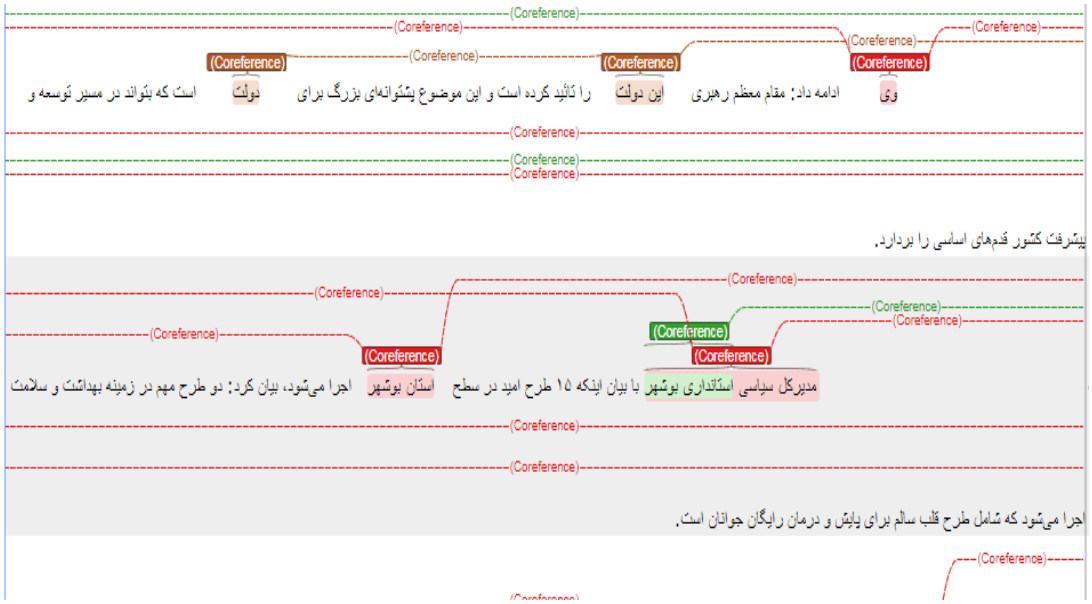


Fig1. A frame of WebAnno annotation tool

The Mehr corpus employs a dual-labeling strategy, distinguishing between "gold labels" and "system labels".

Gold labels:

In the Mehr corpus, the coreference labels for pronouns, noun phrases, named entities, and nested named entities are assigned through manual annotation. Only the mentions of at least one coreference chain are labeled in the corpus. It is worth noting that some non-referential pronouns are labeled among the pronouns. In Persian, some pronouns do not have a direct reference in the text. In fact, in Persian, some pronouns do not refer to anything. Non-referential pronouns are also labeled in the Mehr corpus. Some pronouns, such as nobody and others, which do not have any reference in the text, are not labeled.

In Persian, some pronouns have an indirect reference. These pronouns refer to the speaker of a quote or an utterance, which may have been omitted in the document. In the Mehr corpus, these types of pronouns with indirect reference are also labeled and included in the coreference annotations. In the Mehr corpus, the focus is on labeling pronominal singletons with coreference relations. Specifically, only pronominal mentions with no explicit antecedents within the corpus are labeled. The coreference relation type assigned to these pronominal singletons is the identity relation, indicating that they refer to the same entity. The following is an example of each coreference relation in the Mehr corpus.

- **Coreference Relations**

(محمد) به مدرسه آمد. بچه ها از دیدن (او) تعجب کردند.

(Muhammad) came to school. The children were surprised to see (him).

(ابن سینا) دانشمند ایرانی است. (این دانشمند برجسته) در همدان مدفون است.

IbnSina is an Iranian scientist. (This outstanding scientist) is buried in Hamadan.

- **Indirect Reference**

(ما) باید به تیم ملی خود افتخار کنیم.

(We) should be proud of our national team.

- **Non-Referential Pronouns**

(این)، مساله مهمی است.

(It) is an important issue.

System labels:

Text processing tools in Persian automatically generate labels for various linguistic elements. The accuracy of these automatically generated labels may vary compared to manual labels and relies on the precision of the preprocessing tools employed. The Mehr corpus's automatic labeling encompasses elements such as named entities, noun phrases, POS tags, and syntactic groups. This corpus uses the Hazm preprocessing tool [1], the comprehensive platform of intelligent content analysis [2], and the FarsiYar text preprocessing tool [3] .Table 6 shows the accuracy of preprocessing tools used in this corpus.

Including gold and system labels in the Mehr corpus aims to provide a comprehensive and multifaceted resource for advancing research in Persian coreference resolution. The manual annotations ensure high-quality training and evaluation data, while the automatic labels offer complementary information for further linguistic exploration.

[1] https://www.sobhe.ir/hazm/
[2] http://catotalservices.com/UIHome/Index.aspx
[3] https://text-mining.ir/

Table 6: Accuracy of preprocessing tools used in this paper

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

4- Proposed Method

In light of previous research [40] established a benchmark for Persian coreference resolution by demonstrating the superior performance of their system, which implemented state-of-the-art English models on the Persian RCDAT corpus. Building upon this foundation, the present work aims to advance the state of the art by introducing a novel hybrid model that synergistically integrates rule-based and machine-learning components. This innovative approach seeks to leverage the strengths of both paradigms, overcoming the limitations of purely rule-based or machine-learning systems and ultimately achieving superior performance in Persian coreference resolution.

This section establishes a foundation for our investigation by introducing two baseline models for Persian pronoun resolution. The first is a rule-based module-based system inspired by the work of [27], adapted and tailored for the nuances of Persian grammar and syntax. The second baseline is a mention-pair machine learning model that leverages hand-crafted features to capture pronoun, antecedent, and relational information. To further advance the state of the art, we propose a novel hybrid entity-centric model that integrates these two baselines and incorporates additional cluster-level features, along with GLOVE and Persian BERT embeddings, to enrich contextual representation. Our empirical evaluation demonstrates that this hybrid approach significantly outperforms existing Persian systems, highlighting the potential of combining rule-based and machine-learning paradigms for this challenging task.

4-1- Persian Module-Based Baseline Model

Inspired by the widely adopted architecture of [27], our Persian module-based baseline system leverages a precision-ranked cascade of rule-based modules to resolve coreference chains incrementally. Each module operates by merging a mention into an existing partial cluster or deferring antecedent identification to subsequent modules. This approach facilitates incorporating local (mention-level) and global (cluster-level) information in the resolution process.

The system consists of eight modules specifically adapted for Persian, acknowledging the linguistic nuances and structural differences compared to English. Before module application, input documents undergo preprocessing using existing NLP tools to identify mentioned boundaries and extract relevant features such as number and animacy. These identified mentions are then sequentially processed by the rule-based modules, each contributing to the construction of coreference chains hierarchically.

This modular architecture enables the seamless integration of additional rule-based components and ensures transparency and interpretability in the coreference resolution process. While effective for capturing entity-level features, this system's rule-based nature may pose challenges in handling more complex referential relationships, particularly those involving pronouns that require deeper semantic understanding. However, this baseline is a valuable starting point for our subsequent exploration of hybrid models that integrate machine-learning techniques for enhanced pronoun resolution.

4-1-1- Model Architecture

The architecture of mention selection within our rule-based system closely follows the approach proposed by[26], with a critical adaptation to accommodate the right-to-left reading direction of Persian text. Since partial coreference chains are incrementally formed within each module, selecting representative mentions becomes crucial. Our system prioritizes the examination of the first mention within each partial chain, as this mention typically provides a more explicit entity definition and is less likely to be a pronoun or contain morphological ambiguities [27]. This selection strategy also tends to result in a smaller pool of potential antecedents, leading to increased efficiency.

