# هم‌مرجع‌یابی مبتنی بر پیکره در متون فارسی

**زینب رحیمی* و شادی حسین نژاد**

پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی، گروه پردازش صوت و زبان طبیعی، تهران، ایران



## چکیده

**مرجع‌یابی یا مرجع‌گزینی یا پیداکردن واژگان هم‌مرجع در متن، یکی از وظایف مهم در پردازش زبان طبیعی است که یک بخش عملیاتی مهم در مسائلی مانند خلاصه‌سازی خودکار، پرسش و پاسخ خودکار و استخراج اطلاعات به‌شمار می‌رود. طبق تعاریف زمانی، دو واژه زمانی هم‌مرجع هستند که هر دو به موجودیت واحدی در متن یا جهان حقیقی ارجاع بدهند. تاکنون برای حل این مسأله تلاش‌های متعددی صورت گرفته است که بنابر نتایج این مطالعات، عملیات مرجع‌گزینی را می‌توان با روش‌های متفاوتی مانند روش‌های قاعده‌مند، مبتنی بر قوانین مکاشفه‌ای و روش‌های یادگیری ماشین (بانظارت یا بی‌ناظر) انجام داد. نکته قابل توجه این است که در سال‌های اخیر استفاده از پیکره‌های برچسب‌گذاری‌شده در این زمینه رواج زیادی داشته و منجر به تولید نتایج مناسبی هم شده است. با تکیه بر این موضوع، در پژوهش حاضر، یک پیکره از واژگان هم‌مرجع تولید شده که حدود یک‌میلیون واژه به‌همراه برچسب موجودیت نامدار دارد. در بخش مرجع‌گزینی تمام گروه‌های اسمی، ضمایر و موجودیت‌های نامدار برچسب‌گذاری شده‌اند و برچسب‌های موجودیت نامدار پیکره شامل هفت برچسب است. در پژوهش حاضر با استفاده از این پیکره، یک ابزار مرجع‌گزینی خودکار با استفاده از ماشین بردار پشتیبان تولید شده که دقت آن بر روی داده‌های آزمایش طلایی در حدود شصت درصد است.**

**واژگان کلیدی: هم‌مرجع یابی خودکار، مرجع‌گزینی، تحلیل مرجع ضمیر، عبارات ارجاعی.**

# Corpus based coreference resolution for Farsi text

**Zeinab Rahimi* & Shadi Hossein Nejad**

Research Center for Development of Advanced Technology (RCDAT),
Speech and natural language processing department, Tehran, Iran

## Abstract

**"Coreference resolution" or "finding all expressions that refer to the same entity" in a text, is one of the important requirements in natural language processing. Two words are coreference when both refer to a single entity in the text or the real world. So the main task of coreference resolution systems is to identify terms that refer to a unique entity. A coreference resolution tool could be used in many natural language processing tasks such as machine translation, automatic text summarization, question answering, and information extraction systems. Adding coreference information can increase the power of natural language processing systems.**

**The coreference resolution can be done through different ways. These methods include heuristic rule-based methods and supervised/unsupervised machine learning methods. Corpus based and machine learning based methods are widely used in coreference resolution task in recent years and has led to a good performance. For using such these methods, there is a need for manually labeled corpus with sufficient size. For Persian language, before this research, there exists no such corpus. One of the important targets here,**

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۷/۳/۱۹ • تاریخ پذیرش: ۱۳۹۸/۳/۱۱ • تاریخ انتشار: ۱۳۹۹/۰۴/۰۱ • نوع مطالعه: پژوهشی

**was producing a through corpus that can be used in coreference resolution task and other associated fields in linguistics and computational linguistics.**

**In this coreference resolution research, a corpus of coreference tagged phrases has been generated (manually annotated) that has about one million words. It also has named entity recognition (NER) tags. Named entity labels in this corpus include 7 labels and in coreference task, all noun phrases, pronouns and named entities have been tagged. Using this corpus, a coreference tool was created using a vector space machine, with precision of about 60% on golden test data.**

**As mentioned before, this article presents the procedure for producing a coreference resolution tool. This tool is produced by machine learning method and is based on the tagged corpus of 900 thousand tokens. In the production of the system, several different features and tools have been used, each of which has an effect on the accuracy of the whole tool. Increasing the number of features, especially semantic features, can be effective in improving results. Currently, according to the sources available in the Persian language, there are no suitable syntactic and semantic tools, and this research suffers from this perspective.**

**The coreference tagged corpus produced in this study is more than 500 times bigger than the previous Persian language corpora and at the same time it is quite comparable to the prominent ACE and Ontonotes corpora.**

**The system produced has an f-measure of nearly 60 according to the CoNLL standard criterion. However, other limited studies conducted in Farsi have provided different accuracy from 40 to 90%, which is not comparable to the present study, because the accuracy of these studies has not been measured with standard criterion in the coreference resolution field.**

**Keywords: Automatic coreference resolution, Anaphora resolution, mention.**



## ۱- مقدمه

مسأله مرجع‌گزینی و یا به‌عبارت دیگر یافتن کلیه عباراتی که به یک موجودیت دلالت دارند، یک بخش عملیاتی مهم در مسائلی مانند خلاصه‌سازی خودکار، پرسش و پاسخ خودکار و استخراج اطلاعات به‌شمار می‌رود. تاکنون روش‌های متنوعی برای حل مسأله مرجع‌گزینی پیشنهاد شده است که به دو دسته تقسیم می‌شوند: «روش‌های مبتنی بر قانون و روش‌های یادگیری ماشین».

روش‌های مبتنی بر قانون مجموعه‌ای از قوانین مکاشفه‌ای را در نظر می‌گیرند که از اطلاعات نحوی و معنایی واژگان برای پیداکردن مرجع ضمیر استفاده می‌کنند. در طرف دیگر روش‌های یادگیری ماشین قرار دارند که در آنها از یک پیکره برچسب‌گذاری‌شده -که در آن مراجع عبارت‌ها مشخص شده است- استفاده می‌شود تا یک الگوریتم یادگیری بتواند به شکل خودکار قوانین مرجع‌گزینی را بیاموزد. بزرگ‌ترین مشکل این روش‌ها برای زبان‌هایی مانند زبان فارسی کمبود داده برچسب‌گذاری‌شده است.

در این مقاله به معرفی یک پیکره واژگان هم‌مرجع برای زبان فارسی و سامانه مرجع‌گزینی متناظر آن پرداخته شده است. پیکره عنوان‌شده شامل حدود یک‌میلیون واژهٔ فارسی است که برچسب دستی[۱] (دقیق) هم‌مرجع و موجودیت نامدار دارند. همچنین علاوه بر برچسب‌های دقیق، چند برچسب غیردقیق[۲] به پیکره اضافه شده است که این برچسب‌ها خروجی ابزارهای پیش‌پردازشی تهیه شده در پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی هستند. برای تهیه سامانه مرجع‌یابی از روش یادگیری ماشین مبتنی بر پیکره استفاده شده است.

معماری مورد استفاده، معماری جدیدی نیست که برای نخستین‌بار در این مقاله پیشنهاد شده باشد؛ اما در این پژوهش در وهله نخست منبع ارزشمند دادگان فارسی برای مسأله مرجع‌گزینی تهیه شده است (هدف اصلی) و دوم این که برای هر یک از ماژول های یادشده در معماری، تلاش شده تا با استفاده از ابزارهای موجود و روش‌ها و الگوریتم‌های موجود بهترین انطباق با دادگان انجام و بهترین ترکیب ممکن ایجاد شود. بدین ترتیب قابل استفاده بودن پیکره اثبات شده است. همچنین این سامانه خودکار انتهابه‌انتها[۳] (واقعی و در دسترس) با کاربری شناسایی عبارات هم‌مرجع (ضمیر و عبارات ارجاعی) برای زبان فارسی مطابق اطلاع نگارنده جدید و نوآورانه است.

در این مقاله که علاوه‌بر مقدمه در شش بخش تهیه شده است، به معرفی پیکره واژگان هم‌مرجعِ ایجادشده و الگوریتم مورد استفاده در ابزار هم‌مرجع‌یاب (مرجع‌گزینی) پرداخته می‌شود. پس از مقدمه، در بخش دوم این مقاله پیکره‌های واژگان هم‌مرجع در زبان فارسی و سایر زبان‌ها

[۱] برچسب دقیق برچسبی است که توسط انسان تولید شده است.

[۲] برچسب غیردقیق خروجی ابزارهای پردازشی است و توسط ماشین تولید شده است.

[3] end-to-end

معرفی و همچنین سامانه‌های مرجع‌گزینی معتبر در سطح جهانی و سامانه‌های مرجع‌گزینی موجود در زبان فارسی نیز معرفی می‌شوند. در بخش سوم پیکره واژگان هم‌مرجع به طور کامل توصیف و در بخش چهارم الگوریتم مورد استفاده تشریح می‌شود. بخش پنجم به ارزیابی و نتایج سامانه مرجع‌گزینی اختصاص دارد و در بخش ششم جمع‌بندی مطالب ارائه می‌شود و کارهای پیش‌رو بررسی خواهند شد.

## ۲- مروری بر کارهای انجام‌شده

### ۲-۱- مروری بر پیکره‌های تولیدشده

از دهه ۹۰ میلادی بیشتر رویکردها در تولید سامانه‌های مرجع‌گزینی، به سمت روش‌های یادگیری ماشین تمایل پیدا کرده است. در این راستا رقابت‌های مختلفی مانند رقابت‌های MUC [6,7] و CoNLL با موضوع مرجع‌گزینی برای زبان انگلیسی و سایر زبان‌ها برگزار شده است. در خلال این رقابت‌ها پیکره‌هایی نیز تولید شده است. از معروف‌ترین پیکره‌های هم‌مرجع می‌توان پیکره‌های MUC را نام برد. این پیکره‌ها برخلاف حجم محدودی که دارند، به‌صورت گسترده در پژوهش‌های مختلف به کار رفته‌اند.

از طرف دیگر پیکره ACE [9] فقط به عباراتی که موجودیت نامدار هستند، برچسب هم‌مرجع اختصاص داده است. این پیکره‌ها شامل زبان‌های انگلیسی، عربی و چینی می‌شوند. در پیکره ACE-2 از داده‌های خبری انگلیسی استفاده شده است و در پیکره ACE-5 که زبان‌های عربی و چینی را نیز در بر می‌گیرد از داده‌های مختلفی همچون داده‌های تلفنی، وبلاگ‌ها و غیره نیز علاوه‌بر داده خبری استفاده شده است.

در پیکره OntoNote که براساس درخت‌بانک PennTreeBank تهیه شده، زیربخش زبان انگلیسی و چینی دارای حدود یک‌میلیون واژه است که از منابعی شامل خبرگزاری‌ها، مجلات، اخبار تلویزیونی و محاورات تلویزیونی و ... جمع‌آوری شده است. در قسمت انگلیسی این پیکره دویست هزار واژه از ترجمه انگلیسی «عهد جدید» نیز اضافه شده است. زیربخش عربی، کمی کوتاه‌تر است و فقط شامل سیصد هزار واژه از منابع خبری عربی است.

در جدول (۱) آمار مقایسه‌ای پیکره‌های مختلف دیده می‌شود.

(جدول-۱): مقایسه پیکره‌های واژگان هم‌مرجع

(Table-1): Comparison of coreferenced word corpora

| نام پیکره | زبان | حجم واژگان پیکره |
|---|---|---|
| MUC-6 [6] | English | 25 K |
| MUC-7 [7] | English | 40 K |
| ACE [9] | English<br>Arabic<br>Chinese | 960 K<br>500 K<br>615 K |
| Ontonotes [20] | English<br>Arabic<br>Chinese | 1.6 M<br>300 K<br>950 K |

#### ۲-۱-۱-۲- پیکره‌های مرجع‌گزینی در زبان فارسی

مرجع‌گزینی در زبان فارسی موضوع پژوهشی جدیدی است که تنها در چند سال اخیر به آن توجه شده است. بنا به اطلاع نویسندگان، پیش از این به‌طور سازمان‌یافته و جامع تلاشی برای تولید ابزار یا داده مرجع‌گزینی در زبان فارسی صورت نگرفته؛ لذا در زبان فارسی دادگان برای پیکره هم‌مرجع به‌صورت بسیار محدود تولید شده است. موارد زیر تنها دادگانی هستند که در زبان فارسی تولید شده‌اند:

##### ۲-۱-۱-۱- پیکره PCAC-2008[1]

این پیکره مجموعه‌ای است شامل ۳۱ متن برگرفته از پیکره بی‌جن‌خان که در آن نزدیک‌ترین مرجع اسمی ۲۰۷۹ ضمیر مشخص شده است [18]. دسترسی به اطلاعات مربوط به مرجع ضمیر در بسیاری از کاربردهای پردازش زبان طبیعی چون ترجمه ماشینی، پرسش و پاسخ خودکار و خلاصه‌سازی خودکار دارای اهمیت است. پیکره استفاده‌شده برای آزمایش سامانه [11] از پنج وبلاگ فارسی که به شکل اتفاقی از هر کدام بیست صفحه بارگیری‌شده ساخته شده است [10].

