

هم مرجع یابی 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

مرجع یابی 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 به شمار می‌رود. طبق تعاریف زمانی، دو واژه زمانی هم‌مرجع هستند که هر دو به موجودیت واحدی در متن یا جهان حقیقی ارجاع بدھند. تاکنون برای حل این مسئله 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 برچسب‌گذاری شده 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 به همراه 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 دقیق 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

*نویسنده عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۴۰۷/۲/۱۹

• تاریخ پذیرش: ۱۴۰۸/۳/۱۱

• تاریخ انتشار: ۱۴۰۹/۰۴/۰۱

• نوع مطالعه: پژوهشی

• سال ۱۴۰۹ شماره ۱ پیاپی ۴۳

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.

چند برچسب غیردقيق^۲ به پيکره اضافه شده است که اين برچسبها خروجي ابزارهاي پيشپردازشی تهييه شده در پژوهشگاه توسعه فناوري هاي پيشرفته خواجه نصيرالدين طوسى هستند. برای تهييه سامانه مرجع يابي از روش يادگيري ماشين مبتنى بر پيکره استفاده شده است.

معماري مورد استفاده، معماری جديدي نيست که برای نخستين بار در اين مقاله پيشنهاد شده باشد؛ اما در اين پژوهش در وهله نخست منبع ارزشمند دادگان فارسي برای مسئله مرجع گزيني تهييه شده است (هدف اصلی) و دوم اين که برای هر يك از مازوّل هاي يادشده در معماری، تلاش شده تا با استفاده از ابزارهاي موجود و روشها و الگوريتمهاي موجود بهترین انطباق با دادگان انجام و بهترین ترتيب ممکن ايجاد شود. بدین ترتيب قابل استفاده بودن پيکره اثبات شده است. همچنين اين سامانه خودکار انتهاهاتها^۳ (واقعي و در دسترس) با کاريри شناسايي عبارات هم مرجع (ضمير و عبارات ارجاعي) برای زبان فارسي مطابق اطلاع نگارنده جديد و نواورانه است.

در اين مقاله که علاوه بر مقدمه در شش بخش تهييه شده است، به معرفی پيکره واژگان هم مرجع ايجاد شده و الگوريتم مورد استفاده در ابزار هم مرجع ياب (مرجع گزيني) پرداخته می شود. پس از مقدمه، در بخش دوم اين مقاله پيکره هاي واژگان هم مرجع در زبان فارسي و ساير زبانها

^۲ برچسب غيردقيق خروجي ابزارهاي پردازشی است و توسط ماشين توليد شده است.

^۳ end-to-end

۱- مقدمه

مسئله مرجع گزیني و يا به عبارت ديگر يافتن کليه عباراتي که به يك موجوديت دلالت دارند، يك بخش عملياتي مهم در مسائلی مانند خلاصه سازي خودکار، پرسش و پاسخ خودکار و استخراج اطلاعات به شمار می رود. تاکنون روش هاي متنوعی برای حل مسئله مرجع گزيني پيشنهاد شده است که به دو دسته تقسيم می شوند: «روش هاي مبتنى بر قانون و روش هاي يادگيري ماشين».

روش هاي مبتنى بر قانون مجموعه ای از قوانين مکاشفه ای را در نظر می گيرند که از اطلاعات نحوی و معنایی واژگان برای پیدا کردن مرجع ضمیر استفاده می کنند. در طرف ديگر روش هاي يادگيري ماشين قرار دارند که در آنها از يك پيکره برچسب گذاري شده - که در آن مراجع عبارت ها مشخص شده است - استفاده می شود تا يك الگوريتم يادگيري بتواند به شكل خودکار قوانين مرجع گزيني را بيموزد. بزرگ ترين مشکل اين روش ها برای زبان هايی مانند زبان فارسي کمبود داده برچسب گذاري شده است.

در اين مقاله به معرفی يك پيکره واژگان هم مرجع برای زبان فارسي و سامانه مرجع گزيني متناظر آن پرداخته شده است. پيکره عنوان شده شامل حدود يك ميليون واژه فارسي است که برچسب دستي^۱ (دقیق) هم مرجع و موجودیت نامدار دارند. همچنان علاوه بر برچسب هاي دقیق،

^۱ برچسب دقیق برچسبی است که توسط انسان تولید شده است.