To ensure that the most relevant antecedents are considered, the first mentions within each sentence are sorted from right to left, reflecting the natural reading order of Persian. If no suitable antecedent is found within the current sentence, the algorithm examines previous sentences in the document, again considering mentions from right to left. This systematic approach maximizes the likelihood of identifying the correct antecedent while minimizing computational overhead. The algorithm for the mentioned selection in the current module is outlined as follows.

Selecting Mentions within the Current Module:
Selecting mentions within the current module involves considering the partial chains generated by the previous modules. Only the first mention within each partial chain is selected for examination in the current module. For

instance, let us assume that the partial chains $[m_1^1.m_2^2.m_3^2.m_4^3.m_5^1.m_6^2]$ have been constructed for the current document. The lower index denotes the mention number, while the upper index represents the cluster number. In the current module, the model focuses on finding the antecedent for m_2^2 and m_4^3 (m_1^1 is not checked as it is located at the beginning of the document). This selection is based on the observation that first mentions within partial chains provide better entity definitions and have a lower probability of being pronouns or containing prefixes and suffixes[27]. Additionally, first mentions are typically closer to the beginning of the text, resulting in a smaller pool of potential antecedents to consider.

Antecedent Selection for a Current Mention: Each mention in the current module may be merged with a previous partial chain, or its antecedent may be assigned in the next module. For each mention in the current module, the mentions in the current sentence are sorted from right to left and checked in order. The document sentences are traversed from left to right if there is no answer in the current sentence for the current mention. Then, in each sentence, the mentions are sorted from right to left to reach the answer. Sorting the list of candidates is important because the algorithm stops at the first confluence. Weak sorting causes the accuracy of the system to decrease. A sentence threshold window should be set in each module so the algorithm does not check more than that threshold limit.

Sharing Features in the Entity-Centric Model: When constructing a partial entity that consists of one or more mentions, the system performs a union operation on the features of those mentions, such as number, gender, and animacy. The resulting feature values are then shared among all the mentions within that entity. In cases where the feature values differ among the mentions of the entity, all the values are preserved. For example, suppose the number detection module identifies the mention "group of students" as singular and "five students" as plural. In that case, the resulting partial entity will have singular and plural values for the number feature. This allows for including singular and plural pronouns in the pronoun resolution module when referring to this partial entity.

4-1-2- Model Modules

Our rule-based system employs eight distinct modules, each designed to address specific coreference scenarios in Persian. These modules operate sequentially, each contributing to the incremental construction of coreference chains. The order of the modules is carefully determined, prioritizing high-precision rules in earlier stages and gradually transitioning to more relaxed rules for improved recall. This hierarchical approach ensures a systematic and efficient resolution process while accounting for the varying complexities of different coreference phenomena. We will now detail the functionality of each module within this framework.

Module 1 - Speaker Module: The initial module in our system focuses on resolving pronouns within quotations, a common linguistic construct in news articles and other written discourse. The speaker is typically explicitly mentioned in conversational texts, facilitating straightforward pronoun resolution. However, the speaker may be implicitly or indirectly referenced in non-conversational texts, such as those found in the Mehr and RCDAT corpora. To address this, Module 1 employs a set of rules designed to identify quoted verbs (e.g., "said," "stated," "announced") and their associated subjects, which often correspond to the speaker of the quotation. Once the speaker is identified, the module resolves pronouns within the quotation to this antecedent, establishing a coreference link based on the discourse context.

Module 2 – Exact String Match: Module 2, the Exact String Match module, is a high-precision component that operates on the principle of string identity. In this module, if two mentions match precisely, including any prefixes or suffixes, they are deemed coreferent and merged into a single chain. This approach is computationally efficient and particularly effective in resolving mentions with identical surface forms, such as repeated proper names or nominal mentions. While exact string matching does not consider deeper linguistic or semantic information, its high precision makes it a valuable initial step in the coreference resolution pipeline, resolving straightforward cases before resorting to more complex analysis in subsequent modules.

Module 3 – Strict Head Match: Building upon the string-based matching of Module 2 and Module 3, the Strict Head Match module applies more stringent criteria for resolving coreferences based on head noun agreement. While mentioning sharing the same head noun as a potential candidate for coreference, this module acknowledges that head noun overlap alone is insufficient to guarantee referential identity. For instance, "Tehran University" and "Washington University" share the head noun "University" but refer to distinct entities. Stricter conditions for merging mentions based on head noun matching are imposed to mitigate such errors. The head noun of the current mention must match the head noun of a previous mention and be a complete substring of that mention. For example, the mentions "Tehran Court" and "Tehran High Court" would be considered coreferent as the former is a substring of the latter. This approach ensures higher precision in coreference resolution by considering lexical and structural cues.

Module 4 – Proper Name Match: Module 4 is dedicated to resolving coreference involving proper names. This module targets mentions that share the same proper noun head, such as "Beckham" and "David Beckham," to merge them into a single coreference chain. To ensure accuracy, certain conditions are imposed to prevent erroneous merges. For instance, the module may consider the relative positions of the mentions within the text or the presence of additional modifiers that could differentiate the referents. This approach helps refine the coreference resolution process by addressing the unique challenges posed by proper name mentions in Persian.

Module 5 – Location Match: Module 5 focuses on the resolution of location-based coreference. It specifically targets mentions tagged as "LOC" (location) by the named entity recognition (NER) component of the pre-processing pipeline. In this module, a substring matching heuristic is applied to determine coreference. If one location mention is a complete substring of another, they are considered to refer to the same geographical entity and are merged into a single chain. For instance, the mentions "Tehran" and "Tehran City" would be linked under this rule. This module aims to capture coreference relationships that arise from the hierarchical nature of location names, where more specific locations are often embedded within broader geographical entities. By identifying and merging such mentions, Module 5 enhances the accuracy of coreference resolution for location-related entities in Persian texts.

Module 6 – Title Match: Module 6 addresses coreference relationships involving titles, appositives, and descriptive phrases associated with person names. In Persian news texts, it is common for individuals to be referred to by their titles or appositives, either preceding or following their names. For example, a sentence might mention "President Hassan Rouhani" followed by "Rouhani" in a later sentence, or vice-versa. This module leverages simple pattern-matching rules to identify and merge such pairs into a single coreference chain. By recognizing the relationship between names and their associated titles or descriptions, Module 6 enhances the accuracy and completeness of coreference resolution for person mentions in Persian texts.

Module 7: Demonstrative Phrase Match: Module 7 addresses coreference resolution involving both demonstrative phrases and pronouns. In Persian, demonstratives often play a crucial role in establishing coreference relationships, as they explicitly point to previously mentioned entities. This module identifies mentions containing demonstrative expressions (e.g., " این نمایشگاه" meaning "this exhibition") and attempts to link them with their corresponding antecedents based on semantic and syntactic cues. In addition to demonstratives, this module handles pronoun resolution for pronouns not resolved by earlier modules. It uses syntactic and semantic features to identify the most likely antecedent for each pronoun within a defined context window. This dual functionality enhances the system's ability to capture a broader range of coreference phenomena in Persian text.

Module 8: Pronoun Resolution: In our proposed module-based system, the final module is dedicated to resolving unresolved pronoun mentions after applying the previous seven modules. This module leverages a feature-matching approach to link each pronoun to one of the existing partial chains, thereby identifying its most likely antecedent. Table 7 compares the features used in the proposed system and the system introduced in reference [27].