##### ۲-۱-۱-۲- پیکره لوتوس[2]

پیکره لوتوس، مجموعه‌ای از پنجاه متن به‌نسبه بلند برگفته از پیکره بی‌جن‌خان است که عبارات اسمی هم‌مرجع در آن مشخص شده است. برای مثال در جمله «[پروفسور عسکرزاده] [بنیان‌گذار منطق فازی] است و از [او] آثار بسیاری در این زمینه منتشر شده است.» مواردی که با قلاب مشخص شده‌اند به یک موجودیت واحد اشاره دارند. دسترسی به چنین اطلاعاتی در بسیاری از کاربردهای پردازش زبان و از جمله استخراج اطلاعات دارای اهمیت است [21].

با توجه به حجم دادگان موجود در زبان فارسی، واضح است که نیاز به پیکره جامع واژگان هم‌مرجع با حجم

[1] PCAC-2008 Persian Coreferentially Annotated Corpus
[2] Lotus

واژگان مناسب وجود دارد. همچنین لازم است، پیکره تولیدشده دارای برچسب برای موارد بیشتر واژگان هم‌مرجع علاوه‌بر ضمایر باشد؛ امری که تاکنون به آن توجه نشده است.

**۲-۱-۲- معیار سنجش صحت برچسب‌گذاری در پیکره‌ها**

در تولید پیکره‌ها توجه به میزان هماهنگی برچسب‌گذارها ضروری است. در برچسب‌گذاری دستی یک پیکره با حجم بالا، اختلاف نظر بین برچسب‌گذارها اجتناب‌ناپذیر است؛ لذا در تولید هر پیکره‌ای تلاش بر این است که برچسب‌گذاری‌ها به‌صورت یک‌دست و با دقت مناسب انجام شود. در این راستا از معیارهایی جهت بررسی میزان هماهنگی استفاده می‌شود. یکی از معروف‌ترین، معتبرترین و پرکاربردترین معیارها در زمینه تولید پیکره‌های متنی معیار کاپا[1] [8] است.

این معیار، معیاری آماری است که در زبان‌شناسی پیکره‌ای برای بررسی میزان تفاهم میان دو برچسب‌گذاری بر روی پیکره به کار می‌رود. در این معیار تلاش شده است که برچسب‌گذاری‌هایی که به‌طور اتفاقی شبیه به یکدیگر شده‌اند تأثیری در نتیجه نداشته باشند. به این منظور کاپا به‌صورت فرمول (۱) تا (۳) محاسبه می‌شود [12].

در این فرمول Pr(a) میزان تشابه صحیح میان دو برچسب‌گذاری است و Pr(e) میزان تشابه اتفاقی میان دو برچسب‌گذاری است. برای استفاده از معیار کاپا در مواردی که تعداد برچسب‌گذارها بیشتر از دو نفر است، این مقدار به‌ازای هر دو برچسب‌گذار محاسبه شده و از مقادیر آن میانگین گرفته می‌شود. مثال زیر (جدول ۲) نمونه‌ای از برچسب‌گذاری بر روی داده است که در دو گروه مثبت و منفی برچسب‌گذاری شده است.

**(جدول-۲): نمونه‌ای از برچسب‌گذاری برای محاسبه کاپا**
**(Table-2): Example of tagging to calculate kappa**

| | | برچسب گذار نخست | | |
|---|---|---|---|---|
| | | منفی | مثبت | کل |
| برچسب‌گذار دوم | مثبت | P11 | P12 | P1_ |
| | منفی | P21 | P22 | P2_ |
| | کل | P_1 | P_2 | 1 |

گروه مثبت و منفی در جدول بالا به‌عنوان مثال از نام دو برچسب (رده) فرضی (برای بررسی تعداد مواردی که برچسب‌گذارها مثل هم برچسب یا اتفاق نظر نداشته‌اند) برای محاسبه کاپا آورده شده است. محاسبه فرمول کاپا به‌صورت زیر است:

$$\Pr(a) = P11 + P22 \quad (۱)$$

$$\Pr(e) = P\_1 * P1\_ + P\_2 * P2\_ \quad (۲)$$

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (۳)$$

**۳-۱-۲- معیارهای ارزیابی سامانه‌های مرجع‌گزینی**

انتخاب معیار مناسب برای ارزیابی سامانه‌هایی که عملیات مرجع‌گزینی را به‌طور خودکار انجام می‌دهند، همواره مورد توجه پژوهش‌گران این حوزه بوده است. از سال ۱۹۹۵ میلادی تاکنون معیارهای مختلفی ارائه شده که هر کدام به‌نوعی تلاش کرده است کیفیت سامانه مرجع‌گزینی را بهتر نمایش دهد. معیار ارزیابی باید دارای دو خاصیت مهم باشد. نخست این که عدد خروجی باید با استنباط کاربر از خروجی سامانه هم‌خوانی داشته و از طرف دیگر یک معیار خوب باید قدرت تمایز بالایی داشته باشد و بتواند به‌خوبی یک سامانه کارآمد را از یک سامانه ضعیف جدا کند. چهار معیار پرکاربرد MUC، B3، CEAF و CoNLL در زمینه مرجع‌گزینی به‌طور گسترده مورد استفاده قرار می‌گیرند.

**۱-۳-۱-۲- معیار MUC**

این روش ارزیابی در سال ۱۹۹۵ میلادی ارائه شد [29]. در این معیار key یا کلید، خوشه‌هایی از عبارات هم‌مرجع هستند که توسط انسان مشخص شده‌اند و response یا پاسخ، خوشه‌هایی هستند که سامانه مرجع‌گزینی تشخیص داده است. مقدار بازیابی به‌صورت نسبت تعداد پیوندهای مشترک در کلید و پاسخ بر تعداد پیوندهای کلید محاسبه می‌شود و دقت به‌صورت نسبت تعداد پیوندهای مشترک در کلید و پاسخ بر تعداد پیوندهای پاسخ محاسبه می‌شود.

**۲-۳- ۱-۲- معیار B cube**

این معیار در سال ۱۹۹۸ ارائه شد [2]. هدف از ارائه این معیار رفع برخی از مشکلاتی است که در معیار ارزیابی MUC وجود داشت. یکی از مشکلات معیار MUC این است که تفاوتی بین میزان بدی اشتباهات قائل نیست. در معیار B3 تلاش شده که این حس شهودی در نظر گرفته شود. در این معیار به جای این که به پیوندهای بین عبارات نگاه شود، به خود عبارات و حضور و یا عدم حضور آنها در یک رده هم‌ارزی توجه می‌شود. در نتیجه مقدار بازیابی و دقت برای هر عبارت محاسبه می‌شود و سپس با هم ادغام می‌شوند که بازیابی و دقت نهایی را تولید کنند.

[1] Kappa

(۴)

$$Precision_i = \frac{number\ of\ correct\ elements\ in\ the\ output\ chain\ containing\ entity_i}{number\ of\ elements\ in\ the\ output\ chain\ containing\ entity_i}$$

(۵)

$$Recall_i = \frac{number\ of\ correct\ elements\ in\ the\ output\ chain\ containing\ entity_i}{number\ of\ elements\ in\ the\ truth\ chain\ containing\ entity_i}$$

رابطه‌های (۴) و (۵) برای محاسبه دقت و بازیابی هر یک از عبارات استفاده می‌شود؛ سپس با استفاده از مقادیر بالا، مقدار بازیابی و دقت نهایی سامانه محاسبه می‌شود.

### ۳-۳-۱-۲- معیار CEAF

معیار CEAF[1] در سال ۲۰۰۵ میلادی ارائه شده است [17]. در تعریف این معیار تلاش شده که ایراد معیار B3 برطرف شود. در معیار B3 ممکن است عبارات پاسخ یا کلید بیش از یک بار در محاسبه دقت نقش داشته باشند. برای رفع این مشکل در معیار CEAF تلاش شده است که رابطه بهینه یک به یکی میان عبارات پاسخ و کلید پیدا شود. به‌دلیل این‌که این رابطه یک‌به‌یک است، در این معیار از تمام عبارات موجود در مجموعه پاسخ و کلید استفاده نمی‌شود.

در این معیار $K_{min}(d)$ و $S_{min}(d)$ به‌صورت خوشه‌هایی از عبارات هم‌مرجع که انسان تولید کرده است (K) و خوشه‌هایی از عبارات هم‌مرجع که سامانه مرجع‌گزینی تولید کرده است (S) تعریف می‌شوند. تابع یک به یکی که میان این دو مجموعه ساخته می‌شود (برهم‌نهی[2] میان دومجموعه) به‌صورت رابطه (۶) است:

$$g(K_i) = S_j, K_i \in K_{min}(d) and\ S_j \in S_{min}(d). \quad (۶)$$

### ۴-۳- ۱-۲- معیار CoNLL

نتایج سامانه‌های مرجع‌گزینی ارائه‌شده در رقابت CoNLL2012 [21] می‌تواند معیار مناسبی برای سنجش نتایج باشد؛ زیرا در این رقابت‌ها علاوه‌بر این‌که از پیکره‌ای با حجم قابل‌قبول (پیکره Ontonotes با حجم یک میلیون واژه) استفاده شده است، معیار ارزیابی منصفانه‌ای نیز لحاظ شده است. این معیار CoNLL نام دارد که به‌صورت میانگین حسابی (بدون وزن) سه معیار پرکاربرد B3،MUC و CEAF محاسبه می‌شود.

[1] Constrained Entity – Alignment F-measure
[2] alignment

## ۲-۲- مروری بر روش‌های مرجع‌گزینی

### ۱-۲-۲- روش‌های مطرح‌شده برای انگلیسی

در این بخش به بررسی روش‌های مختلف مورد استفاده در امر مرجع‌گزینی می‌پردازیم.

مقاله لی و همکاران عملیات مرجع‌گزینی را با استفاده از مجموعه‌ای از قوانین قطعی و دقیق انجام می‌دهد. ایده اصلی استفاده از مجموعه‌ای از قوانین، که از قوانین با دقت بالا شروع می‌شود و با اضافه‌کردن قانون‌های جدید سعی در بالابردن بازیابی دارد، نخستین‌بار توسط برک و کالینز [4] ارائه شد [16]. او پیشنهاد داده بود که از هفت قانون به‌عنوان فیلترهایی برای استخراج روابط هم‌مرجعی استفاده شود. ایده استفاده از قوانین بعدتر توسط تعداد زیادی از پژوهش‌گران به‌کار گرفته شد [14,19]. اما تا قبل از [13] هیچ کدام از سامانه‌های مرجع‌گزینی مبتنی بر قانون موفق به رسیدن به بازدهی‌های گزارش‌شده توسط سامانه‌های مبتنی بر یادگیری ماشین نشدند؛ اما در رویکرد ارائه‌شده توسط حقیقی در [13] برای نخستین‌بار بازدهی یک سامانه مبتنی بر قوانین در حد و اندازه‌های سامانه‌های یادگیری ماشین با ناظر و حتی جلوتر از آنها بود. پس از حقیقی، پژوهش‌گران دانشگاه استنفورد در طی چند مقاله که در سال‌های ۲۰۱۰، ۲۰۱۱ و ۲۰۱۲ منتشر کردند، سامانه مبتنی بر قانونی را ارئه کردند که از کلیه سامانه‌های ارائه‌شده تاکنون پاسخ بهتری داشت. قوانین استفاده‌شده در این سامانه سلسله‌مراتبی به‌نسبه ساده و مستقل از زبان انگلیسی هستند و به‌دلیل همین خصوصیت اقبال زیادی به آن شد.

در روش مورد استفاده در سامانه تولیدشده در دانشگاه برکلی [10] با استفاده از یک مدل log-linear تلاش شده است تا بدون استفاده از ویژگی‌های زیاد مدل مرجع‌گزینی تولید شود. در مرجع‌گزینی مجموعه‌ای از عبارات نامزد وجود دارند که می‌توانند با یکدیگر هم‌مرجع باشند. این عبارات را در این مقاله «نامیده[3]» می‌نامیم. در این مدل ذکر شده است که برای هر نامیده تنها یک مرجع قبلی[4] پیدا می‌شود یا این‌که آن نامیده یک زنجیره جدید می‌سازد.