(جدول-۱): مقایسه پیکره‌های واژگان هم مرجع
 (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]	Arabic Chinese	960 K 500 K 615 K
Ontonotes [20]	English Arabic Chinese	1.6 M 300 K 950 K

۱-۱-۲- پیکره‌های مرجع گزینی در زبان فارسی
 مرجع گزینی در زبان فارسی موضوع پژوهشی جدیدی است که تنها در چند سال اخیر به آن توجه شده است. بنا به اطلاع نویسنده‌گان، پیش از این به طور سازمان یافته و جامع تلاشی برای تولید ابزار یا داده مرجع گزینی در زبان فارسی صورت نگرفته؛ لذا در زبان فارسی دادگان برای پیکره هم مرجع به صورت بسیار محدود تولید شده است. موارد زیر تنها دادگانی هستند که در زبان فارسی تولید شده‌اند:

۱-۱-۲-۱- پیکره^۱ PCAC-2008^۲

این پیکره مجموعه‌ای است شامل ۳۱ متن برگرفته از پیکره بی جن‌خان 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 استفاده شده برای آزمایش سامانه [11] از پنج وبلاگ فارسی که به شکل اتفاقی از هر کدام بیست صفحه بارگیری شده ساخته شده است [10].

۱-۱-۲-۲- پیکره لوتوس^۲

پیکره لوتوس، مجموعه‌ای از پنجاه متن به نسبه بلند برگفته از پیکره بی جن‌خان 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

معرفی و همچنین سامانه‌های مرجع گزینی معتبر در سطح جهانی و سامانه‌های مرجع گزینی 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 مرجع گزینی اختصاص دارد و در بخش ششم جمع‌بندی مطالب ارائه می‌شود و کارهای پیش‌رو بررسی خواهد شد.

۲- مروری بر کارهای انجام شده

۱- مروری بر پیکره‌های تولید شده

از دهه ۹۰ میلادی بیشتر رویکردها در تولید سامانه‌های مرجع گزینی، به سمت روش‌های یادگیری ماشین تمايل پیدا کرده است. در این راستا رقابت‌های مختلفی مانند رقابت‌های MUC [6,7] و CoNLL با موضوع مرجع گزینی 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، به صورت 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.

برای محاسبه کاپا آورده شده است. محاسبه فرمول کاپا به صورت ذی است:

$$\Pr(q) = P_{11} + P_{22} \quad (1)$$

$$\Pr(e) = P_1 * P1 + P_2 * P2 \quad (Y)$$

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (3)$$

۳-۱-۲- معیارهای ارزیابی سامانه‌های مرجع گزینی

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

MUC - معيار ١-٣-٢

این روش ارزیابی در سال ۱۹۹۵ میلادی ارائه شد [۲۹]. در این معیار key یا کلید، خوشه‌هایی از عبارات هم‌مرجع هستند که توسط انسان مشخص شده‌اند و response یا پاسخ، خوشه‌هایی هستند که سامانه مرجع گزینی تشخیص داده است. مقدار بازیابی به صورت نسبت تعداد پیوندهای مشترک در کلید و پاسخ بر تعداد پیوندهای کلید محاسبه می‌شود و دقت به صورت نسبت تعداد پیوندهای مشترک در کلید و پاسخ بر تعداد پیوندهای پاسخ محاسبه می‌شود.

B cube ، معيار ٢-١-٣-٢

میان معيار در سال ۱۹۹۸ ارائه شد [۲]. هدف از ارائه این معيار رفع بحرخی از مشکلاتی است که در معيار ارزیابی وجود داشت. یکی از مشکلات معيار MUC این است که تفاوتی بین میزان بدی اشتباها قائل نیست. در معيار B3 تلاش شده که این حس شهودی در نظر گرفته شود. در این معيار به جای این که به پیوندهای بین عبارات نگاه شود، به خود عبارات و حضور و یا عدم حضور آنها در یک رد 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.

وازگان مناسب وجود دارد. همچنین لازم است، پیکره تولیدشده دارای برچسب برای موارد بیشتر واژگان هم مرجع علاوه بر 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 به صورت 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^۱ [۸] است.