Table 7: Features related to the pronoun module in the system [27] and the presented system

Feature name	***Proposed system***	***English module-based system***
Number	1- Static list for pronouns 2- NER labels 3- POS labels	1- Static list for pronouns 2- NER labels 3- POS labels 4- Static dictionary from [73]
Gender	1- Static lexicon 2- Static pronouns	1- Static lexicon from [73]
Person	-	1- Use for only 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 [۷۵]	1- Static list for pronouns 2- NER labels 3- Dictionary extracted from WEB [74]
Distance	Less than 4 sentence	Less than 4 sentence

4-2- Persian Mention-Pair Baseline Model

Our second baseline model employs a mention-pair architecture, a classical coreference resolution approach focusing on pairwise comparisons between potential mention-antecedent pairs. This model utilizes a suite of hand-crafted features meticulously designed to capture the linguistic nuances of the Persian language. These features encompass various aspects of pronouns (e.g., type, grammatical role), their potential antecedents (e.g., entity type, syntactic position), and the relationship between them (e.g., distance, agreement). We introduce a distance window hyper-parameter to enhance the model's generalization capabilities, which restricts the maximum distance between a pronoun and its potential antecedent. This helps mitigate the risk of erroneous long-range coreference links.

During training, we employ a strategy to address class imbalance, a common issue in coreference datasets where negative examples (non-coreferent pairs) significantly outnumber positive examples (coreferent pairs). We create a positive example for each pronoun by pairing it with its nearest true antecedent. To balance the dataset, we generate negative examples by including all intervening mentions between the pronoun and its true antecedent [41].

A comprehensive feature vector encompassing mention-level and cluster-level attributes (Table 8) is constructed for each mention-antecedent pair. The features are mention-level (M) and cluster-level (E). The "+" sign indicates features exclusive to the hybrid model presented in this article, while the mention-pair system incorporates features without the "+" sign.

These features include lexical, syntactic, semantic, and positional information, providing a rich representation for the machine learning classifier. Through rigorous experimentation involving five repetitions of 10-fold cross-validation on the training data, we evaluated the performance of various classifiers, including decision trees, random forests, multi-layer perceptron, and support vector machines. The results presented in Section 5 demonstrate that the random forest classifier consistently outperforms the others for this mention-pair baseline model.

Table 8: The features used in the proposed hybrid models and the mention-pair model

Pronoun features		
row	Description	M/E
1	Is it a personal pronoun?	M
2	Is it a demonstrative pronoun?	M
3	Is it a reflexive pronoun?	M
4	Is it a third-person pronoun?	M
5	Is it a speech pronoun?	M
6	Pos label of the first token on the left side of the pronoun	M
7	Pos label of the second token on the left side of the pronoun	M
8	Pos label of the third token on the left side of the pronoun	M
9	Pos label of the first token on the right side of the pronoun	M
10	Pos label of the second token on the right side of the pronoun	M
11	Pos label of the third token on the right side of the pronoun	M
12	Is pronoun subject?	M
13	Is pronoun object?	M
14	Pronoun number	M
15	Pronoun animacy	M +
16	Pronoun person	M +
Antecedent features		
17	How many tokens does the antecedent have?	M
18	Is the antecedent a pronoun?	M
19	Is the antecedent a demonstrative phrase?	M
20	Pos label of the first token on the left side of the antecedent	M
21	Pos label of the second token on the left side of the antecedent	M
22	Pos label of the third token on the left side of the antecedent	M
23	Pos label of the first token on the right side of the antecedent	M
24	Pos label of the second token on the right side of the antecedent	M
25	Pos label of the third token on the right side of the antecedent	M
26	Antecedent number	M
27	Is Antecedent subject?	M

28	Is Antecedent object?	M
29	How many mentions are there in the antecedent chain?	E+
30	Is the antecedent a reflexive pronoun?	M+
31	Antecedent type (pronoun, proper noun, common noun)	M+
32	The sentence Index of the first mention in the antecedent chain	E+
33	Antecedent animacy	M+
34	Antecedent person	M+
35	Named-entity label	M+
36	Antecedent chain number	E+
37	Antecedent chain animacy	E+
Relational features		
38	Sentence distance	M
39	Token distance	M
40	Number agreement	M
41	Subject agreement	M
42	Object agreement	M
43	String match	M
44	Is the distance between the pronoun and antecedent candidate less than three?	M
45	Are the pronoun and the antecedent candidate in the same sentence?	M
46	What is the minimum sentence distance between the pronoun and the antecedent candidate chain?	E+
47	Is the antecedent string token count more than the pronoun string?	M+
48	Animacy agreement	M+
49	Person agreement	M+
50	If a pronoun is an object and the antecedent is a subject, both are in the same sentence.	M+
51	If a pronoun is reflexive and the antecedent candidate is a subject, both are in the same sentence.	M+
Word embedding features		
52	Euclidean distance of two mention heads	M+
53	Euclidean distance between the mean of two mentions	M+
54	The Euclidean distance between the antecedent candidate and the sentence in which the pronoun is placed.	M+
-	word embedding for pronoun	M+
-	word embedding for antecedent	M+

4-3- The Proposed Hybrid Model

Building upon the rule-based and machine-learning baselines presented in previous sections, we introduce a novel hybrid model for Persian pronoun resolution that combines the strengths of both paradigms. This hybrid system leverages the efficiency and interpretability of rule-based approaches, exemplified by the seven high-precision modules described in Section 4.1 while harnessing the power of machine learning to model complex linguistic relationships and contextual dependencies. The final module in our pipeline, designed explicitly for pronoun resolution, employs a random forest classifier that has been shown to handle high-dimensional feature spaces and provide interpretable results effectively.

The model operates entity-centric, incrementally gathering information about coreference clusters as the text is processed. As depicted in

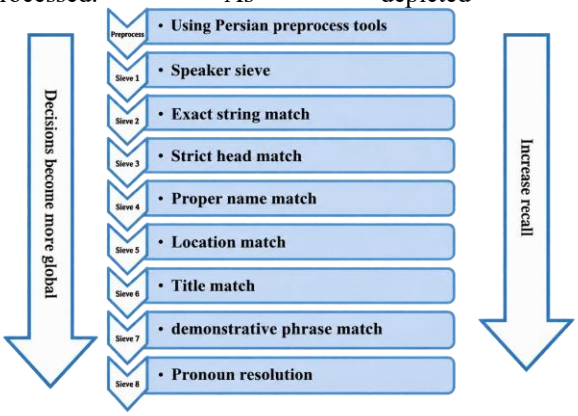


Fig 2 , this incremental approach allows the model to incorporate cluster-level features, such as the number of mentions in a cluster and the animacy of the cluster's referent, into the decision-making process for pronoun resolution. By utilizing both hand-crafted features and contextualized word embeddings derived from Persian BERT and GloVe, the machine-learning module effectively learns to identify the most likely antecedent for each pronoun, considering both local and global contextual information.

Our hybrid model offers several key advantages over existing approaches. By incorporating rule-based modules, we can quickly leverage linguistic knowledge and

heuristics to resolve unambiguous coreferences. This reduces the burden on the machine-learning module, allowing it to focus on more challenging cases where contextual information and complex linguistic patterns are critical. Additionally, using a random forest classifier ensures computational efficiency and interpretability. The subsequent sections will delve deeper into the model's training and testing procedures and a comprehensive evaluation of the Mehr and RCDAT corpora, demonstrating its superior performance compared to previous Persian coreference resolution systems.