در این سامانه از شش ویژگی نوع نامیده (اسمی، ضمیری، اسم خاص)، شباهت تمام رشته نامیده (شباهت دقیق)، هسته معنایی نامیده، واژهٔ ابتدایی و انتهایی نامیده، واژگان بلافاصله قبل و بعد از نامیده و فاصله بین نامیده‌ها

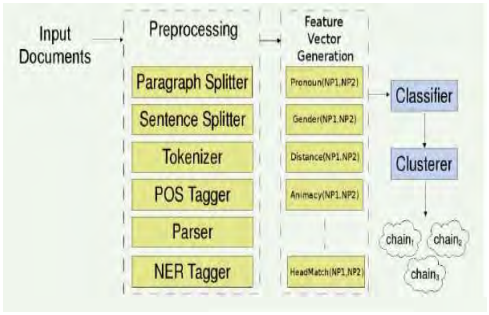
[3] mention
[4] antecedent

(تعداد جمله‌ها و تعداد نامیده‌ها بین دو نامیده) استفاده شده است. این سامانه بر روی داده CoNLL 2011 سنجیده شده که بهترین نتیجه آن ۶۵٪ براساس معیار MUC بوده است.

در سامانه دانشگاه ایلی‌نویز و مقاله منتشرشده مربوط به آن نشان داده شده است که استفاده از یک مدل ساده با کمک مجموعه ویژگی‌های قوی می‌تواند به نتایج خوبی در زمینه مرجع‌گزینی منجر شود. این رویکرد مسأله مرجع‌گزینی را به‌صورت مسأله گراف مدل کرده است. هر گره[1] گراف و شاید متن همجوار آن یک نامیده در نظر گرفته شده است. در صورتی‌که دو نامیده با یکدیگر هم‌مرجع باشند یک پیوند در گراف بین آن دو برقرار می‌شود. مدل هم‌مرجع در این مقاله براساس جفت[2] است؛ یعنی هر نامیده با نامیده‌های قبلی خود مقایسه می‌شود تا مشخص شود کدامیک، هم مرجع آن است [3].

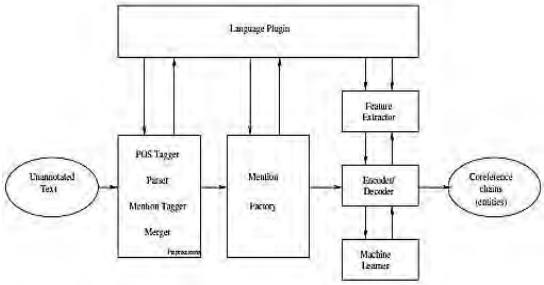
ویژگی‌های استفاده‌شده در این سامانه شامل نوع نامیده (اسم خاص، ضمیر یا گروه اسمی)، رابطه بین دو رشته، ویژگی‌های معنایی (جنس، شمار، رابطه معنایی در ورد نت و غیره)، رابطه میان توصیف‌گرهای[3] دو نامیده (هم‌معنایی، تضاد، شمول و ... ) و فاصله بین دو نامیده است. این سامانه بر روی داده ACE آزمایش شده و f-measure آن براساس معیار B3 در حدود ۷۸٪ است.

سامانه ریکانسیل[4] یک سکوی[5] پیمانه‌ای[6] است که برای طراحی سامانه‌های مرجع‌گزینی ساخته شده است. طراحی کلی این سامانه در شکل (۱) نشان داده شده است. هر یک از ماژول‌ها می‌توانند با ماژول دیگری (که همان کارکرد را دارد) جایگزین شوند. این سامانه بر روی داده‌های آزمون مختلف آزمایش شده است، دقت آن براساس معیار B3 برروی پیکره ACE05 معادل ۷۳ درصد است [26].




**(شکل-۱): طراحی کلی سامانه ریکانسیل [24]**
**(Figure-1): General design of recouncil system**

سامانه BART برای زبان انگلیسی طراحی شده است اما نوع طراحی ماژولار آن استفاده برای هر زبانی را ممکن می‌کند. این ابزار شامل پنج بخش اصلی پایپ‌لاین پیش‌پردازش، تولید نامیده‌ها، ماژول استخراج ویژگی، دیکودر و انکودر است. شکل (۲) شمای کلی این سامانه را نمایش می‌دهد. این سامانه بر روی داده‌های ACE-2 آزمایش شده و بهترین نتیجه آن f-measure معادل ۷۹ درصد است [27]. در سال‌های بعد سامانه بارت تغییراتی برای زبان‌های عربی و چینی ایجاد شده است. بهترین نتیجه آن برای زبان چینی و عربی به‌ترتیب f-measure معادل ۷۳ و ۶۷ درصد است [26].




**(شکل-۲): شمای کلی سامانه BART [26]**
**(figure-2): Outline of the BART system**

در مقاله [30] روش پیشنهادی استفاده از شبکه‌های عصبی بازگشتی[7] (RNN) است که اطلاعات را به‌طورمستقیم از نامیده‌ها استخراج کنند. این روش بالاخص در مورد نامیده‌های ضمیری کارایی دارد. با وجود اینکه سامانه‌های مرجع‌گزینی که از اطلاعات غیر محلی استفاده می‌کنند می‌توانند از بسیاری از خطاها جلوگیری کنند؛ اما درعمل ثابت شده است که سامانه‌هایی که از ویژگی‌های به‌طورکامل محلی استفاده می‌کنند، نتایج بهتری داشته‌اند. در این مقاله ادعا می‌شود که استفاده از ویژگی‌های جهانی، نتیجه بهتری خواهد داد؛ اما استخراج این ویژگی‌ها از کلاسترها کاری پیچیده است که تأثیر منفی بر روی نتایج دارد؛ لذا در این مقاله ویژگی‌های خوشه‌های نامیده‌ها[8] با استفاده از شبکه‌های عصبی بازگشتی یاد گرفته می‌شود.

مقاله [5] ترکیبی از روش قاعده‌مند و یادگیری ماشین است. مقاله مسأله مرجع‌گزینی را بر روی زبان انگلیسی و عربی و چینی انجام داده است. تشخیص نامیده‌ها در دو مرحله انجام می‌شود. نخست با استفاده از مکاشفه‌ها و موجودیت‌های نامدار، نامیده‌ها با بازخوانی[9] بالا استخراج

[1] node
[2] pair-wise
[3] modifier
[4] recouncil
[5] platform
[6] modular

[7] recurrent
[8] mention cluster
[9] recall

می‌شوند؛ سپس با هرس‌کردن[1] (براساس مکاشفه) نامیده‌های اضافی دور ریخته می‌شوند و دقت بالا می‌رود. در مقاله برای هر زبان قوانین مکاشفه‌ای خاصی مورد استفاده قرار گرفته است.

در مقاله [25] به بررسی تأثیر نقش روابط درخت تجزیه وابستگی در مرجع‌گزینی پرداخته می‌شود. این مقاله این کار را برای سه زبان عربی، چینی و انگلیسی انجام داده است. به‌طور سنتی معماری کلی یک سامانه مرجع‌یابی شامل چهار بخش تشخیص نامیده، استخراج ویژگی، تولید جفت‌ها و پس‌پردازش است. این سامانه نیز از همین معماری پیروی می‌کند. در این مقاله سعی شده است از روابط درخت تجزیه وابستگی برای تشخیص هم‌مرجع‌بودن یک جفت نامیده استفاده شود؛ لذا باید به‌دنبال روش‌های بهینه برای تبدیل سازه‌ها به درخت وابستگی بود. در این مقاله برای تبدیل به درخت وابستگی در زبان انگلیسی از LTH، در عربی از CATiB و در زبان چینی از Penn2Malt استفاده شده است. دقت B3 این سامانه f-measure معادل ۷۲٪ و براساس معیار CoNLL در حدود ۶۸ درصد است.

### ۲-۲-۲- مرجع‌گزینی در زبان فارسی

مسأله مرجع‌گزینی در زبان فارسی نیز در چند سال اخیر مورد توجه قرار گرفته ولی بیشتر تمرکز بر روی مسأله پیداکردن مرجع ضمیر بوده است. در ادامه به بررسی پژوهش‌های انجام‌شده در این زمینه در زبان فارسی پرداخته می‌شود.

در [18] از روش‌های یادگیری ماشین برای پیداکردن مرجع ضمیر استفاده شده است. این پژوهش با استفاده از پیکره PCAC-2008 انجام گرفته است. این پیکره مجموعه‌ای است شامل ۳۱ متن برگرفته از پیکره بی‌جن‌خان که در آن نزدیک‌ترین مرجع اسمی ۲۰۷۹ ضمیر مشخص شده است. در این روش از ۲۵ ویژگی استفاده شده است که به سه مجموعه تقسیم شده‌اند. مجموعه ویژگی‌های مرتبط با ضمیر، مجموعه ویژگی‌های مرتبط با عبارت اسمی و مجموعه ویژگی‌هایی که روابط بین این دو را نشان می‌دهند. برای یادگیری از چهار الگوریتم بیشینه آنتروپی، درخت تصمیم، ماشین بردار پشتیبان و شبکه عصبی چندلایه استفاده شده است. بهترین دقت مربوط به استفاده از روش درخت تصمیم (C4.5) بوده که معیار F1-measure آن عدد ۴۴.۷٪ است.

در پژوهشی دیگر که در زبان فارسی انجام شده، مجموعه‌ای از قوانین برای پیداکردن مرجع ضمیر ارائه شده است. این سامانه از دو قسمت پیش‌پردازش و مرجع‌گزینی تشکیل شده است. در گام پیش‌پردازش عملیات تعیین مرز واژگان، توکن‌بندی، ریشه‌یابی و برچسب‌زنی اجزای کلام[2] انجام و سپس در گام مرجع‌گزینی با استفاده از مجموعه‌ای از قوانین که تا سه جمله قبل از جمله جاری را جستجو می‌کنند، مرجع ضمیر پیدا می‌شود [11].

پیکره استفاده‌شده برای آزمایش این سامانه از پنج وبلاگ فارسی که به شکل اتفاقی از هر کدام بیست صفحه بارگیری‌شده ساخته شده است. برای مقایسه این روش مبتنی بر قانون با روش‌های دیگر از روش [18] استفاده شده است. نتیجه ارائه‌شده براساس معیار بازیابی و دقت است. این سامانه مبتنی بر قانون دقت ۹۵٪ و بازیابی ۹۰٪ را گزارش کرده است. و نتایج برای سامانه مبتنی بر یادگیری ماشینی - که برای مقایسه استفاده شده است- دقت ۹۲٪ و بازیابی ۸۷.۷٪ گزارش شده است.

در [1] سامانه‌ای مبتنی بر قوانین برای مرجع‌گزینی پیشنهاد شده است. این سامانه از مجموعه‌ای از قوانین خاص زبان فارسی و مجموعه‌ای از ویژگی‌ها برای تشخیص گروه‌های ارجاعی استفاده می‌کند. نتیجه این سامانه مرجع‌گزینی که منحصراً از روش مبتنی بر قانون استفاده می‌کند با استفاده از معیار ارائه‌شده در رقابت CoNLL بر روی بخش آزمون از پیکره درخت وابستگی اوپسالا سنجیده شده است. نامیده‌ها در این ارزیابی به‌صورت gold در نظر گرفته شده‌اند و نتیجه f-measure با معیار CoNLL عدد ۴۸.۳٪ است. همچنین درصورتی‌که سامانه به‌صورت end-to-end در نظر گرفته شود، نتیجه f-measure برمبنای معیار CoNLL ۹.۳۸٪ است. علت این افت شدید، ضعف در قسمت ابزارهای زیرساختی پردازش زبان طبیعی فارسی است که برای تشخیص نامیده‌ها به کار رفته‌اند.

در [20] پیکره لوتوس[3] معرفی شده است. این پیکره مجموعه‌ای از پنجاه متن به‌نسبه بلند برگفته از پیکره بی‌جن‌خان است که عبارات اسمی هم‌مرجع در آن مشخص شده است. در این پژوهش برمبنای پیکره لوتوس و با استفاده از سه الگوریتم درخت تصمیم، ماشین بردار پشتیبان و شبکه عصبی چندلایه سامانه مرجع‌گزینی تولید شده و بر روی ۲۰٪ از داده پیکره لوتوس -که به‌عنوان داده آزمون انتخاب شده است- ارزیابی شده است. بهترین نتایج مربوط به شبکه عصبی بوده که مقدار f-measure آن ۳۹.۴٪ گزارش شده است.