این معیار، معیاری آماری است که در زبان‌شناسی پیکره‌ای برای برسی میزان تفاهم میان دو برجسب‌گذاری بر روی پیکره به کار می‌رود. در این معیار تلاش شده است که برجسب‌گذاری‌هایی که به‌طور اتفاقی شبیه به یکدیگر شده‌اند تأثیری در نتیجه نداشته باشند. به این منظور کاپا

به صورت فرمول (۱) تا (۳) محاسبه می شود [12].

در این فرمول (a) میزان تشابه صحیح میان دو برچسب‌گذاری است و (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 بهزادی 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 مشبت و منفی در حس، گذاری شده است.

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

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

گروه مثبت و منفی در جدول بالا به عنوان مثال از نام دو برچسب (رده) فرضی (برای بررسی تعداد مواردی که برچسب‌گذارها مثل هم برچسب یا اتفاق نظر نداشته‌اند)

1 Kappa

۲-۲-۲- مروري بر روش‌های مرجع‌گزینی

۲-۲-۱- روش‌های مطرح شده 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 اضافه کردن قانون‌های جدید سعی در بالابردن بازیابی دارد، نخستین بار توسط برک و کالینز [4] ارائه شد [16]. او پیشنهاد داده بود که از هفت قانون به عنوان فیلترهایی برای استخراج روابط هم‌مرجعی استفاده شود. ایده استفاده از قوانین بعدتر توسط تعداد زیادی از پژوهش‌گران به کار گرفته شد [14, 19]. اما تا قبل از [13] هیچ کدام از سامانه‌های مرجع‌گزینی مبتنی بر قانون موفق به رسیدن به بازدهی‌های گزارش شده توسط سامانه‌های مبتنی بر یادگیری ماشین نشده؛ اما در رویکرد ارائه شده توسط حقیقی در [13] برای نخستین بار 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 ارائه کردند که از کلیه سامانه‌های ارائه شده تاکنون پاسخ بهتری داشت. قوانین استفاده شده در این سامانه سلسله‌مراتبی به نسبه ساده و مستقل از زبان انگلیسی هستند و به دلیل همین خصوصیت اقبال زیادی به آن شد.

در روش مورد استفاده در سامانه تولید شده 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 قبلی⁴ 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، واژه ابتدایی و انتهایی نامیده، واژگان بلافصله قبل و بعد از نامیده و فاصله بین نامیده‌ها

³ mention
⁴ antecedent

$$\text{Precision}_i = \frac{\text{number of correct elements in the output chain containing entity}_i}{\text{number of elements in the output chain containing entity}_i} \quad (4)$$

$$\text{Recall}_i = \frac{\text{number of correct elements in the output chain containing entity}_i}{\text{number of elements in the truth chain containing entity}_i} \quad (5)$$

رابطه‌های (4) و (5) 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.

۲-۳-۱- معيار CEAFF

معیار CEAFF¹ 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 CEAFF تلاش شده است که رابطه بهینه یک به یکی میان عبارات پاسخ و کلید پیدا شود. به دلیل این که این رابطه یک‌یکی است، در این معیار از تمام عبارات موجود در مجموعه پاسخ و کلید استفاده نمی‌شود.

در این معیار $S_{min}(d)$ و $K_{min}(d)$ به صورت 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 (برهمنه² میان دومجموعه) به صورت رابطه (6) است:

$$g(K_i) = S_j, K_i \in K_{min}(d) \text{ and } S_j \in S_{min}(d). \quad (6)$$

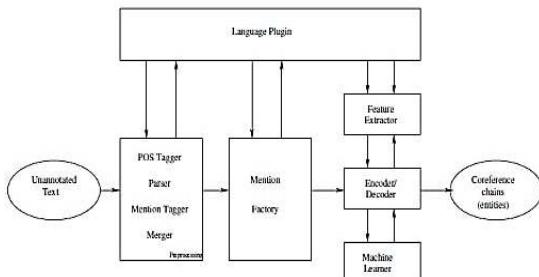
۲-۳-۲- معيار CoNLL

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

¹ Constrained Entity – Alignment F-measure

² alignment

سامانه BART برای زبان انگلیسی طراحی شده است اما نوع طراحی مژو لار آن استفاده برای هر زبانی را ممکن می کند. این ابزار شامل پنج بخش اصلی پایپ لاین پیش پردازش، تولید نامیده ها، مژول استخراج ویژگی، دیکودر و انکودر است. شکل (۲) شماتیکی این سامانه را نمایش می دهد. این سامانه بر روی داده های ACE-2 آزمایش شده و بهترین نتیجه آن f-measure معادل ۷۹ درصد است [۲۷]. در سال های بعد سامانه بارت تغییراتی برای زبان های عربی و چینی ایجاد شده است. بهترین نتیجه آن برای زبان چینی و عربی به ترتیب f-measure معادل ۷۳ و ۶۷ درصد است [۲۶].