Feature Vector: A key component of our hybrid model is carefully engineering features that capture the multifaceted nature of pronoun resolution. Our approach combines hand-crafted features, specifically designed to capture the linguistic nuances of Persian, with rich contextualized representations derived from both GloVe and Persian BERT embeddings. These hand-crafted features encompass a wide range of attributes about pronouns (e.g., type, grammatical role), their potential antecedents (e.g., entity type, syntactic position), and their relationships (e.g., distance, agreement). By integrating GloVe embeddings, we capture semantic and syntactic relationships between words, while Persian BERT embeddings provide a deeper understanding of contextual nuances. This multi-faceted approach to feature engineering enhances the model's ability to disambiguate pronoun references and make informed predictions.

Table 8 provides a comprehensive overview of the 54 features employed in our hybrid model, encompassing mention-level and cluster-level attributes. This diverse set of features, combined with the power of contextual embeddings, enables the random forest classifier to effectively model the complex interactions between pronouns and their potential antecedents, leading to superior performance in Persian pronoun resolution. This comprehensive feature engineering approach, coupled with the hybrid model architecture, contributes to the superior performance of our system on the Persian coreference resolution task.

Training: Our entity-centric approach aims to establish coreference links between pronouns and their potential antecedents by utilizing a machine learning module trained on pre-existing coreference chains. The training process generates positive examples by associating each pronoun with its closest actual antecedent and negative examples by incorporating all intervening mentions within the specified distance window. This strategy aims to mitigate the class imbalance issue commonly encountered in coreference datasets.

Crucially, our model incorporates cluster-level features, such as the number of mentions in a cluster and the animacy of the cluster's referent, to provide a richer contextual representation for pronoun resolution. We evaluate the performance of various classifiers, including random forests, multi-layer perceptron, support vector machines, and gradient boosting trees, on two distinct Persian corpora: the newly developed Mehr corpus and the established RCDAT corpus. The results of this comparative analysis, including the optimal hyperparameters for the best-performing classifier, are presented in Section 5.

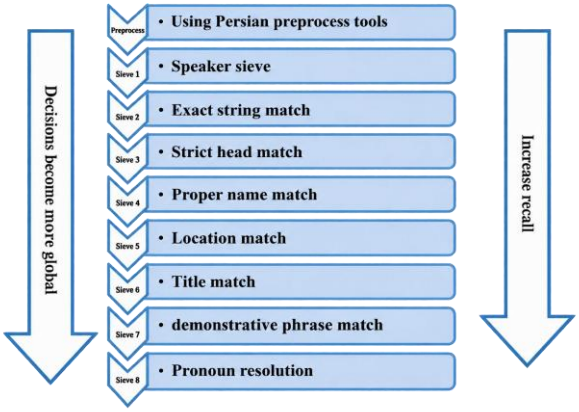


Fig 2. The introduced hybrid architecture.

Test: Our entity-centric hybrid model is evaluated through a systematic testing procedure on unseen data. The process begins with the application of the rule-based modules, generating partial coreference chains based on pre-defined linguistic rules and heuristics. For unresolved pronouns, the machine-learning module, powered by a random forest classifier enhanced with BERT embeddings, assesses potential coreference links with existing partial chains within a predefined sentence window. The random forest classifier assigns scores to each candidate antecedent, selecting the highest-scoring candidate that exceeds a predetermined threshold. If no candidate meets this criterion, the pronoun is classified as non-anaphoric.

5- Experiments

To rigorously evaluate the effectiveness of our proposed hybrid model, we conducted experiments on two distinct Persian corpora: the newly developed Mehr corpus and the established RCDAT corpus. We followed a standard train-test split for the Mehr corpus, utilizing the designated training documents for model development and the corresponding test documents for evaluation. In the case of the RCDAT corpus, due to the absence of a predefined test set, we adopted the methodology employed in previous Persian coreference resolution studies, utilizing the Uppsala test documents [76] for evaluation. This corpus, comprising 415 sentences and 16,274 tokens, offers diverse narratives for assessing the model's performance

on unseen data. It is important to note that our evaluation of the RCDAT corpus focuses solely on pronouns that are part of an existing coreference chain, excluding non-referential pronouns and pronouns initiating new chains, as these are not annotated in the RCDAT corpus. This decision ensures consistency with previous work and directly compares our results with existing systems.

This paper focuses on the specific challenge of resolving detached pronouns in Persian. Detached pronouns in Persian include personal, demonstrative, and reflexive pronouns, each contributing unique complexities to the coreference resolution task. Compared to English, pronoun resolution in Persian poses additional challenges due to the language's distinct linguistic properties.

One prominent challenge stems from Persian's pro-drop or null-subject nature, which allows for omitting nominal subjects in sentences. This phenomenon often obscures the intended referent of a pronoun, making antecedent identification more difficult. Additionally, unlike English, third-person pronouns in Persian lack inherent gender marking, rendering gender an unreliable cue for resolution. Furthermore, Persian's complex system of politeness and formality introduces further ambiguity. Plural pronouns may refer to singular entities out of respect, while singular pronouns may address plural entities in formal contexts. These intricate sociolinguistic nuances further underscore the necessity for specialized models and techniques to navigate Persian pronoun resolution's intricacies effectively.

5-1- Experiments on MEHR Corpus

This section comprehensively evaluates our proposed hybrid model for Persian pronoun resolution on the Mehr corpus. Table 9 summarizes the results of our experiments, directly comparing various baseline systems and previously established Persian coreference resolution approaches. The reported performance metrics reflect the optimal model configurations identified during our experimentation. We begin by establishing baseline results using a rule-based system and a mention-pair model (Lines 1-2), followed by a comparison with the current state-of-the-art Persian system (Lines 3-4), as reported by [40], which notably outperformed contemporary English systems on the Persian RCDAT corpus.

Using a random forest classifier, the proposed hybrid system demonstrates promising results for Persian pronoun resolution across various settings. The model achieves the highest F1 score (66.8) when utilizing gold cluster-based features and BERT embeddings, underscoring the importance of accurate annotations and contextualized language representations for this task. The system maintains competitive performance even with automatically derived cluster features, particularly when combined with either GloVe or BERT embeddings (F1 scores of 64.75 and 65.15, respectively). These results highlight the effectiveness of our hybrid approach and the potential for further improvement by refining automatic feature extraction methods.

Table 9: The reported results of the presented system on Mehr corpus compared to previous works and various settings

	System	***Using Rule-Based Modules***	***GloVe/BERT***	***Cluster-Based Features***	***precision***	***recall***	***F1***
1	Rule-based baseline	+	-	-	36.4	35.6	36
2	Mention-pair baseline (random forest)	-	-	system	60	58.82	59.4
3	[40] (random forest)	-	GloVe	system	55.08	54.54	55.08
4	(MLP) [40]	-	GloVe	system	60.47	59.88	60.47
	Proposed system with different settings (random forest)						
5		+	GloVe	gold	66.56	66.27	66.41
6		+	-	system	64.2	62.94	63.56
7		+	GloVe	system	65.4	64.11	64.75
8		-	-	gold	64.28	63.01	63.64
9		-	GloVe	gold	66.28	64.98	65.62
10		+	BERT	gold	67.2	66.4	66.8
11		+	BERT	system	65.9	64.42	**65.15**

	System	*Using Rule-Based Modules*	*GloVe/BERT*	*Cluster-Based Features*	*precision*	*recall*	*F1*
	Proposed system with different settings (MLP)						
12		+	GloVe	gold	64.2	62.97	63.56
13		+	GloVe	system	60.8	59.64	60.2
14		+	BERT	gold	64.9	63.81	64.35
15		+	BERT	system	61.3	63.04	**62.16**
16	Changing the modules' order of system line 7	+	GloVe	system	64.9	63.9	64.4
17	Using logistic regression model instead of random forest in system line 7	+	GloVe	system	60.42	58.44	59.41

5-1-1- Mehr Corpus Result Analysis

Line 11 of Table 9 demonstrates the results of the proposed system on the Mehr corpus. This line indicates that the proposed system outperforms the previous best Persian system, increasing the 4.68 F1 score. This highlights the effectiveness of implementing a hybrid architecture using cluster-level features and leveraging the simplicity of the module-based model. Below is an analysis of the results presented in Table 9.