[1] prunning

[2] Part Of Speech Tagging

[3] Lotus

## ۳-۲- مروری بر روش‌های مرجع‌گزینی بر اساس مقایسه مؤلفه‌های رایج مورد استفاده

### ۱-۳-۲- پیش‌پردازش

در کارهای فارسی انجام‌شده به‌طور معمول اطلاعات دقیقی در مورد پیش‌پردازش و ابزارهای استفاده‌شده وجود ندارد. در [12] تنها به ابزار برچسب‌گذار اجزای کلام اشاره شده که از [24] استفاده شده است و در [1] از ابزارهای پیش‌پردازشی تولیدشده در پژوهشگاه خواجه نصیر (بخشی از ابزارهای معرفی‌شده در مقاله حاضر) استفاده شده است. ابزارهای غیرفارسی به‌طور معمول از ابزارهای OpenNLP و ابزارهای دانشگاه استنفورد و برکلی استفاده کرده‌اند. در این مقاله از ابزارهای پیش‌پردازشی تولیدشده در پژوهشگاه خواجه نصیرالدین طوسی استفاده شده که دقت هر یک از آنها در جدول (۴) آورده شده است. این ابزارها از کیفیت و دقت بسیار بالایی برخوردار هستند و در زمره بهترین ابزارهای پیش‌پردازشی زبان فارسی هستند.

### ۲-۳-۲- تشخیص نامیده

در مقالات فارسی تنها به مرجع‌یابی برای ضمایر پرداخته شده است و تشخیص ضمایر با استفاده از برچسب‌های اجزای کلام (برچسب‌های دستی یا ماشینی) صورت می‌گیرد به استثنای مقاله [1] که در آن گروه‌های اسمی با استفاده از تجزیه‌گر وابستگی استخراج می‌شوند و گروه‌هایی که قابلیت هم‌مرجع بودن را ندارند، طبق قوانینی حذف می‌شوند. در سامانه‌های مرجع‌گزینی غیرفارسی تشخیص نامیده‌ها مشابه مقاله حاضر با استخراج ضمایر (براساس برچسب اجزای کلام و یا فهرستی از ضمایر) استخراج موجودیت‌های نامدار به‌جز موجودیت‌هایی که قابلیت هم‌مرجع‌بودن ندارند، مانند اعداد و استخراج گروه‌های اسمی با استفاده از تجزیه‌گر نحوی و یا تجزیه‌گر وابستگی صورت می‌گیرد. در این مقاله بر خلاف بیش‌تر سامانه‌های مرجع‌گزینی فارسی از ضمایر (براساس برچسب اجزای کلام) و موجودیت‌های نامدار (با استفاده از ابزار تشخیص موجودیت نامدار پژوهشگاه خواجه نصیرالدین طوسی) و گروه‌های اسمی (با استفاده از روش‌های قطعه‌بندی مختلف) استفاده شده است.

### ۳-۳-۲- استخراج ویژگی ها

موسوی و قاسم‌ثانی [18] از ۲۵ ویژگی که در سه دسته ویژگی‌های معرف ضمیر – ویژگی‌های معرف هم مرجع قبلی – ویژگی‌های معرف رابطه بین ضمیر و هم‌مرجع قبلی دسته‌بندی می‌شوند؛ استفاده می‌کنند. در مقاله [1] نویسندگان از ویژگی‌های نقش هسته نامیده - جمع و مفرد - جانداری - نوع نامیده (ضمیر یا گروه اسمی) استفاده می‌کنند. در کارهای غیرفارسی مجموعه متنوعی از ویژگی‌ها یاد شده است. از این بین می‌توان به موارد نوع نامیده (اسمی، ضمیری، اسم خاص)،شباهت نامیده به سایر نامزدها، ویژگی‌های هسته معنایی نامیده، نوع واژه ابتدایی و انتهایی نامیده، واژگان بلافاصله قبل و بعد از نامیده، رابطه بین دو رشته، ویژگی‌های معنایی (جنس، شمار، رابطه معنایی در ورد نت و ...) و فاصله بین دو نامیده اشاره کرد. در مقاله حاضر تلاش شده است از همه ویژگی‌های یادشده در مقالات مختلف استفاده شود به‌جز مواردی که برای زبان فارسی قابل استفاده نیستند، مانند جنسیت ضمایر.

### ۴-۳-۲- دسته‌بندی

سامانه‌های مرجع‌گزینی به دو دسته قاعده‌مند و مبتنی بر یادگیری ماشینی تقسیم‌بندی می‌شوند. بیش‌تر سامانه‌های مرجع‌گزینی موجود چه در فارسی و چه در زبان‌های دیگر بر مبنای یادگیری ماشین عمل می‌کنند. بیشترین روش‌های یادگیری ماشین به کار رفته در سامانه‌های مختلف مرجع‌گزینی شامل ماشین بردار پشتیبان، درخت تصمیم، شبکه‌های عصبی چندلایه، آنتروپی بیشینه، گراف، شبکه‌های عصبی بازگشتی هستند که از این بین روش‌های ماشین بردار پشتیبان، درخت تصمیم و شبکه‌های عصبی چندلایه بیشترین استفاده را داشته و نتایج مطلوبی ارائه کرده‌اند. در مقاله حاضر نیز پس از بررسی روش‌های مختلف دسته‌بندی درنهایت از ماشین بردار پشتیبان که روش بسیار رایجی در زمینه مرجع‌گزینی است، استفاده شده است.

### ۵-۳-۲- تولید زنجیره خروجی

طراحی برخی از سامانه‌های مرجع‌گزینی به‌صورتی است که پس از مرحله دسته‌بندی نیاز به اعمال روش‌های خوشه‌بندی بوده است از جمله این روش‌ها می‌توان به روش خوشه‌بندی مبتنی بر آنتروپی بیشینه، single link – best first و most recent first اشاره کرد. در مقاله حاضر روش مورد استفاده بی‌نیاز از مرحله خوشه‌بندی بود.

### ۶-۳-۲- داده مورد استفاده

در سامانه‌های مرجع‌گزینی زبان فارسی از پیکره‌های کوچک مانند PCAC و Lotus و در سامانه‌های غیرفارسی از

پیکره‌های MUC, CoNLL و ACE استفاده شده است. پیکره تولیدشده و مورد استفاده در مقاله حاضر از لحاظ کمیت و انواع واژگان هم‌مرجع و تعدد برچسب‌های مربوط به مرجع‌گزینی از تمام پیکره‌های فارسی کامل‌تر و قابل مقایسه با پیکره‌های معروف این زمینه مانند ACE و Ontonotes است.

### ۷-۳-۲- روش ارزیابی

روش ارزیابی در مقالات فارسی به هیچ عنوان استاندارد نیست؛ درحالی‌که تمام مقالات غیرفارسی از معیارهای استاندارد MUC, B3, CEAF و CoNLL برای ارزیابی سامانه مرجع‌گزینی استفاده می‌کنند. مقاله حاضر نیز دقت سامانه مرجع‌گزینی را براساس معیارهای استاندارد یادشده، در بخش ۵ ارائه کرده است.

## ۳- پیکره واژگان هم‌مرجع و موجودیت‌های نامدار

پیکره واژگان هم‌مرجع شامل یک‌میلیون واژه و در قالب استاندارد CoNLL منتشر شده است. متون این پیکره از سایت‌های خبری در بازه ماه ۱۲ میلادی سال ۲۰۱۶ تهیه شده‌اند. دسته‌بندی موضوعی دادگان شامل موضوعات مختلف از جمله سیاسی، فرهنگی، اقتصادی، ورزشی و ... است. موضوعات مختلف پیکره در جدول (۳) ارائه شده است. این پیکره دارای دو دسته برچسب است.

- برچسب‌های دستی

این برچسب‌ها شامل برچسب هم‌مرجع‌بودن، نوع عبارت (ضمیر، موجودیت نامدار، گروه اسمی)، موجودیت نامدار (۷ برچسب شخص، مکان، سازمان، زمان، تاریخ، پول و درصد) و جاندار/بی‌جان بودن عبارت هستند.

- برچسب‌های خودکار

برچسب‌های خودکار شامل برچسب اجزای کلام[1]، ریشه واژگان، گروه‌های نحوی و ... است. این برچسب‌ها با استفاده از ابزارهای تولیدشده در پژوهشگاه خواجه نصیر به پیکره اضافه شده‌اند، لذا دقت این برچسب‌ها با دقت برچسب‌های دستی متفاوت خواهد بود. میزان توافق برچسب‌گذاری در بخش موجودیت‌های نامدار ۹۴٪ درصد براساس معیار کاپا است. برای محاسبه توافق برچسب‌گذاری در بخش برچسب‌های هم‌مرجعی از معیار MUC استفاده شده است. در تشخیص نامیده‌ها توافق ۸۷٪ و در تشخیص هم مرجع‌ها توافق ۸۵٪ است.

[1] POS

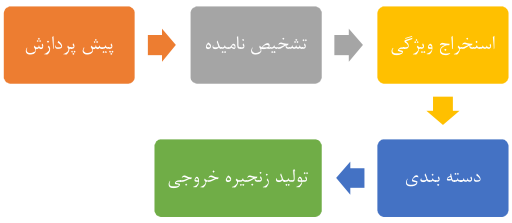
(جدول-۳): دسته‌بندی متون خام پیکره

(Table-3): topic categorization of raw corpus

| موضوع | تعداد واژه‌ها | تعداد | موضوع | تعداد واژه‌ها | تعداد |
|---|---|---|---|---|---|
| مذهبی | ۱۹۹۵۲ | ۲۱ | سیاسی | ۱۸۵۳۶۰ | ۲۴۸ |
| فناوری | ۲۳۰۰۴ | ۲۰ | فرهنگی | ۲۱۶۶۳۴ | ۱۹۵ |
| علمی | ۱۴۴۹۸ | ۱۱ | اقتصادی | ۱۳۲۸۷۰ | ۱۳۹ |
| آموزشی | ۱۲۱۸۷ | ۹ | ورزشی | ۴۸۸۵۰ | ۴۷ |
| اجتماعی | ۲۸۵۳۶۰ | ۸ | پزشکی | ۳۰۴۹۳ | ۳۰ |
| جمع | ۹۹۳۴۱۴ | ۷۵۳ | حوادث | ۲۴۲۱۶ | ۲۵ |

## ۴- سامانه مرجع‌گزینی

سامانه تشخیص عبارات هم‌مرجع یا به‌اختصار هم‌مرجع‌یاب، از پنج بخش اصلی تشکیل شده است. بخش نخست ماژول پیش‌پردازش است که در آن کارهای پیش‌پردازشی متن فارسی از قبیل نرمال‌سازی، تقطیع[2] و... انجام می‌شود. بخش دوم ماژول تشخیص نامیده[3] است. در این ماژول عبارت‌هایی که امکان قرارگیری در زنجیره‌های هم‌مرجعی را دارند، انتخاب می‌شوند. بخش سوم ماژول استخراج ویژگی است که در آن با استفاده از ویژگی‌های تولیدشده در ماژول نخست (با ابزارهای پیش‌پردازشی) بردار ویژگی تولید می‌شود. در بخش چهارم با استفاده از یک روش دسته‌بندی، داده‌های خروجی بخش چهارم دسته‌بندی و درنهایت در بخش پنجم داده‌های دسته‌بندی‌شده در بخش قبل با یکدیگر ترکیب می‌شوند و زنجیره‌های هم‌مرجعی را تشکیل می‌دهند. معماری کلی ابزار هم‌مرجع‌یاب، در شکل (۳) نمایش داده شده است. در ادامه این بخش به توضیح کامل‌تر هر یک از پنج بخش پرداخته خواهد شد.




(شکل-۳): معماری کلی ابزار هم‌مرجع‌یاب.

(Figure-3): general architecture of proposed coreference resolution system

[2] tokenization
[3] Mention detection

## ۱-۴- پیش‌پردازش

ورودی مرحله پیش‌پردازش متن خام و بدون برچسب است. در مرحله پیش‌پردازش عملیات‌های پیش‌پردازشی از قبیل: نرمال‌سازی، غلط‌یابی، تقطیع، جداسازی جملات، برچسب‌گذاری اجزای کلام، تشخیص ضمایر، قطعه‌بندی (تشخیص گروه‌های اسمی)، برچسب‌گذاری موجودیت‌های نامدار و... بر روی داده انجام می‌شوند. در شکل (۴)، مراحل مختلف پیش پردازش نمایش داده‌شده‌اند. هر یک از ابزارهای مورد استفاده در بخش پیش‌پردازش در ادامه توصیف شده‌اند.