[۲۶]: شماتیکی سامانه BART
(figure-2): Outline of the BART system

در مقاله [۳۰] روش پیشنهادی استفاده از شبکه های عصبی بازگشتی^۷ (RNN) است که اطلاعات را به طور مستقیم از نامیده ها استخراج کنند. این روش بالا خص در مورد نامیده های ضمیری کارایی دارد. با وجود اینکه سامانه های مرجع گزینی که از اطلاعات غیر محلی استفاده می کنند می توانند از بسیاری از خطاهای جلوگیری کنند، اما در عمل ثابت شده است که سامانه هایی که از ویژگی های به طور کامل محلی استفاده می کنند، نتایج بهتری داشته اند. در این مقاله ادعا می شود که استفاده از ویژگی های جهانی، نتیجه بهتری خواهد داد، اما استخراج این ویژگی ها از کلاس ها کاری پیچیده است که تأثیر منفی بر روی نتایج دارد؛ لذا در این مقاله ویژگی های خوش های نامیده ها^۸ با استفاده از شبکه های عصبی بازگشتی یاد گرفته می شود.

مقاله [۵] ترکیبی از روش قاعده مند و یادگیری ماشین است. مقاله مسئله مرجع گزینی را بر روی زبان انگلیسی و عربی و چینی انجام داده است. تشخیص نامیده ها در دو مرحله انجام می شود. نخست با استفاده از مکاشفه ها و موجودیت های نامدار، نامیده ها با بازخوانی^۹ بالا استخراج

⁷ recurrent

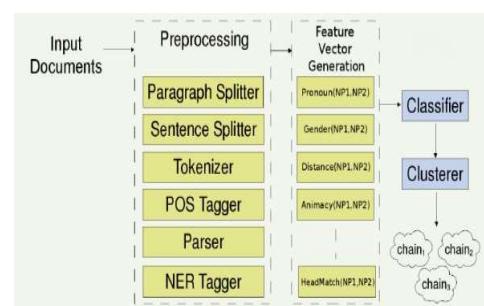
⁸ mention cluster

⁹ recall

(تعداد جمله ها و تعداد نامیده ها بین دو نامیده) استفاده شده است. این سامانه بر روی داده CoNLL 2011 معيار MUC سنجیده شده که بهترین نتیجه آن ۶۵٪ براساس معيار MUC بوده است. در سامانه دانشگاه ايلی نویز و مقاله منتشر شده مربوط به آن نشان داده شده است که استفاده از یک مدل ساده با کمک مجموعه ویژگی های قوی می تواند به نتایج خوبی در زمینه مرجع گزینی منجر شود. این رویکرد مسئله مرجع گزینی را به صورت مسئله گراف مدل کرده است. هر گره^۱ گراف و شاید متن هم جوار آن یک نامیده در نظر گرفته شده است. در صورتی که دو نامیده با یکدیگر هم مرجع باشند یک پیوند در گراف بین آن دو بقرار می شود. مدل هم مرجع در این مقاله براساس جفت^۲ است، یعنی هر نامیده با نامیده های قبلی خود مقایسه می شود تا مشخص شود کدامیک، هم مرجع آن است [۳].

ویژگی های استفاده شده در این سامانه شامل نوع نامیده (اسم خاص، ضمیر یا گروه اسمی)، رابطه بین دو رشته، ویژگی های معنایی (جنس، شمار، رابطه معنایی در ورد نت و غیره)، رابطه میان توصیف گرهای^۳ دو نامیده (هم معنایی، تضاد، شمول و ...) و فاصله بین دو نامیده است. این سامانه بر روی داده ACE آزمایش شده و f-measure و براساس معيار B3 در حدود ۷۸٪ است.