Hybrid Model Superiority: The hybrid model consistently outperforms both the rule-based baseline (Line 1) and the mention-pair baseline (Line 2), underscoring the benefits of integrating linguistic knowledge with machine learning. Notably, the highest F1 score of 66.8 is achieved when combining the hybrid architecture with gold cluster features and BERT embeddings (Line 10), showcasing the synergistic effect of leveraging both rule-based insights and contextualized representations for enhanced pronoun resolution accuracy.

Impact of BERT and GloVe Embeddings: Including BERT embeddings consistently improves performance over models relying solely on GloVe or no embeddings. Combining both embeddings further elevates results, suggesting a complementary relationship in capturing different aspects of linguistic information. BERT's ability to model contextual nuances is particularly evident in its superior performance compared to GloVe in configurations with gold cluster features (Lines 10 vs. 5).

Significance of Cluster-Level Features: The consistent improvement observed when using gold cluster-level features (e.g., Lines 5, 8, 9, 10, 12, 14) emphasizes their importance in providing a richer representation of coreference chains, enabling the model to make more informed decisions about pronoun resolution. This highlights the need to develop accurate automatic feature extraction methods to bridge the performance gap between gold and system-generated features.

Rule-Based Modules Enhance Performance: While the impact of rule-based modules is less pronounced when using BERT embeddings, they still contribute to improved performance, particularly in recall. This suggests that these modules effectively capture certain coreference relationships that the machine learning component alone cannot fully address. Notably, removing rule-based modules in configurations using gold features (Lines 8 and 9) decreases efficiency, indicating their value in resolving specific linguistic patterns.

Classifier Performance: The random forest classifier consistently outperforms the MLP classifier across comparable settings, demonstrating its superior ability to model complex interactions among the diverse features utilized in our hybrid system. This aligns with the established strengths of random forests in handling high-dimensional data and providing interpretable models.

Model Robustness and Adaptability: The hybrid model's performance remains relatively stable across different configurations, indicating its robustness to feature extraction and module order variations. Additionally, the slight decrease in performance observed when altering the module order (Line 16) suggests that the proposed architecture is flexible and can be adapted to different pipelines without significantly sacrificing accuracy.

Module Order Independence: Changing the order of the rule-based modules has a minimal impact on performance, indicating the robustness and flexibility of the hybrid architecture.

Superiority of Non-Linear Models: The significant drop in performance when using logistic regression (a linear model) instead of random forest (a non-linear model) highlights the importance of capturing non-linear

relationships in the feature space for accurate pronoun resolution.

Challenges in System-Generated Features: The performance gap between systems using gold and system-generated features underscores the need for further research on improving automatic feature extraction methods in Persian NLP.

Potential for Future Improvement: While the hybrid model with BERT embeddings shows promising results, there remains room for further improvement, particularly in refining automatic feature extraction and exploring alternative neural architectures.

These analyses highlight the strengths and improvements of the proposed system, demonstrating the importance of the hybrid architecture, cluster-level features, rule-based modules, and the choice of machine learning algorithms in achieving efficient pronoun resolution in Persian. Table 10 summarizes the main observations and their corresponding analyses, providing a concise overview of the key findings from the experiment results.

Table 10: Analysis of Experimental Results for Pronoun Resolution in the Mehr Corpus

Observation	***Analysis***
Rule-based baseline (Line 1) exhibits low accuracy.	Rule-based modules alone are insufficient for accurate Persian pronoun resolution, highlighting the need for machine learning integration
Mention-pair baseline (Line 2) shows significant improvement over rule-based baseline.	Random forest effectively models complex feature interactions in pronoun resolution, surpassing rule-based approaches.
Hybrid model with system-based features and BERT (Line 11) achieves the highest F1 score.	This configuration demonstrates the synergistic effect of combining linguistic rules, machine learning, and contextual embeddings (GloVe or BERT).
Hybrid models with gold cluster features (Lines 5, 8, 9, 10, 12, 14) outperform those with system features.	Gold-standard annotations for cluster features yield superior results, highlighting the need for high-quality training data and accurate feature extraction.
Removing rule-based modules (Lines 8 and 9) generally decreases performance.	The hybrid architecture is flexible and robust to variations in module sequencing, maintaining consistent performance.
Random forest consistently outperforms MLP (Lines 7, 11, 13, 15 vs. lines 6, 10, 12, 14).	Random forest's ability to model non-linear relationships in high-dimensional feature spaces makes it more suitable for this task than MLP.
Logistic regression underperforms compared to random forest (Line 17).	The superior performance of random forest emphasizes the importance of capturing non-linear patterns in pronoun resolution.
Including GloVe embeddings improves performance in both automatic and gold models.	Word embeddings provide valuable semantic and syntactic information, enhancing the model's ability to disambiguate pronoun references.
BERT embeddings consistently outperform GloVe embeddings when combined with gold cluster features (Lines 10 & 14).	BERT's contextualized representations are particularly effective in capturing nuanced linguistic relationships in Persian pronoun resolution.

5-1-2- Hyper-Parameter Settings for Implemented Systems on Mehr Corpus

The performance of machine learning models is heavily influenced by the selection of appropriate hyperparameters. In this section, we detail the hyper-parameter tuning process employed for our mention-pair baseline and hybrid models on the Mehr corpus.

Mention-pair baseline :A grid search was conducted to identify the optimal hyper-parameter configuration for the random forest classifier. The hyper-parameter space, as outlined in Table 11 included various values for maximum depth, the number of estimators (trees), and the splitting criterion.

Table 11: 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

The grid search revealed that the optimal configuration for the mention-pair baseline model on the Mehr corpus was:

- **Max Depth:** 1000
- **Number of Estimators:** 20
- **Criterion:** Gini

To evaluate the model's robustness and generalization capabilities, 10-fold cross-validation was performed on the training data using these optimal hyperparameters,

yielding an average F1 score of 73.52%, as detailed in Table 12.

Table 12: The results of 10-fold cross-validation for the mention-pair model using the best hyper-parameters

precision	*recall*	*F1*
85.65	64.46	73.52

Hybrid Model: The same grid search procedure was employed for the hybrid model with the random forest classifier. The optimal hyperparameters for this model were found to be:

- **Max Depth:** 30
- **Number of Estimators:** 500
- **Criterion:** Entropy

Table 13 presents the results of 10-fold cross-validation on the training data, demonstrating the impact of incorporating word embeddings on model performance.

Table 13: 10-fold cross-validation results for the hybrid model using the best hyper-parameters

	precision	*recall*	*F1*
Feature vector results without considering word embedding	88.71	81.33	84.85
Feature vector results with considering word embedding	91.17	79.63	84.96

A two-hidden-layer architecture was employed for the hybrid model utilizing the MLP classifier, with the first layer consisting of 1500 nodes and the second layer containing 300 nodes. The ReLU activation function was used in both hidden layers. The optimal hyperparameters for this model, obtained through experimentation and fine-tuning, are summarized in Table 14. The training results for this configuration, as reported in Table 15 , demonstrate the effectiveness of the MLP classifier in conjunction with our hybrid approach.