**(شکل-۴): مراحل مختلف پیش‌پردازش**
**(Figure-4): preprocessing steps**

### ۱-۱-۴- نرمال‌سازی، غلط‌یابی، تقطیع و جداسازی جملات

این ابزار علاوه‌بر نرمال‌سازی، تقطیع، غلط‌یابی، تصحیح نقطه‌گذاری، تشخیص مرز جملات (شکستن متن به جملات) و ... را نیز انجام می‌دهد. این ماژول برای یکسان‌سازی نویسه‌ها در متن، تغییر کدگذاری[1]، حذف علائم، اشکال و اضافات متن، و به‌صورت کلی ایجاد یک متن تمیز و آماده برای اعمال پردازش‌های متنی تولید شده است. جمله براساس علائم نگارشی مانند نقطه، علامت تعجب و علامت سؤال مرزبندی می‌شود. همچنین قواعد مکاشفه‌ای و استثناهای مناسبی در نظر گرفته می‌شود که تقطیع جمله اشتباه نباشد.

### ۲-۱-۴- برچسب‌گذاری اجزای کلام

برچسب‌گذار اجزای کلام با استفاده از پیکره متنی زبان فارسی[2] و مدل مخفی مارکف تولید شده است. پیکره متنی زبان فارسی شامل بیش از هشت میلیون واژه است که پیش از این در پژوهشگاه خواجه‌نصیر تهیه شده است. در گروه پردازش متن، پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی، برچسب‌گذارهای اجزای کلام متعددی تولید شده‌اند که مجموعه برچسب آنها بسته به کاربرد مورد نظر با یکدیگر تفاوت دارد.

در مرحله پیش‌پردازش در مراحل مختلف از دو برچسب‌گذار با مجموعه برچسب شانزده و صدتایی استفاده شده است. مجموعه شانزده برچسبی شامل برچسب‌های درشت‌دانه موجود در پیکره متنی (برچسب‌های مرحله نخست) هستند. (اسم، فعل، صفت و ... ) است و برچسب‌های مجموعه صدتایی علاوه‌بر مقوله کلی هر واژه جزئیات بیشتری را در بر می‌گیرند. در این مجموعه برچسب تأکید بر تشخیص جزئیات فعل است. برچسب‌گذار قابلیت تشخیص زمان، نوع و شخص فعل را دارد. اطلاعات اضافی این برچسب‌گذار شامل: پی‌چسب‌های اسم و فعل، جزئیات زیاد فعل، نوع قید و صفت، جمع و مفردبودن و عام و خاص‌بودن اسم و ... است.

### ۳-۱-۴- تشخیص ضمیر

برای تشخیص ضمیر، از برچسب‌گذار اجزای کلام استفاده می‌شود و سپس با اعمال چند قانون بر روی خروجی آن ضمایر استخراج می‌شوند. ضمایر دو دسته هستند، دسته نخست ضمایر منفصل مانند : «خود، من، تو، او و ...» این ضمایر به هیچ واژه‌ای نچسبیده‌اند و خود یک واژه (یا توکن) مستقل محسوب می‌شوند. دسته دوم ضمایر متصل مانند «ش، م، ت، مان، تان، شان و ...» هستند. این ضمایر به‌عنوان واژه مستقل در متن وجود ندارند و به‌حتم به واژه دیگری چسبیده‌اند. نقش اصلی ماژول تشخیص ضمیر، جداکردن این واژگان است.

در مرحله نخست، تابع تشخیص ضمیر از برچسب‌گذار ۱۰۰ (اجزای کلام) استفاده می‌کند. همان‌طور که در بخش ۲-۱-۴ یاد شد، منظور برچسب‌گذار اجزای کلام آموزش‌دیده با مجموعه صد برچسبی است. پیکره اصلی دکتر بی‌جن‌خان دارای بیش از ششصد نوع برچسب است که به فراخور نیاز و با توجه به میزان دقت و جزئیات مورد نظر در کاربرد، چند مدل مختلف با تعداد برچسب مختلف (مثل ۱۶، ۳۲ و ۱۰۰) آموزش دیده‌اند. استخراج ضمایر دسته نخست ساده است. هر واژه‌ای که برچسب خروجی آن «PRO» باشد، ضمیر منفصل محسوب می‌شود؛ اما برای استخراج ضمایر دسته دوم در ابزار تشخیص ضمیر واژگانی که برچسب nclitic داشته باشند، به‌عنوان ضمایر متصل استخراج می‌شوند.

[1] encoding
[2] Bijankhan corpus

در پیکره متنی مواردی که در این پژوهش با نام ضمیر متصل شناخته می‌شوند (مان، تان، شان و ... ) به‌عنوان پی‌چسب[1] برچسب‌گذاری شده است. برای مثال واژه «مدادش» در پیکره متنی برچسب «اسم-مفرد-عام-پی‌چسب» دارد. برچسب‌گذار صد با استفاده از این برچسب پیکره متنی، طوری طراحی شده است که در چنین مواردی برچسب nclitic[2] به واژه اختصاص می‌دهد.

در این ماژول هر واژه‌ای که برچسب nclitic داشته باشد، انتخاب و سپس از انتهای آن ضمیر متصل جدا می‌شود و به‌عنوان یک توکن جدید در سطر بعدی قرار می‌گیرد. یعنی واژه «مدادش» به دو واژه «مداد» و «ش» تبدیل می‌شود. واژه نخست همان برچسب «اسم-مفرد-عام-پی‌چسب» را می‌گیرد و واژه دوم برچسب «ضمیر» می‌گیرد. دقت این ماژول ۹۲٪ است.

در پیکره متنی، شمار ضمایر مشخص نیست؛ لذا برچسب‌گذار صدبرچسبی، نمی‌تواند شمار ضمایر را مشخص کند؛ اما در این پژوهش برای افزایش دقت شمار ضمایر نیز مشخص می‌شود. این کار با استفاده از فهرست ضمایر جمع و مفرد انجام می‌گیرد.

### ۴-۱-۴- تشخیص جانداری[3]

برای تشخیص جانداری از برچسب‌گذار جانداری استفاده می‌شود. این برچسب‌گذار با استفاده از مدل مخفی مارکف بر روی پیکره دادگان تعلیم داده شده است. پیکره دادگان دارای برچسب جانداری برای واژگان خود است [22]. برای مثال عبارت «کتاب سبز علی» با استفاده از برچسب‌گذار جانداری به‌صورت «بی‌جان، بی‌جان، جاندار» برچسب‌گذاری می‌شود؛ اما با توجه به اینکه واحدهای مورد استفاده در این پژوهش قطعات[4] هستند، لازم است برای هر قطعه یک برچسب جانداری واحد وجود داشته باشد. جانداری کل یک قطعه براساس جانداری هسته آن قطعه تعیین می‌شود؛ لذا در عبارت «کتاب سبز علی» چون هسته «کتاب» است و بی‌جان است، کل عبارت بی‌جان در نظر گرفته می‌شود. دقت این ابزار در حدود ۸۰٪ است.

[1] clitic

[2] این برچسب به معنی N+clitic است. یعنی پی‌چسبی که به اسم چسبیده است.

[3] animacy

[4] chunk

### ۵-۱-۴- تشخیص هسته گروه اسمی

تشخیص هسته گروه اسمی[5] با استفاده از چند قانون صورت می‌گیرد. در زبان فارسی تعداد وابسته‌های پسین در گروه اسمی، بسیار بیشتر از وابسته‌های پیشین است. درواقع در بیش‌تر موارد به‌جز اعداد و ضمایر اشاره وابسته‌های اسم، پسین هستند. در این ماژول نخستین واژه‌ای که در گروه اسمی (از راست به چپ) برچسب اسم داشته باشد، هسته گروه اسمی محسوب می‌شود. بنابراین به‌دلیل سادگی قانون مورد استفاده عملکرد این ماژول به عملکرد POStagger و قطعه‌بند مورد استفاده وابسته است. (۹۴ و ۷۰ در معیار F)

### ۶-۱-۴- قطعه‌بندی گروه‌های اسمی

برای قطعه‌بندی گروه‌های اسمی از دو قطعه‌بند استفاده شده است. یکی از این دو با استفاده از روش یادگیری ماشین طراحی شده و دیگری به‌صورت قاعده‌مند و براساس تجزیه وابستگی تولید شده است. دقت این دو قطعه‌بند به‌ترتیب ۷۰ و ۷۶ است.

#### ۱-۶-۱-۴- قطعه‌بند مبتنی بر روش‌های یادگیری ماشین

این قطعه‌بند، با استفاده از پیکره گروه‌های نحوی استاندارد تولید شده است [15]. این پیکره در پژوهشگاه خواجه نصیر تولید شده و شامل بیش از ۲۵۰ هزار توکن است. تعداد برچسب‌های به‌کاررفته در پیکره هفت برچسب گروه اسمی است. این گروه‌ها شامل گروه اسمی، گروه فعلی، گروه حرف اضافه، گروه قیدی، گروه صفتی، گروه فهرست و گروه «را» است. در تولید این ابزار از روش مبتنی بر یادگیری میدان تصادفی شرطی[6] استفاده شده است. در تولید این ابزار از دو ویژگی اصلی برچسب اجزای کلام (شانزده برچسب) و واژه استفاده شده است [15].

براساس ابزار قطعه‌بند، قطعه‌بند گروه اسمی تولید شده است. در تولید این ابزار در داده آموزش تنها از گروه‌های اسمی استفاده شده است. و درنتیجه خروجی فقط گروه‌های اسمی را مشخص می‌کند. دقت این ماژول با معیار F در حدود هفتاد است.

[5] با توجه به اینکه در این پژوهش تمامی نامیده‌ها از جنس گروه اسمی (موجودات نامدار نیز گروه اسمی هستند) یا ضمیر هستند دو عبارت قطعه و گروه اسمی گاه به جای هم به کار برده می‌شوند.

[6] Conditional random fields

**۲-۶-۱-۴- قطعه‌بند مبتنی بر تجزیه‌گر وابستگی**

قطعه‌بند ب، بر پایه پارسر وابستگی کار می‌کند. از مالت پارسر[1] تعلیم‌یافته بر روی مجموعه دادگان با برچسب‌های استاندارد[2] برای تجزیه روابط وابستگی اجزای جمله استفاده می‌شود. داده ورودی پس از پیش‌پردازش‌های لازم وارد پارسر وابستگی شده و سپس از روابط وابستگی موجود و برچسب نقش‌های دستوری مانند فاعل، مفعول، مبتدا، برای استخراج قطعات مرتبط استفاده می‌شود. در این سامانه از نقش‌های دستوری (از قبیل فاعل، مفعول، گروه اسمی، گروه حرف اضافه، گروه قیدی و پیونددهنده‌های گفتمانی و غیره) استفاده می‌شود.

دقت شناسایی مرزهای آرگومانی درست با کاهش طول جمله افزایش می‌یابد. خطای حاصل از عدم شناسایی مرزهای آرگومانی دقیق، ناشی از وجود خطای پارسر وابستگی است؛ زیرا فرآیند شناسایی مرز آرگومان با استفاده از مدل دادگان در مالت پارسر انجام شده است و عملکرد مالت پارسر مطابق با تحقیقات آقای خلاش در پایان‌نامه کارشناسی ارشد ایشان با افزایش طول جمله، کاهش می‌یابد. پس از تشخیص قطعات موجود در متن، یک‌بار کل قطعات بررسی می‌شوند، اگر یک قطعه زیرمجموعه قطعه دیگری بود، قطعه کوچک‌تر حذف می‌شود. با این کار می‌توان جلوی خطاهای زیادی را گرفت. برای مثال اگر قطعه یک شامل abc باشد و قطعه دو شامل ab باشد، قطعه دوم حذف می‌شود.

**۷-۱-۴- تشخیص موجودیت‌های نامدار**

برچسب‌گذار موجودیت‌های نامدار موجود، با استفاده از پیکره اعلام تولید شده است [15]. دادگان مورد استفاده برای تولید پیکره موجودیت نامدار بیش از ۵۵۰ هزار توکن است. تعداد برچسب‌های موجودیت نامدار به‌کاررفته در پیکره سیزده موجودیت است. این موجودیت‌ها شامل: شخص، مکان، سازمان، رخداد، تاریخ، بازه، زمان، عدد اصلی، عدد ترتیبی، LNORP[3]، درصد، پول و اندازه[4] است. در تولید ابزار تشخیص موجودیت نامدار از روش مبتنی بر یادگیری میدان تصادفی شرطی استفاده شده است. ابزار تشخیص موجودیت نامدار با استفاده از دو ویژگی اصلی برچسب اجزای کلام (بیش از صد برچسب) و واژه و یک فهرست واژگان استفاده شده است. این فهرست واژگان شامل اسامی اشخاص (۵۲۰۰ اسم) و مکان‌ها (۱۸۰ کشور و شهر) است. دقت برچسب‌گذار اجزای کلام مورد استفاده ۹۴/۸ است. دقت ابزار تشخیص موجودیت نامدار براساس معیار f1 ۸۷/۱۶ است. در جدول (۴) دقت ابزار برچسب‌گذاری تشخیص موجودیت نامدار مشاهده می‌شود.