سامانه ریکانسیل^۴ یک سکوی^۵ پیمانه های^۶ است که برای طراحی سامانه های مرجع گزینی ساخته شده است. طراحی کلی این سامانه در شکل (۱) نشان داده شده است. هر یک از مژو ل ها می توانند با مژو ل دیگری (که همان کار کرد را دارد) جایگزین شوند. این سامانه بر روی داده های آزمون مختلف آزمایش شده است، دقت آن براساس معيار B3 بروی پیکره ACE05 معادل ۷۳ درصد است [۲۶].



[۲۶]: طراحی کلی سامانه ریکانسیل
(Figure-1): General design of recouncil system

¹ node

² pair-wise

³ modifier

⁴ recouncil

⁵ platform

⁶ modular

در پژوهشی دیگر که در زبان فارسی انجام شده، مجموعه‌های 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 برجسبزی اجزای کلام² 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 استفاده شده برای آزمایش این سامانه از پنج وبلاگ فارسی که به شکل اتفاقی از هر کدام بیست صفحه پارگیری شده ساخته شده است. برای مقایسه این روش مبتنی بر قانون با روش‌های دیگر از روش [18] استفاده شده است. نتیجه ارائه شده 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 ارائه شده در رقابت CoNLL بر روی بخش آزمون از پیکره درخت وابستگی اوپسالا سنجیده شده است. نامیده‌ها در این ارزیابی به صورت gold در نظر گرفته شده‌اند و نتیجه f-measure CoNLL عدد end-to-end ۴۸.۳٪ است. همچنین در صورتی که سامانه به صورت end-to-end در نظر گرفته شود، نتیجه f-measure بر مبنای 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 لوتوس³ معروفی شده است. این پیکره مجموعه‌ای از پنجاه متن به نسبه بلند برگفته از پیکره بی جن‌خان است که عبارات اسمی هم‌مرجع در آن مشخص شده است. در این پژوهش بر مبنای پیکره لوتوس و با استفاده از سه الگوریتم درخت تضمیم، ماشین بردار پشتیبان و شبکه عصبی چندلایه سامانه مرجع‌گزینی تولید شده و بر روی ۲۰٪ از داده پیکره لوتوس - که به عنوان 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.

² Part Of Speech Tagging

³ Lotus

می‌شوند؛ سپس با هرسکردن⁴ (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 مرجع یابی 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 هم‌مرجع بودن یک جفت نامیده استفاده شود؛ لذا باید بینال 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 ۷۲٪ و براساس معیار CoNLL در حدود ۶۸ درصد است.

۲-۲-۲- مرجع‌گزینی در زبان فارسی

مسئله مرجع‌گزینی در زبان فارسی نیز در چند سال اخیر مورد توجه قرار گرفته ولی بیشتر تمرکز بر روی مسئله پیداکردن مرجع ضمیر بوده است. در ادامه به بررسی پژوهش‌های انجام شده 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 بی جن‌خان 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.

⁴ pruning



۲-۳-۳- استخراج ویژگی ها

موسوی و قاسمیانی [۱۸] از ۲۵ ویژگی که در سه دسته ویژگی های معرف ضمیر - ویژگی های معرف هم مرجع قبلی

۲-۳- مروری بر روش های مرجع گزینی بر اساس

مقایسه مؤلفه های رایج مورد استفاده

۲-۳-۱- پیش پردازش

در کارهای فارسی انجام شده به طور معمول اطلاعات دقیقی در مورد پیش پردازش و ابزارهای استفاده شده وجود ندارد. در [۱۲] تنها به ابزار برچسب گذار اجزای کلام اشاره شده که از [۲۴] استفاده شده است و در [۱] از ابزارهای پیش پردازشی تولید شده در پژوهشگاه خواجه نصیر (بخشی از ابزارهای معروفی شده در مقاله حاضر) استفاده شده است. ابزارهای غیرفارسی به طور معمول از ابزارهای OpenNLP و ابزارهای دانشگاه استنفورد و برکلی استفاده کردند. در این مقاله از ابزارهای پیش پردازشی تولید شده در پژوهشگاه خواجه نصیر الدین 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 زمرة بهترین ابزارهای پیش پردازشی زبان فارسی هستند.