Table 14: Optimal hyper-parameters of MLP hybrid model

First hidden layer	
Kernel initializer	identity
Regularization value	0.001
dropout	0.6
Second hidden layer	
Kernel initializer	identity
Regularization value	0.001
dropout	0.7
Other model settings	
Learning rates	0.0001
Optimization Algorithm	Adam

Table 15: 10-fold cross-validation results for the hybrid MLP model using the best parameters

precision	*recall*	*F1*
80.22	86.89	81.89

5-2- Experiments on RCDAT Corpus

This section evaluates our proposed hybrid model on the established RCDAT corpus, focusing on pronouns that are part of an existing coreference chain. Following the methodology of previous studies on this corpus, we utilize the Uppsala test documents for evaluation, comprising 415 sentences and 16,274 tokens. As with the Mehr corpus, we assess the model's performance using random forest and MLP classifiers across various configurations, with the results summarized in Table 16. Section 5.2.1 provides an in-depth analysis of the RCDAT results, while Section 5.2.2 details the hyper-parameter tuning process. It is crucial to emphasize that the evaluation of the RCDAT corpus expressly excludes non-referential pronouns and pronouns initiating new chains, as these annotations are not available in the corpus. This focused approach allows for direct comparison with previous research on the RCDAT corpus and provides a clear assessment of the model's ability to resolve pronouns within established coreference chains.

Table 16: The reported results of the presented system on RCDAT corpus compared to previous works and various settings

	System	*Using Rule-Based Modules*	*GloVe/BERT*	*Cluster-Based Features*	*precision*	*recall*	*F1*
1	Rule-based baseline	+	-	-	42.66	41.15	41.89
2	Mention-pair baseline (random forest)	-	-	-	50.91	48.67	49.5

	System			*Using Rule-Based Modules*	*GloVe/BERT*	*Cluster-Based Features*	*precision*	*recall*	*F1*
3	(random forest) [40]			-	GloVe	-	43.66	42.2	42.88
4	(MLP) [40]			-	GloVe	-	49.75	46.63	48.14
	Proposed system with different settings (random forest)								
5				-	-	system	41.83	40.35	41.08
6				+	-	system	49.8	48.2	49
7				+	GloVe	system	51.37	49.55	**50.45**
8				-	-	gold	53.76	51.85	52.79
9				+	-	gold	55.22	53.27	54.23
10				+	GloVe	gold	56.05	54.07	**55.04**
11				+	BERT	gold	58.48	56.09	**57.26**
12				+	BERT	system	53.15	51.19	**52.15**
	Proposed system with different settings (MLP)								
13				+	GloVe	gold	49.34	48.5	48.92
14				+	GloVe	system	46.78	45.13	45.94
15				+	BERT	system	48.90	46.48	47.66

5-2-1- RCDAT Corpus Result Analysis

The experimental results on the RCDAT corpus, as presented in Table 16, offer valuable insights into the efficacy of our hybrid model and its various configurations in resolving pronouns within established coreference chains. It is important to note that the RCDAT corpus lacks annotations for non-referential pronouns and pronouns initiating new chains. Thus, our evaluation focuses on pronouns with existing antecedents within the corpus. This allows for a direct comparison with previous work on the RCDAT corpus and provides a clear assessment of the model's performance in resolving pronouns within existing discourse contexts.

Hybrid Model's Consistent Superiority: The proposed hybrid model outperforms both the rule-based baseline (Line 1) and the mention-pair baseline (Line 2) across all configurations. This underscores the effectiveness of combining rule-based knowledge with machine-learning techniques for Persian pronoun resolution, even on a different corpus (RCDAT), compared to the previous section's focus on Mehr.

Impact of BERT Embeddings (Gold Clusters): When using gold cluster features, the addition of BERT embeddings consistently leads to substantial improvements in F1 score (57.26 vs. 52.79, Line 11 vs. Line 8), highlighting the crucial role of contextualized representations in accurately resolving pronouns within established coreference chains.

Significance of GloVe Embeddings (System Clusters): In configurations utilizing system-generated cluster features, the inclusion of GloVe embeddings provides a notable boost to F1 score (50.45 vs. 49.00, Line 7 vs. Line 6), indicating that these embeddings capture valuable semantic and syntactic relationships that aid in pronoun resolution, even when gold-standard cluster information is not available.

Rule-Based Modules' Continued Relevance: While BERT embeddings offer significant advantages, the inclusion of rule-based modules still positively impacts performance, especially when using system-generated cluster features. This suggests that rule-based knowledge remains valuable for resolving certain types of pronoun references, even when advanced contextual embeddings are available.

Random Forest vs. MLP: Similar to the results on the Mehr corpus, the random forest classifier outperforms the MLP classifier across comparable settings. This reinforces that the random forest's ability to model complex feature

interactions is particularly well-suited for pronoun resolution tasks.

Challenges of Automatic Feature Extraction: The performance gap between configurations using gold and system-generated cluster features underscores the continued need for improvement in automatic feature extraction techniques for Persian coreference resolution. This suggests that the quality of automatically generated features still lags behind gold standard annotations, impacting the performance of the hybrid model in real-world scenarios.

Potential for Further Improvement: While our hybrid model demonstrates significant advancements in Persian pronoun resolution, there remains room for further improvement. This could involve exploring more sophisticated feature engineering techniques, refining the rule-based modules, and investigating alternative machine learning architectures that can better leverage the information provided by contextualized embeddings like BERT.

5-2-2- Hyper-Parameter Settings for Implemented Systems on RCDAT Corpus

Hyper-parameter tuning is a critical step in optimizing the performance of machine learning models. In this section, we detail the hyper-parameter optimization process for our Persian pronoun resolution models on the RCDAT corpus. For all experiments, the antecedent candidate selection was restricted to a sentence window of 3, and the merging threshold was set to 0.2, as determined during the development phase.

Mention-Pair Model: The optimal hyperparameters for the random forest classifier in the mention-pair baseline were determined through a grid search, as outlined in Table 11. The best configuration, yielding an average F1 score of 69.00 in 10-fold cross-validation on the training data (Table 17), was:

- **Max Depth:** 40
- **Number of estimators:** 1000
- **Criterion:** Entropy

Table 17: The results of 10-fold cross-validation for the mention-pair model using the best hyper-parameters

precision	*recall*	*F1*
84.19	58.46	69

Hybrid Model: Similarly, a grid search was employed to optimize the hyperparameters of the random forest classifier in the hybrid model. The optimal configuration, resulting in an average F1 score of 84.96 in 10-fold cross-validation (Table 18), was:

- **Max Depth:** 50
- **Number of Estimators:** 1000
- **Criterion:** Entropy

Table 18: 10-fold cross-validation results for the hybrid model using the best hyper-parameters

	precision	*recall*	*F1*
Feature vector results without considering word embedding	88.75	81.48	84.95
Feature vector results with considering word embedding	88.79	81.46	84.96

The hybrid model with the MLP classifier utilized a three-layer architecture with 850, 680, and 440 nodes, respectively, using the ReLU activation function. Table 19 details the optimal hyperparameters for this model, including kernel initializers, regularization values, and dropout rates for each layer. The learning rate was set to 0.0001, and the Adam optimizer was employed for training.