پس از اعمال ابزارهای یادشده بر روی داده ورودی خروجی به‌صورت فایل متنی تمیز[5] به‌همراه ویژگی‌ها به‌دست می‌آید. توضیحات مربوط به ابزارهای پیش‌پردازشی مورد استفاده در جدول (۴) مشاهده می‌شود.

**(جدول-۴):خلاصه توضیحات ابزارهای پیش پردازشی**
**(Table-4): Summary of preprocessing tools descriptions**

| ابزار | دقت/معیارF | روش |
|---|---|---|
| نرمال سازی، قطعه‌بندی[6]، جداسازی | بیش از ۹۰ | قاعده‌مند |
| برچسب‌گذار اجزای کلام | ۹۴ | آماری |
| قطعه‌بند گروه اسمی | ۷۰ | آماری و قاعده |
| تشخیص ضمیر | ۹۲ | قاعده مند |
| تشخیص موجودیت نامدار | ۸۸ | آماری |
| تشخیص جانداری | ۸۰ | قاعده مند |

## ۲-۴- تشخیص نامیده‌ها

در عملیات مرجع‌گزینی مجموعه‌ای از عبارات نامزد که در این جا به آنها «نامیده» می‌گوییم از درون متن انتخاب می‌شوند و باید برای هر کدام مرجع مناسب پیدا شود. این نامیده‌ها می‌توانند دارای سه نوع مختلف باشند. موجودیت‌های نامدار (مانند: Microsoft)، عبارت اسمی که هسته آن اسم عام است (مانند : corporation) یا ضمیر [13]. برای تشخیص اسامی خاص از ابزار تشخیص موجودیت‌های نامدار، برای تشخیص گروه‌های اسمی از قطعه‌بند عبارت و برای تشخیص ضمایر از ابزار تشخیص ضمیر استفاده شده است.

ترتیب انتخاب این نامیده ها می‌تواند به دو صورت باشد:

**انتخاب به‌ترتیب:** در این روش نخست ضمایر انتخاب، از متن حذف، سپس موجودیت‌های نامدار انتخاب و از متن حذف و در انتها گروه‌های اسمی از متن باقی‌مانده انتخاب می‌شوند. در این روش تعداد نامیده‌ها محدودتر از روش دوم است و پردازش سریع‌تر صورت می‌گیرد.

[1] Malt parser
[2] Gold
[3] مخفف چهار واژه ملیت، دین، گروه سیاسی و زبان
[4] اندازه‌گیری‌های استاندارد مانند: سن، مساحت، فاصله، انرژی، سرعت، دما، حجم، وزن و سایر عبارت و ساختارهای بیان‌گر اندازه
[5] منظور از فایل متنی تمیز، فایل بدون خطا و تقطیع شده است.
[6] tokenization

**انتخاب توأم:** در این روش هیچ یک از موارد انتخاب شده، از متن حذف نمی‌شوند. لذا ممکن است یک عبارت هم به‌عنوان موجودیت نامدار در نامیده‌ها حاضر شود و هم به‌عنوان گروه اسمی. همچنین ممکن است یک موجودیت نامدار بخشی از یک گروه اسمی باشد. در این روش تعداد نامیده‌ها بیشتر از روش نخست و پردازش کندتر است. بازخوانی این روش بیشتر از روش قبل ولی دقت آن کمتر است. همچنین در خروجی نهایی ممکن است عبارات استخراج‌شده با یکدیگر هم‌پوشانی داشته باشند.

مثال: تشخیص نامیده‌ها در جمله «مهاجم تیم فوتبال استقلال گفت: در بازی دیروز در ورزشگاه مملو از تماشاگر، مس تمام توانش را برای بردن ما گذاشت اما ما با وجود خستگی پیروز شدیم.» به‌ترتیب زیر انجام می‌شود.

حالت نخست: انتخاب به‌ترتیب:

۱- در مرحله نخست ضمایر انتخاب می‌شوند و از متن حذف می‌شوند.

ضمایر در این جمله شامل «ش – ما – ما» است. در پایان مرحله نخست مجموعه نامزدها و متن خام به‌صورت زیر است:

- مجموعه نامزدها: «ش – ما – ما»
- متن باقی‌مانده[1]: مهاجم تیم فوتبال استقلال گفت: در بازی دیروز در ورزشگاه مملو از تماشاگر، مس تمام توانش را برای بردن ما گذاشت اما ما پیروز شدیم.

۲- در مرحله دوم موجودیت‌های نامدار انتخاب و از متن حذف می‌شوند.

موجودیت‌های نامدار در این جمله شامل «استقلال - مس» است. در پایان مرحله دوم مجموعه نامزدها و متن خام به‌صورت زیر است:

- مجموعه نامزدها: «ش – ما – ما – استقلال - مس»
- متن باقی‌مانده : مهاجم تیم فوتبال ~~استقلال~~ گفت: در بازی دیروز در ورزشگاه مملو از تماشاگر، ~~مس~~ تمام توانش را برای بردن ما گذاشت اما ما پیروز شدیم.

۳- در مرحله سوم گروه‌های اسمی انتخاب و از متن حذف می‌شوند.

گروه‌های اسمی در این جمله شامل «مهاجم تیم فوتبال - بازی دیروز - ورزشگاه مملو از تماشاگر» است.

در پایان مرحله سوم مجموعه نامزدها و متن خام به‌صورت زیر است:

- مجموعه نامزدها: «ش – ما – ما – استقلال – مس – مهاجم تیم فوتبال – بازی دیروز – ورزشگاه مملو از تماشاگر – تمام توان - بردن»
- متن باقی‌مانده[2]: ~~مهاجم تیم فوتبال استقلال~~ گفت: در ~~بازی دیروز~~ در ~~ورزشگاه مملو از تماشاگر، مس~~ تمام ~~توانش~~ را برای ~~بردن~~ ~~ما~~ گذاشت اما ~~ما~~ پیروز شدیم.

در این مثال دیدیم که استخراج و حذف موجودیت‌های نامدار قبل از گروه‌های اسمی موجب شد که واژه استقلال قبل از انتخاب گروه اسمی «مهاجم تیم فوتبال استقلال» از متن حذف شود و گروه اسمی ناقص «مهاجم تیم فوتبال» را در متن باقی بگذارد؛ اما در حالت دوم تمام این سه دسته هم‌زمان استخراج می‌شوند و استخراج یکی از آنها مانع استخراج دیگری نمی‌شود.

- مجموعه نامیده‌ها شامل «ضمایر: {ش، ما، ما}موجودیت‌های نامدار: {استقلال، مس}- گروه‌های اسمی: { مهاجم تیم فوتبال استقلال، بازی دیروز، ورزشگاه مملو از تماشاگر، مس، تمام توانش ، بردن ما}» است.

## ۳-۴- استخراج ویژگی

برای تولید داده آموزش در این مرحله از میان نامیده‌های انتخاب‌شده در بخش قبلی نمونه‌های مثبت و منفی انتخاب می‌شوند. هر نمونه در بخش «داده آموزش» شامل دو نامیده، یک مجموعه ویژگی (که در ادامه یاد خواهد شد) و یک رده P یا N است. P و N به این معنی است که دو نامیده موجود در این نمونه با یکدیگر هم‌مرجع هستند یا خیر. تفاوت نمونه در بخش آموزش با بخش آزمون، این است که در نمونه‌های بخش آزمون ردهٔ مثبت یا منفی در نمونه وجود ندارد.

### ۱-۳-۴- نحوه انتخاب نمونه‌های مثبت و منفی

این انتخاب براساس مقاله مرجع [19] صورت گرفته است. برای انتخاب نمونه‌های مثبت، تنها دو نامیده هم‌مرجع که در زنجیره هم‌مرجعی با یکدیگر همسایه هستند، انتخاب می‌شوند. برای اینکه نمونه‌های منفی در داده آموزش خیلی زیاد نشوند، نمونه‌های منفی نیز شامل هر آنچه بین این دو نمونه است و نامیده دوم می‌شود. برای روشن‌شدن مطلب به

[1] واژگانی که قرمز شده‌اند و روی آنها خط کشیده شده است از متن حذف شده‌اند.

[2] واژگانی که روی آنها خط کشیده شده است از متن حذف شده‌اند.

مثال زیر توجه کنید. در ستون نخست عبارت‌ها و در ستون دوم شماره زنجیره هم‌مرجعی و در ستون سوم شماره جایگاه واژه در متن آمده است.

**(جدول-۵): مثال از نحوه انتخاب نمونه‌های مثبت و منفی**
**(Table-5): Example of how to select positive and negative samples**

| عبارت | زنجیره هم‌مرجعی | جایگاه متن |
|---|---|---|
| A | 1 | 1 |
| B | 2 | 2 |
| C | - | 3 |
| D | 1 | 4 |
| E | - | 5 |
| F | - | 6 |
| G | 2 | 7 |
| H | 1 | 8 |

این متن شامل دو زنجیره هم‌مرجعی است. یکی {A,D,H} و دیگری {B,G}. در زنجیره نخست A,D و H,D با یکدیگر همسایه محسوب می‌شوند و دو نمونه B,G در زنجیره دوم نیز با یکدیگر همسایه هستند. در چنین متنی نمونه‌های مثبت و منفی به‌صورت زیر انتخاب می‌شوند.

P: {A,D}, {H,D} , {G,B}
N:
{D,C},{D,B},{H,E},{H,F},{H,G},{G,C},{G,D},{G,E},{G,F}

به مثالی از زبان فارسی توجه کنید. در جمله «علی دیروز به مدرسه رفته بود. او در آنجا مدادش را گم کرده بود.» در زنجیره هم مرجعی واژگان علی با او و او با ش همسایه هستند لذا نمونه‌های مثبت و منفی به‌صورت زیر انتخاب می‌شوند.

نمونه‌های مثبت: {علی/او} ، {او/ش} ، {مدرسه/آنجا}
نمونه‌های منفی: {او/دیروز} ، {او/مدرسه}
{ش/آنجا} ، {ش/مداد}
{آنجا/او}

پس از انتخاب نمونه‌های مثبت و منفی لازم است، ویژگی‌های مورد استفاده استخراج شوند.

برای انتخاب ویژگی، ابتدا ویژگی‌های معرفی‌شده در مقالات و سامانه‌های موجود به‌صورت جامع بررسی شد؛ سپس از میان مجموعه موجود، انتخاب ویژگی هم بر اساس ادعای بهترین ویژگی‌های مشترک مطرح‌شده در مقالات (مقاله مرجع [19]) و هم با روش تجربی و از طریق گزینه Feature selection در ابزار وکا، متمایزکننده‌ترین ویژگی‌ها به‌عنوان مجموعه نهایی مورد استفاده قرار گرفتند؛ همچنین با توجه به ویژگی‌های خاص زبان فارسی، برخی ویژگی‌ها مانند جنسیت، هرس شدند. مجموعه ویژگی مورد استفاده در این مقاله به‌صورت یک مجموعه ویژگی مرجع مبنای بیش‌تر ابزارهای هم‌مرجع‌یابی است [19].