۲-۳-۲- تشخیص نامیده

در مقالات فارسی تنها به مرجع یابی برای ضمایر پرداخته شده است و تشخیص ضمایر با استفاده از برچسب های اجزای کلام (برچسب های دستی یا ماشینی) صورت می گیرد به استثنای مقاله [۱] که در آن گروه های اسمی با استفاده از تجزیه گر وابستگی استخراج می شوند و گروه هایی که قابلیت هم مرجع بودن را ندارند، طبق قوانینی حذف می شوند. در سامانه های مرجع گزینی غیرفارسی تشخیص نامیده ها مشابه مقاله حاضر با استخراج ضمایر (براساس برچسب اجزای کلام و یا فهرستی از ضمایر) استخراج موجودیت های نامدار به جز موجودیت هایی که قابلیت هم مرجع بودن ندارند، مانند اعداد و استخراج گروه های اسمی با استفاده از تجزیه گر نحوی و یا تجزیه گر وابستگی صورت می گیرد. در این مقاله برخلاف بیشتر سامانه های مرجع گزینی فارسی از ضمایر (براساس برچسب اجزای کلام) و موجودیت های نامدار (با استفاده از ابزار تشخیص موجودیت نامدار پژوهشگاه خواجه نصیر الدین طوسی) و گروه های اسمی (با استفاده از روش های قطعه بندی مختلف) استفاده شده است.

۲-۳-۴- دسته بندی

سامانه های مرجع گزینی به دو دسته قاعده مند و مبتنی بر یادگیری ماشینی تقسیم بندی می شوند. بیشتر سامانه های مرجع گزینی موجود چه در فارسی و چه در زبان های دیگر بر مبنای یادگیری ماشین عمل می کنند. بیشترین روش های یادگیری ماشین به کار رفته در سامانه های مختلف مرجع گزینی شامل ماشین بردار پشتیبان، درخت تصمیم، شبکه های عصبی چند لایه، آنتروپی بیشینه، گراف، شبکه های عصبی بازگشته هستند که از این بین روش های ماشین بردار پشتیبان، درخت تصمیم و شبکه های عصبی چند لایه بیشترین استفاده را داشته و نتایج مطلوبی ارائه کرده اند. در مقاله حاضر نیز پس از بررسی روش های مختلف دسته بندی درنهایت از ماشین بردار پشتیبان که روش بسیار رایجی در زمینه مرجع گزینی است، استفاده شده است.

۲-۳-۵- تولید زنجیره خروجی

طراحی برخی از سامانه های مرجع گزینی به صورتی است که پس از مرحله دسته بندی نیاز به اعمال روش های خوش بندی بوده است از جمله این روش ها می توان به روش single link – best و most recent first اشاره کرد. در مقاله حاضر روش first و most recent first مورد استفاده بی نیاز از مرحله خوش بندی بود.

۲-۳-۶- داده مورد استفاده

در سامانه های مرجع گزینی زبان فارسی از پیکره های کوچک مانند PCAC و Lotus و در سامانه های غیرفارسی از

- ویژگی های معرف رابطه بین ضمیر و هم مرجع قبلی دسته بندی می شوند؛ استفاده می کنند. در مقاله [۱] نویسنده گان از ویژگی های نقش هسته نامیده - جمع و مفرد - جانداری - نوع نامیده (ضمیر یا گروه اسمی) استفاده می کنند. در کارهای غیرفارسی مجموعه متنوعی از ویژگی ها یاد شده است. از این بین می توان به موارد نوع نامیده (اسمی، ضمیری، اسم خاص)، شبهات نامیده به سایر نامزدها، ویژگی های هسته معنایی نامیده، نوع واژه ابتدایی و انتهایی نامیده، واژگان بالا فاصله قبل و بعد از نامیده، رابطه بین دو رشته، ویژگی های معنایی (جنس، شمار، رابطه معنایی در ورد نت و ...) و فاصله بین دو نامیده اشاره کرد. در مقاله حاضر تلاش شده است از همه ویژگی های یادشده در مقالات مختلف استفاده شود به جز مواردی که برای زبان فارسی قابل استفاده نیستند، مانند جنسیت ضمایر.