Table 19: Optimal hyper-parameters of hybrid MLP model

First hidden layer	
Kernel initializer	glorot_normal
Regularization value	0.001
dropout	0.5
Second hidden layer	
Kernel initializer	identity
Regularization value	0.01
dropout	0.1
Third hidden layer	
Kernel initializer	identity
Regularization value	0.001
dropout	0.1
Other model settings	
Learning rates	0.0001
Optimization Alghorithm	Adam

6- Discussion and Implications

6-1- Comparative Analysis and Key Findings

Our experimental results on the Mehr and RCDAT corpora (Tables 9 & 16) provide compelling evidence for the effectiveness of our hybrid model in Persian pronoun resolution. Our approach consistently outperforms both rule-based and mention-pair baselines, underscored by the value of integrating linguistic knowledge with machine learning, especially when enhanced with contextualized embeddings like BERT and GloVe.
The results further highlight several key findings:

The Impact of Contextual Embeddings:
The incorporation of BERT embeddings, in particular, significantly boosts performance, demonstrating the importance of capturing contextual nuances for accurate pronoun resolution in Persian. Using embeddings further enhances accuracy, suggesting a complementary relationship in capturing different linguistic information.

The Results Further Highlight Several Key Findings:
The Impact of Contextual Embeddings: The incorporation of BERT embeddings, in particular, significantly boosts performance, demonstrating the importance of capturing contextual nuances for accurate pronoun resolution in Persian. The combination of BERT with GloVe embeddings further enhances accuracy, suggesting a complementary relationship in capturing different linguistic information.

The Value of Rule-Based Modules: While BERT embeddings play a crucial role, incorporating rule-based modules remains beneficial, especially for handling specific linguistic patterns and more straightforward coreference cases.

The Superiority of Random Forest: The random forest classifier consistently outperforms the MLP classifier, highlighting its effectiveness in modeling the complex interactions among diverse features.

6-2- Broader Implications and Contributions

The implications of our research extend beyond the Persian language. The hybrid approach presented in this paper offers a versatile framework for pronoun resolution that can be adapted to other under-resourced languages. By combining the strengths of rule-based and machine-learning methods and incorporating contextualized embeddings, our model provides a practical and interpretable solution for addressing the challenges of pronoun resolution in diverse linguistic contexts.
Furthermore, the development of the Mehr corpus represents a significant contribution to the Persian NLP community. This comprehensive resource, meticulously annotated with gold-standard coreference chains and enriched with automatically labeled linguistic elements, provides a valuable benchmark for future research and development in Persian coreference resolution.

6-3- Limitations and Future Directions

While our hybrid model demonstrates promising results, several limitations warrant further investigation. The performance gap between configurations using gold and system-generated cluster features highlights the need for continued research in improving automatic feature extraction methods. Additionally, exploring alternative machine learning architectures, such as transformer-based models specifically tailored for Persian pronoun resolution, may yield further performance gains.
In future work, we plan to investigate the effectiveness of our hybrid approach in other less-studied languages and explore its applicability to specific domains, such as legal and medical texts. We also aim to develop more sophisticated techniques for integrating external knowledge sources, such as knowledge graphs, to enhance the model's ability to resolve ambiguous pronoun references. Furthermore, we will explore methods to mitigate potential biases in the model's predictions, ensuring fairness and equity in its application.

7- Conclusion

This paper presents a novel hybrid entity-centric model for Persian pronoun resolution, effectively addressing the challenges posed by the need for more specialized resources and complex linguistic features in under-resourced languages. Our approach synergistically integrates rule-based modules, capturing linguistic patterns, with a machine-learning module enhanced by Persian BERT embeddings to leverage contextual information. This combination has resulted in significant performance improvements, surpassing existing Persian systems on both the Mehr and RCDAT corpora, with a notable 4.68 F1-score increase on the Mehr corpus and a 4.01 increase on the RCDAT corpus.
The introduction of the Mehr corpus, a comprehensive Persian coreference dataset annotated with a wide range of linguistic phenomena, significantly contributes to the research community. This corpus addresses the limitations of previous resources and provides a valuable benchmark for future research in Persian coreference resolution.
The performance of our hybrid model, particularly when incorporating BERT embeddings and cluster-level features, demonstrates the efficacy of combining diverse linguistic cues and contextual information for accurate pronoun resolution. Furthermore, the model's modularity, interpretability, and computational efficiency make it a practical solution for real-world applications.

Future work will focus on refining automatic feature extraction methods, exploring alternative machine-learning architectures, and incorporating additional knowledge sources to enhance performance further. We also aim to investigate the generalizability of our approach to other under-resourced languages and to adapt it to specific domains, such as legal and medical texts. By addressing these challenges, we strive to advance the state of the art in pronoun resolution, particularly for languages with limited resources, and contribute to developing more comprehensive and accurate natural language processing systems.

References

[1] Lee, K., et al. End-to-end Neural Coreference Resolution. in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017. Association for Computational Linguistics.

[2] Ng, V. and C. Cardie. Improving machine learning approaches to coreference resolution. in Proceedings of the 40th annual meeting on association for computational linguistics. 2002. Association for Computational Linguistics.

[3] Bahdanau, D., K.H. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. in 3rd International Conference on Learning Representations, ICLR 2015. 2015.

[4] Kriman, S. and H. Ji. Joint detection and coreference resolution of entities and events with document-level context aggregation. in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop. 2021.

[5] Liu, Z., K. Shi, and N. Chen. Coreference-Aware Dialogue Summarization. in Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2021.

[6] Mao, R. and X. Li. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. in Proceedings of the AAAI conference on artificial intelligence. 2021.

[7] Vu, H.H., H. Kamigaito, and T. Watanabe, Context-Aware Machine Translation with Source Coreference Explanation. Transactions of the Association for Computational Linguistics, 2024. 12: p. 856-874.

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

[9] Devlin, J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019.

[10] Hürriyetoğlu, A., et al., Event Coreference Resolution for Contentious Politics Events. arXiv preprint arXiv:.10123, 2022.

[11] Lu, Y., et al., End-to-end neural event coreference resolution. Artificial Intelligence, 2022. 303: p. 103632.

[12] Yu, X., W. Yin, and D. Roth. Pairwise Representation Learning for Event Coreference. in Proceedings of the 11th Joint Conference on Lexical and Computational Semantics. 2022.

[13] Zhang, W., S. Wiseman, and K. Stratos. Seq2seq is All You Need for Coreference Resolution. in The 2023 Conference on Empirical Methods in Natural Language Processing.

[14] Wan, B., et al., A deep neural network model for coreference resolution in geological domain. Information Processing & Management, 2023. 60(3): p. 103268.

[15] Žabokrtský, Z., et al. Findings of the second shared task on multilingual coreference resolution. in Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution. 2023.

[16] Bi, M., et al., Machine reading comprehension combined with semantic dependency for Chinese zero pronoun resolution. Artificial Intelligence Review, 2023. 56(8): p. 7597-7612.

[17] Demir, Ş., Neural coreference resolution for Turkish. Journal of Intelligent Systems: Theory and Applications, 2023. 6(1): p. 85-95.

[18] Lu, J. and V. Ng. Learning antecedent structures for event coreference resolution. in 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). 2017. IEEE.

[19] Hobbs, J.R., Resolving pronoun references. Lingua, 1978. 44(4): p. 311-338.

[20] Carbonell, J.G. and R.D. Brown. Anaphora resolution: a multi-strategy approach. in Proceedings of the 12th conference on Computational linguistics-Volume 1. 1988. Association for Computational Linguistics.

[21] Lappin, S. and H. Leass, An algorithm for pronominal anaphora resolution. Computational linguistics, 1994. 20(4): p. 535-561.

[22] Kennedy, C. and B. Boguraev. Anaphora for everyone: pronominal anaphora resoluation without a parser. in Proceedings of the 16th conference on Computational linguistics-Volume 1. 1996. Association for Computational Linguistics.