### ۲-۳-۴- ویژگی‌ها

- فاصله دو نامیده: فاصله بین جملاتی که دو نامیده در آنها قرار دارند.
- در یک جمله‌بودن دو نامیده: درصورتی‌که دو نامیده در یک جمله باشند، یک و در غیر این صورت صفر است.
- ضمیربودن نامیده: هر یک از نامیده‌ها که ضمیر باشد، این مقدار برایش یک است.
- مطابقت دو نامیده: درصورتی‌که دو نامیده به‌طور کامل مثل هم باشند (هر دو یک اسم یا یک ضمیر یا یک عبارت باشند). مثلا «کشور ایران» و «کشور ایران»
- مطابقت هسته دو نامیده: درصورتی‌که هسته دو نامیده یکی باشد، مقدار یک است. برای استخراج این ویژگی نیازمند ماژولی برای استخراج هسته گروه اسمی هستیم. این ماژول براساس برچسب اجزای کلام و جایگاه واژه در عبارت، هسته را استخراج می‌کند. مثال: «مجلس ایران» و «مجلس شورای اسلامی»
- مطابقت وابسته با هسته : درصورتی‌که هسته نامیده دوم در در وابسته‌های نامیده نخست باشد، این مقدار یک است. مثال: «وزیر کشور» و «کشور ایران»
- جانداری: در صورتی که هر دو جاندار باشند، مقدار یک است.
- شمار: براساس ویژگی برچسب اجزای کلام صد در صورتی که هسته هر دو گروه دارای شمار یکسان باشد، (هر دو جمع یا مفرد) ویژگی یک می‌شود.
- این+ گروه اسمی: در صورتیکه نامیده دوم با یکی از واژگان «این، همین، همان و... » شروع شود، این ویژگی یک می‌شود.
- بخشی از رشته: درصورتی‌که نامیده دوم زیر مجموعه‌ای از نامیده نخست باشد، این ویژگی یک می‌شود.
- موجودیت نامدار: در صورتیکه هر دو نامیده موجودیت نامدار باشند، این ویژگی یک می‌شود.
- نوع موجودیت نامدار: درصورتی‌که نوع موجودیت نامدار هر دو نامیده یکسان باشد، این ویژگی یک می‌شود.

## ۴-۴- دسته‌بندی

### ۴-۴-۱- انتخاب دسته‌بندی‌کننده مناسب

برای انتخاب دسته‌بندی‌کننده مناسب، از ابزار Weka استفاده شد. حدود شصت‌هزار توکن از دادگان انتخاب شدند. این دادگان به قالب ورودی ابزار Weka تبدیل و با استفاده از این ابزار بر روی دسته‌بندی‌کننده‌های مختلف آزمایش شدند.

دسته‌بندی‌کننده‌های ماشین بردار پشتیبان[1]، درخت تصمیم C4.5، درخت تصمیم J48، درخت تصادفی[2] و جنگل تصادفی[3]، از مواردی هستند که در پیشینه مطالعات مرجع‌یابی از آنها استفاده شده است و نتایج مناسب‌تری داشته‌اند. در جدول (۶) نتایج آزمون با استفاده از شصت‌هزار واژه با هر یک از دسته‌بندی‌کننده‌های یادشده مشاهده می‌شود. به جز ماشین بردار پشتیبان بقیه تست‌ها با Weka انجام شده است.

**(جدول-۶): نتایج آزمون بر روی دسته‌بندی‌کننده‌های مختلف در Weka**

**(Table-6): Test results on different categories in Weka**

| دسته‌بندی کننده | دقت | فراخوانی | F-measure |
|---|---|---|---|
| naiveBayes | 47 | 22 | 30 |
| J48 | 92 | 30 | 45 |
| RandomTree | 63 | 66 | 64 |
| RandomForest | 70 | 60 | 68 |
| SVM | 81 | 62 | 70 |

براساس نتایج به‌دست‌آمده دو دسته‌بندی‌کننده جنگل تصادفی و ماشین بردار پشتیبان نتایج بهتری نسبت به سایر دسته‌بندی‌کننده‌ها داشتند؛ لذا ادامه کار بر روی این دو دسته‌بندی‌کننده انجام گرفت. ماشین بردار پشتیبان به دو دلیل بر جنگل تصادفی ترجیح داده شد:

۱- برای ماشین بردار پشتیبان ابزارهای متن‌باز بیشتر، دقیق‌تر و کاراتری وجود دارند.

۲- دخالت‌دادن ویژگی واژگان در ماشین بردار پشتیبان ممکن بود اما در ابزارهای یافت‌شده برای جنگل تصادفی این امکان وجود نداشت.

همچنین از نظر منطقی، دلیل این‌که SVM (از میان طبقه‌بندهای پایه) به‌عنوان طبقه‌بندی‌کننده انتخاب شده است، می‌توان به‌اجمال به‌صورت زیر بیان کرد:

- اگر تعداد ویژگی‌ها زیاد باشد؛ چون SVM دارای حفاظت در برابر با بیش‌برازش[4] است، با کاهش ویژگی‌ها عملکرد SVMتحت تأثیر قرار نمی‌گیرد.
- به‌طور معمول فضای برداری متون تنک است. به‌صورت تئوری و تجربی اثبات می‌شود که الگوریتم‌های افزایشی مثل SVM که بایاس استنتاجی دارند، عملکرد خوبی با ماتریس‌های تنک دارند.
- نشان داده شده است که برای اغلب متون، جداسازی‌ها با کرنل خطی به‌صورت بهینه انجام می‌گیرد.

### ۴-۴-۲- دسته‌بندی‌کننده ماشین بردار پشتیبان

پس از استخراج نمونه‌های مثبت و منفی، یک دسته‌بندی‌کننده[5] با استفاده از این ویژگی‌ها آموزش داده می‌شود. دسته‌بندی‌کننده مورد نظر در این ابزار، ماشین بردار پشتیبان است که با استفاده از ابزار SVMLight و داده استخراج‌شده از بخش قبلی آموزش داده شده است. برای آموزش ماشین بردار پشتیبان یک‌بار از مجموعه ویژگی بالا استفاده شد و در آزمایش بعدی واژگان موجود در نامیده‌ها نیز به‌عنوان ویژگی به ماشین بردار پشتیبان داده شد؛ اما به‌دلیل محدودیت‌های SVMLight برای استفاده از ویژگی واژگان باید واژگان را با استفاده از اعداد نمایش بدهیم. برای این‌کار احتیاج به یک دادگان از واژگان فارسی داریم. حجم فعلی این واژگان سی‌هزار واژه است. با توجه به اینکه هر واژه جدیدی که در داده‌ها مشاهده شود، در فایل‌های آموزش وجود نداشته و درنتیجه در فرایند یادگیری تأثیر نداشته است، این واژه‌نامه تنها از مجموعه واژگان داده آموزش تشکیل شده است. درضمن با توجه به محدودیت‌های SVMLight از مجموعه واژگان به‌صورت کیسه واژگان[6] استفاده شده و ترتیب آنها در بردار ویژگی لحاظ نشده است.

## ۴-۵- تولید زنجیره خروجی

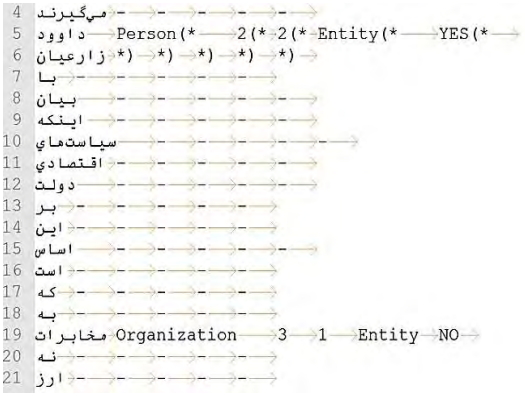
درنهایت خروجی‌های مرحله قبل براساس شماره زنجیره هم‌مرجعی، با یکدیگر ادغام و زنجیره‌های نهایی تولید می‌شوند. درواقع خروجی بخش قبل به‌صورت دوتایی‌هایی از نامیده‌هاست که مشخص می‌شود هم‌مرجع هستند یا خیر. با خوشه‌بندی خروجی، به زنجیره‌های هم‌مرجعی نهایی می‌رسیم.

## ۵- ارزیابی و نتایج

پیکره به‌صورت مجموعه‌ای از فایل‌های متنی تهیه شده است. نمونه‌ای از فایل پیکره برچسب‌گذاری‌شده در شکل (۵) نشان داده شده است. برای استفاده از پیکره از برنامه‌ای

[1] Support vector machine(SVM)
[2] Random Tree
[3] Random Forest
[4] overfitting
[5] classifier
[6] bag of words

که براساس الگوریتم توضیح داده شده در بخش ۴ تهیه شده است، می‌توان استفاده کرد. پیکره به‌گونه‌ای طراحی شده است که اضافه‌کردن برچسب جدید و استخراج اطلاعات از پیکره به‌آسانی صورت می‌گیرد. داده‌های پیکره و برنامه تهیه‌شده به‌ترتیب در بستر پارسیگان[1] و سایت گیتهاب[2] در دسترس عموم قرار دارند.




**(شکل-۵): نمونه‌ای از داده‌های برچسب‌خورده**
**(Figure-5): Example of labelled data**

در تولید مدل‌های دسته‌بندی‌کننده ماشین بردار پشتیبان از ۹۰٪ داده کل پیکره استفاده شده است. و ۱۰٪ از داده نیز به‌عنوان دادهٔ آزمون مورد استفاده قرار گرفته است. داده آموزش نهصدهزار واژه و داده آزمایش در حدود نودهزار واژه بوده است. در تولید مدل یک‌بار از ویژگی واژگان استفاده شده و یک‌بار این ویژگی استفاده نشده است. نتایج نشان می‌دهند که استفاده از ویژگی واژگان تأثیر مثبتی در بازیابی[3] دارد. در جدول (۷) نتایج را برای این دو مدل مشاهده می‌کنید.

**(جدول-۷): نتایج مدل SVM.**
**(Table-7): SVM test results**

| نام مدل | Precision | Recall | F1 |
|---|---|---|---|
| بدون استفاده از ویژگی واژگان | ۰/۹۹ | ۰/۳۸ | ۰/۵۵ |
| با استفاده از ویژگی واژگان | ۰/۸۰ | ۰/۶۴ | ۰/۷۱ |

در ارزیابی کلی سامانه مرجع‌گزینی ارزیابی با استفاده از ابزار ارزیابی استاندارد ارائه‌شده در مسابقات CoNLL صورت گرفته است. این ابزار نتایج را برای معیارهای MUC,B3,CEAF محاسبه می‌کند و براساس استاندارد ارائه‌شده میانگین این معیارها را به‌عنوان معیار CoNLL اعلام می‌کند[4]. ارزیابی بر روی داده آزمایش اوپسالا [23] انجام گرفته است. این پیکره در مجموع شامل ۶۱۴ جمله[5] و ۱۶۲۷۴ واژه (متوسط ۲۶/۵ واژه در هر جمله) است که به‌طور کامل قابل مقایسه با بخش آزمون و توسعه پیکره‌های MUC-6 (پیکره آزمون رقابت MUC-6 دارای ۱۳ هزار توکن است) و MUC-7 (پیکره توسعه رقابت MUC-7 دارای ۱۷ هزار توکن است) است.

این پیکره طبق شیوه‌نامهٔ مرجع‌گزینی پیکره واژگان هم‌مرجع، برچسب‌گذاری شده و دارای برچسب‌های دقیق است. پیکره اوپسالا به‌صورت متون پشت سر هم است. برای اینکه بتوان این داده را مورد استفاده قرار داد، این مجموعه به چهار قسمت تقسیم‌بندی شده است. اساس این تقسیم‌بندی این است که در هر فایل یک روایت وجود داشته باشد تا مرجع‌گزینی معنادار باشد. در جدول (۸) نتایج ارزیابی برروی شانزده‌هزار توکن پیکره اوپسالا مشاهده می‌شود.

**(جدول-۸): نتایج ارزیابی برروی شانزده‌هزار توکن پیکره اوپسالا**
**(Table-8): Evaluation results on 16,000 Uppsala corpus tokens**

| معیار | muc | b3 | ceafe | ceafm | conll |
|---|---|---|---|---|---|
| فایل ۱ | ۷۷/۲۷ | ۵۴/۹ | ۶۴/۵۶ | ۵۶/۷۸ | ۶۱/۶۲ |
| فایل ۲ | ۶۳/۳ | ۵۶/۸۱ | ۵۹/۱۷ | ۵۸/۷۸ | ۵۹/۵۱ |
| فایل ۳ | ۷۷/۰۱ | ۵۱/۱۶ | ۵۲/۶۴ | ۴۶/۰۷ | ۵۶/۷۲ |
| فایل ۴ | ۶۶/۴۲ | ۵۵/۹۳ | ۵۹/۸۸ | ۵۹/۳۶ | ۶۰/۳۹ |
| میانگین | ۶۹/۲۵ | ۵۴/۷ | ۵۹/۰۶ | ۵۵/۲۴ | **۵۹/۵۶** |
| k-fold | ۶۳/۴۲ | ۵۶/۵ | ۵۷/۳۱ | ۵۵/۹۱ | **۵۸/۲۸** |

گفتنی است، برای هر یک از معیارهای ارزیابی، نتایج به‌صورت دقت، فراخوانی و معیار f-measure ارائه می‌شوند. آنچه در جدول بالا آمده، مقدار f-measure برای هر معیار است. برای ارزیابی k-fold برای k=10 فولدها بر روی مجموعه آموزشی ایجادشده و با استفاده از ابزار Conll ارزیابی انجام شد.