[23] Stuckardt, R., Design and enhanced evaluation of a robust anaphor resolution algorithm. Computational Linguistics, 2001. 27(4): p. 479-506.

[24] Haghighi, A. and D. Klein. Simple coreference resolution with rich syntactic and semantic features. in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. 2009. Association for Computational Linguistics.

[25] Raghunathan, K., et al. A multi-pass sieve for coreference resolution. in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010. Association for Computational Linguistics.

[26] Lee, H., et al. 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. 2011. Association for Computational Linguistics.

[27] Lee, H., et al., Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics, 2013. 39(4): p. 885-916.

[28] Van Cranenburgh, A., et al. A Hybrid Rule-Based and Neural Coreference Resolution System with an Evaluation on

Dutch Literature. in Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference. 2021.
[29] Lee, K., L. He, and L. Zettlemoyer. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018.
[30] Joshi, M., et al., Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 2020. 8: p. 64-77.
[31] Zhang, H., et al., Knowledge-aware Pronoun Coreference Resolution. arXiv preprint arXiv:.03663, 2019.
[32] Zhang, H., Y. Song, and Y. Song, Incorporating Context and External Knowledge for Pronoun Coreference Resolution. arXiv preprint arXiv:.10238, 2019.
[33] Hou, F., et al., Anisotropic span embeddings and the negative impact of higher-order inference for coreference resolution: An empirical analysis. Natural Language Engineering, 2024: p. 1-22.
[34] De Langhe, L., O. De Clercq, and V. Hoste, Constructing a cross-document event coreference corpus for Dutch. Language Resources and Evaluation, 2022: p. 1-30.
[35] Rahimi, Z. and S. HosseinNejad, Corpus based coreference resolution for Farsi text Signal and Data Processing2020. 17(1): p. 79-98.
[36] Haji Mohammadi, H., et al., Mehr: A Persian Coreference Resolution Corpus. Journal of AI and Data Mining, 2023.
[37] Biau, G. and E. Scornet, A random forest guided tour. Test, 2016. 25: p. 197-227.
[38] Artari, V.K.P., et al. A Multi-Pass Sieve Coreference Resolution for Indonesian. in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021.
[39] Sahlani, H., M. Hourali, and B. Minaei-Bidgoli, Coreference resolution with deep learning in the persian language. Signal and Data Processing2020.
[40] Sahlani, H., M. Hourali, and B. Minaei-Bidgoli, Coreference Resolution Using Semantic Features and Fully Connected Neural Network in the Persian Language. International Journal of Computational Intelligence Systems, 2020.
[41] Soon, W.M., H.T. Ng, and D.C.Y.J.C.l. Lim, A machine learning approach to coreference resolution of noun phrases. Computational linguistics, 2001. 27(4): p. 521-544.
[42] Ng, V. and C. Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. in COLING 2002: The 19th International Conference on Computational Linguistics. 2002.
[43] Denis, P. and J. Baldridge. Specialized models and ranking for coreference resolution. in Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008. Association for Computational Linguistics.
[44] Denis, P. and J. Baldridge, Global joint models for coreference resolution and named entity classification. Procesamiento del Lenguaje Natural, 2009. 42.
[45] Wiseman, S.J., et al. Learning anaphoricity and antecedent ranking features for coreference resolution. in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. Association for Computational Linguistics.
[46] Lee, K., et al., End-to-end neural coreference resolution. arXiv preprint arXiv:.05365, 2017.
[47] Luo, X., et al. A mention-synchronous coreference resolution algorithm based on the bell tree. in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. 2004. Association for Computational Linguistics.
[48] Nicolae, C. and G. Nicolae. Bestcut: A graph algorithm for coreference resolution. in Proceedings of the 2006 conference on empirical methods in natural language processing. 2006. Association for Computational Linguistics.
[49] Culotta, A., M. Wick, and A. McCallum. First-order probabilistic models for coreference resolution. in Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. 2007.
[50] Lee, H., M. Surdeanu, and D. Jurafsky, A scaffolding approach to coreference resolution integrating statistical and rule-based models. Natural Language Engineering, 2017. 23(5): p. 733-762.
[51] Lai, T.M., 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. IEEE.
[52] Zhang, R., et al., Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. arXiv preprint arXiv:1805.04893, 2018.
[53] Joshi, M., et al. BERT for Coreference Resolution: Baselines and Analysis. in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.
[54] Kantor, B. and A. Globerson. Coreference resolution with entity equalization. in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
[55] Werlen, L.M. and J. Henderson. Graph Refinement for Coreference Resolution. in Findings of the Association for Computational Linguistics: ACL 2022. 2022.
[56] Zhang, W., S. Wiseman, and K. Stratos, Seq2seq is all you need for coreference resolution. arXiv preprint arXiv:2310.13774, 2023.
[57] Guo, Q., et al. Dual Cache for Long Document Neural Coreference Resolution. in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.
[58] Moosavi, N.S. and G. Ghassem-Sani, A ranking approach to Persian pronoun resolution. Advances in Computational Linguistics. Research in Computing Science, 2009. 41: p. 169-180.
[59] Denis, P. and J. Baldridge. A Ranking Approach to Pronoun Resolution. in IJCAI. 2007.
[60] Lee, K., L. He, and L. Zettlemoyer, Higher-order coreference resolution with coarse-to-fine inference. arXiv preprint arXiv:.05392, 2018.
[61] Chai, H. 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.
[62] Miculicich, L. and J. Henderson, Graph Refinement for Coreference Resolution. arXiv preprint arXiv:.16574, 2022.
[63] Wang, Y., Y. Shen, and H. Jin. An End-To-End Actor-Critic-Based Neural Coreference Resolution System. in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. IEEE.
[64] Hajimohammadi, H., A. Talebpour, and A. Mahmoudi Aznaveh. Using decision tree to Persian pronoun resolution. in Twenty-first annual National Conference of the Computer Society of Iran. 2016. Tehran: Computer Society of Iran.
[65] Nourbakhsh, A. and M. Bahrani. Persian Pronoun Resolution Using Data Driven Approaches. in International Conference on Information and Software Technologies. 2017. Springer.
[66] Fallahi, F. and M. Shamsfard, Recognizing anaphora reference in Persian sentences. International Journal of Computer Science Issues (IJCSI), 2011. 8(2): p. 324.
[67] Nazaridoust, M., B.M. Bidgoli, and S. Nazaridoust. Co-reference Resolution in Farsi Corpora. in In Advance Trends in Soft Computing: Proceedings of WCSC 2013. 2013. Cham: Springer International Publishing.
[68] Asgari-Bidhendi, M., A. Hadian, and B. Minaei-Bidgoli, Farsbase: The persian knowledge graph. Semantic Web, 2019. 10(6): p. 1169-1196.
[69] Bijankhan, M., The role of the corpus in writing a grammar: An introduction to a software. Iranian Journal of Linguistics, 2004. 19(2): p. 48-67.
[70] Mirzaei, A. and P. Safari. Persian Discourse Treebank and coreference corpus. in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.
[71] Yimam, S.M., et al. 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.
[72] Hirschman, L., MUC-7 coreference task definition, version 3.0. Proceedings of MUC-7,, 1997.
[73] Bergsma, S. and D. Lin. Bootstrapping path-based pronoun resolution. in Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics. 2006.
[74] Ji, H. and D. Lin. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. in Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1. 2009.
[75] Rasooli, M.S., 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.
[76] Seraji, M., B. Megyesi, and J. Nivre, Bootstrapping a Persian dependency treebank. Linguistic Issues in Language Technology, 2012. 7(18).