## ۶- نتیجه‌گیری

در این مقاله روال تولید یک ابزار مرجع‌گزینی ارائه شده است. این ابزار با روش یادگیری ماشین و براساس پیکره نهصدهزار واژه‌ای واژگان هم‌مرجع تولید شده است. در تولید سامانه مرجع‌گزینی چندین ویژگی و ابزار مختلف به کار

[1] http://parsigan.ir/datasources/Coref/4
[2] https://github.com/rcdat-rcisp/RCDAT-Coreference
[3] recall

[4] Conll = (muc+b$^3$+ceafm+ceafe)/4
[5] لازم به ذکر است که ۱۴ جمله از بخش آموزش به بخش آزمون اضافه شده است تا متن یک‌پارچه باقی بماند.

رفته‌اند که دقت هر یک تأثیری بر دقت کل ابزار می‌گذارد. مهم‌ترین و اصلی‌ترین ابزاری که تاثیر بسیار زیادی بر نتیجه نهایی دارد قطعه‌بند است؛ زیرا قطعه‌بند هسته اصلی بخش تشخیص نامیده‌هاست. تشخیص نامیده‌ها مهم‌ترین بخش ابزار مرجع گزینی است؛ زیرا اگر نامیده‌ای در این بخش وارد روال نشود باعث کم‌شدن بازخوانی می‌شود و هیچ راهی برای بازیابی آن نامیده وجود ندارد. و این مسأله بر دقت نهایی خیلی تأثیرگذار خواهد بود؛ لذا تولید یک ابزار دقیق قطعه‌بند برای مسأله هم‌مرجع‌یابی ضروری به نظر می‌رسد. این یکی از مسائلی است که نتایج ابزار را بهبود خواهد داد. همچنین ابزار تشخیص ضمایر نیز باید دقیق باشد؛ زیرا تأثیری مشابه قطعه‌بند (البته کمتر) بر خروجی نامیده‌ها دارد. ابزار تشخیص جانداری نیز ابزار دیگری است که با افزایش دقت آن می‌توان به بهبود نتایج کمک کرد.

افزایش تعداد ویژگی‌ها و بالاخص ویژگی‌های معنایی به ابزار می‌تواند در بهبود نتایج تأثیرگذار باشد. در حال حاضر با توجه به منابع موجود در زبان فارسی ابزارهای نحوی و معنایی مناسبی وجود ندارند و این پژوهش از این دیدگاه در تنگنا است.

در زبان فارسی موارد زیر تنها دادگانی هستند که در زبان فارسی تولید شده‌اند:

- پیکره PCAC-2008 که است شامل ۳۱ متن برگرفته از پیکره بی‌جن‌خان که در آن نزدیک‌ترین مرجع اسمی ۲۰۷۹ ضمیر مشخص شده است [18].
- پیکره لوتوس که مجموعه‌ای از پنجاه متن به‌نسبه بلند برگفته از پیکره بی‌جن‌خان است که عبارات اسمی هم‌مرجع در آن مشخص شده است [20].

پیکره هم‌مرجع تولیدشده در این پژوهش از نظر حجمی بیش از پانصد برابر پیکره‌های قبلی زبان فارسی است و در عین حال به‌طور کامل قابل مقایسه با پیکره‌های مطرح ACE و Ontonotes است (ر.ک جدول ۱). همچنین از نظر نامیده‌های برچسب‌گذاری‌شده این پیکره سه نوع نامیده را مشخص می‌کند؛ درحالی‌که پیکره‌های دیگر زبان فارسی، تنها یک نوع نامیده را مشخص می‌کنند.

سامانه مرجع‌گزینی تولیدشده براساس معیار استاندارد CoNLL دارای f-measure نزدیک به شصت است. این در حالی است که سایر پژوهش‌های محدود انجام‌شده در فارسی دقت‌های متفاوتی از چهل تا نود درصد ارائه کرده‌اند که به‌دلیل این که سنجش دقت این پژوهش‌ها با معیارهای استاندارد رایج در مرجع‌گزینی صورت نگرفته است، این پژوهش‌ها با پژوهش حاضر قابل مقایسه نیستند. تنها پژوهش قابل مقایسه پژوهش طباطبایی و شکفته است که در آن سامانه‌ای مبتنی بر قوانین برای مرجع‌گزینی پیشنهاد شده است [1]. نتیجه این سامانه مرجع‌گزینی که منحصراً از روش مبتنی بر قانون استفاده می‌کند با استفاده از معیار ارائه‌شده در رقابت CoNLL بر روی بخش آزمون از پیکره درخت وابستگی اوپسالا سنجیده شده است. نامیده‌ها در این ارزیابی به‌صورت gold در نظر گرفته شده‌اند و نتیجه f-measure با معیار CoNLL عدد ۴۸.۳٪ است.در مقایسه بین این دو روش شاهد بهبود نتیجه در سامانه حاضر نسبت به کار ارائه‌شده در [1] هستیم.

## تقدیر و سپاس

در اینجا لازم است از مرکز تحقیقات مخابرات ایران برای حمایت از این پژوهش در قالب طرح جویش‌گر تقدیر و سپاس‌گزاری به‌عمل آوریم.

## ۷- مراجع 7- References

[۱]طباطبایی، ش. شکفته، ی. «سامانه پایه مرجع‌یابی گروه‌های اسمی در زبان فارسی با استفاده از قوانین ساده». اولین همایش جویشگر بومی، تهران، ۱۳۹۴.

[1] Sh. Tabatabaee and Y. Shekofteh, “The basic coreference resolution system for noun phrases in Persian language using simple rules”, The first conference on national search engine, Tehran 2015.

[2] B. Amit and B. Baldwin, "Algorithms for scoring coreference chains", The first international conference on language resources and evaluation workshop on linguistics coreference. Vol. 1. 1998 .

[3] B. Eric and D. Roth, "Understanding the value of features for coreference resolution," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008.

[4] B. Baldwin , M. Collins , J. Eisner , A. Ratnaparkhi , J. Rosenzweig and A. Sarkar, University of Pennsylvania: description of the University of Pennsylvania system used for MUC-6, Proceedings of the 6th conference on Message understanding, November 06-08, 1995, Columbia, Maryland.

[5] Ch. Chen and Ng.Vincent, "Combining the best of two worlds: A hybrid approach to multilingual coreference resolution," Joint Conference on

Methods in Natural Language Processing, Association for Computational Linguistics, 2005.

[18] N. S. Moosavi and Gh. Ghassem-Sani, "A Ranking Approach to Persian Pronoun Resolution," Advances in Computational Linguistics. Research in Computing Science 41, pp. 169-180, 2009.

[19] V. Ng, and C. Cardie, "Improving machine learning approaches to coreference resolution," In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 104-111, 2002.

[20] M. Nazaridoust, B. Minaei Bidgoli, S. Nazaridoust, "Co-reference Resolution in Farsi Corpora", *Advance Trends in Soft Computing Studies in Fuzziness and Soft Computing*, vol. 312, pp.155-162, 2014.

[21] S. Pradhan et al, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes," Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, 2012.

[22] M.S. Rasooli, M. Kouhestani, and A. Moloodi, "Development of a Persian Syntactic Dependency Treebank", In The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), Atlanta, USA.

[23] M. Seraji, B. Megyesi, J. Nivre , "Bootstrapping a Persian Dependency Treebank", Published as a Journal in Special Issue of the Linguistic Issues in Language Technology (LiLT), Heidelberg, Germany, 2012.

[24] M. Shamsfard, H. Fadaee, "A Hybrid Morphology-Based POS Tagger for Persian", In Proceedings of 6th Language Resources and Evaluation Conference (LREC 2008), Morocco, 2008.

[25] M. Stamborg, et a, "Using syntactic dependencies to solve coreferences," Joint Conference on EMNLP and CoNLL-Shared Task. Association for Computational Linguistics, 2012.

[26] V. Stoyanov, et al. "Reconciling ontonotes: Unrestricted coreference resolution in ontonotes with reconcile," Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Association for Computational Linguistics, 2011.

[27] O. Uryupina, M. Alessandro, and Massimo Poesio. "BART goes multilingual: The EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, 2012.

[6] N. Chinchor and S. Beth, "Message understanding conference (MUC) 6," LDC2003T13 (2003), 2013.

[7] N. Chinchor, "Message Understanding Conference (MUC) 7", LDC2001T02. Web Download. Philadelphia: Linguistic Data Consortium, 2001.

[8] C. Jacob , "A coefficient of agreement for nominal scales", Educational and Psychological Measurement, vol. 20, 1960, pp.37–46.

[9] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel , "The automatic content extraction (ace) program-tasks, data, and evaluatio"n"., In LREC, vol. 2, pp. 1, 2004.

[10] D. Greg, D. Leo Wright Hall and D. Klein, "Decentralized Entity-Level Modeling for Coreference Resolution," ACL (1), 2013.

[11] F. Fallahi and M. Shamsfard, "Recognizing anaphora reference in Persian sentences," Int. J. Comput. Sci, vol. 8, pp. 324-329, pp. 2011.

[12] A.M. Green,"Kappa statistics for multiple raters using categorical classifications", In Proceedings of the Twenty, 1997.

[13] A. Haghighi 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 : Association for Computational Linguistics*, Vol. 3, pp. 1152-1161, 2009.

[14] A. Haghighi and K. Dan, "Unsupervised coreference resolution in a nonparametric bayesian model," Annual meeting-Association for Computational Linguistics. vol. 45. No. 1. 2007.

[15] Sh. Hosseinnejad, Y. Shekofteh, & T. Emami Azadi, "A'laam Corpus: A Standard Corpus of Named Entity for Persian Language", *Signal and Data Processing*, vol.14, pp.127-142, 2017.

[16] H. Lee, et al, "Joint entity and event coreference resolution across documents," Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012.

[17] X. Luo, "On coreference resolution performance metrics," Proceedings of the conference on Human Language Technology and Empirical

UniTN/Essex submission to the CoNLL-2012 shared task," Joint Conference on EMNLP and CoNLL-Shared Task, Association for Computational Linguistics, 2012.

[28] Y. Versley, et al, "BART: A modular toolkit for coreference resolution," Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session. Association for Computational Linguistics, 2008.

[29] M. Vilain, et al, "A model-theoretic coreference scoring scheme," Proceedings of the 6th conference on Message understanding, Association for Computational Linguistics, 1995.

[30] S. Wiseman, A. M. Rush and S. M. Shieber, "Learning Global Features for Coreference Resolution," arXiv preprint arXiv:1604.03035, 2016.

[31] A. salimibadr and M.Homayounpour, Phrase chunking in Persian texts . JSDP, vol. 10 (2), pp. 69-86,2014.

**زینب رحیمی** در سال ۱۳۶۶ در شاهرود متولد شد. تحصیلات تا مقطع دیپلم را در شهر شاهرود سپری و دیپلم متوسطه خود را در سال ۱۳۸۳ دریافت کرد. وی تحصیلات خود را در مقطع کارشناسی در رشته مهندسی کامپیوتر (نرم‌افزار) در دانشگاه صنعتی امیرکبیر (۱۳۸۹) و کارشناسی ارشد را در رشته مهندسی فناوری اطلاعات (سامانه‌های چندرسانه‌ای) در همان دانشگاه (۱۳۹۱) به پایان رساند. ایشان هم‌اکنون دانشجوی مقطع دکترا در رشته مهندسی کامپیوتر (هوش مصنوعی) در دانشگاه شهید بهشتی تهران هستند. از موضوعات مورد علاقه ایشان می‌توان به پردازش زبان طبیعی، ترجمه ماشینی و هستان‌شناسی اشاره کرد.

نشانی رایانامه ایشان عبارت است از:

**rahimi-z@rcdat.ir**

**شادی حسین نژاد** فارغ‌التحصیل رشته مهندسی کامپیوتر از دانشگاه تهران در سال ۱۳۹۱ است. همچنین مدرک کارشناسی ارشد خود را در سال ۱۳۹۳ از دانشگاه صنعتی شریف در رشته زبان‌شناسی رایانشی دریافت کرده است. ایشان از سال ۱۳۹۲ تا ۱۳۹۶ در گروه صوت و پردازش متن پژوهشگاه خواجه نصیرالدین طوسی مشغول به فعالیت بوده است. زمینه پژوهشی مورد علاقه ایشان پردازش زبان طبیعی، زبان‌شناسی رایانشی، آهنگ و آواشناسی است.

نشانی رایانامه ایشان عبارت است از:

**shadi.hn@gmail.com**