برچسب‌های هم‌مرجعی از معیار MUC استفاده شده است. در تشخیص نامیده‌ها توافق ۸۷٪ و در تشخیص هم‌مرجع‌ها توافق ۸۵٪ است.

(جدول-۳): دسته‌بندی متون خام پیکره
(Table-3): topic categorization of raw corpus

تعداد	تعداد واژه‌ها	موضوع	تعداد	تعداد واژه‌ها	موضوع
۲۴۸	۱۸۵۳۶۰	سیاسی	۲۱	۱۹۹۵۲	مذهبی
۱۹۵	۲۱۶۶۲۴	فرهنگی	۲۰	۲۳۰۰۴	فنایری
۱۳۹	۱۳۲۸۷۰	اقتصادی	۱۱	۱۴۴۹۸	علمی
۴۷	۴۸۸۵۰	ورزشی	۹	۱۲۱۸۷	آموزشی
۳۰	۳۰۴۹۳	پژوهشی	۸	۲۸۵۲۶۰	اجتماعی
۲۵	۲۴۲۱۶	حوادث	۷۵۳	۹۹۳۴۱۴	جمع

۴- سامانه مرجع‌گزینی

سامانه تشخیص عبارات هم‌مرجع یا به اختصار هم‌مرجع‌یاب، از پنج بخش اصلی تشکیل شده است. بخش نخست مازول پیش‌پردازش است که در آن کارهای پیش‌پردازشی متن فارسی از قبیل نرمال‌سازی، تقطیع^۱ و... انجام می‌شود. بخش دوم مازول تشخیص نامیده^۲ است. در این مازول عبارت‌هایی که امکان قرارگیری در زنجیره‌های هم‌مرجعی را دارند، انتخاب می‌شوند. بخش سوم مازول استخراج ویژگی است که در آن با استفاده از ویژگی‌های تولیدشده در مازول 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 دسته‌بندی شده 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

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 Ontonotes و ACE است.

۲-۳-۷- روش ارزیابی

روش ارزیابی در مقالات فارسی به هیچ عنوان استاندارد نیست؛ در حالی که تمام مقالات غیرفارسی از معیارهای استاندارد CoNLL, MUC, B3, CEAF 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/ی جان بودن عبارت هستند.

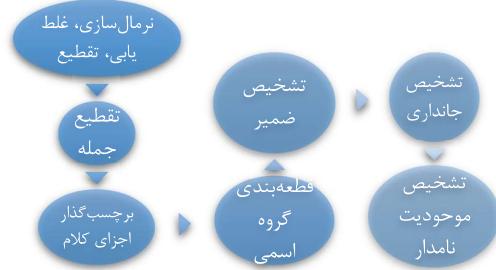
• برچسب‌های خودکار

برچسب‌های خودکار شامل برچسب اجزای کلام^۱، 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

¹ POS

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 پیش‌پردازش نمایش داده شده‌اند. هر یک از ابزارهای مورد استفاده در بخش پیش‌پردازش در ادامه توصیف شده‌اند.



(شکل-۴): مراحل مختلف پیش‌پردازش

(Figure-4): preprocesing steps

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

این ابزار علاوه‌بر نرمال‌سازی، تشخیص مرز جملات (شکستن متن به نقطه‌گذاری، تشخیص مرز جملات (شکستن متن به جملات) و ... را نیز انجام می‌دهد. این مازول برای یکسان‌سازی نویسه‌ها در متن، تغییر کدگذاری^۱، 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 سوال مرزبندی می‌شود. همچنین قواعد مکاشفه‌ای و استثناهای مناسبی در نظر گرفته می‌شود که تقطیع جمله اشتباہ نیاشد.

۱-۴-۲- برچسب‌گذاری اجزای کلام

برچسب‌گذار اجزای کلام با استفاده از پیکره متنی زبان فارسی^۲ 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

¹ encoding

² Bijankhan corpus

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 شائزده و صفتایی استفاده شده است. مجموعه شائزده برچسب‌های شامل برچسب‌های درشت‌دانه موجود در پیکره متنی (برچسب‌های مرحله نخست) هستند. (اسم، فعل، صفت و ...) است و برچسب‌های مجموعه صفتایی 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.

۱-۴-۱-۳- 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 به عنوان واژه مستقل در متن وجود ندارند و به عنوان واژه دیگری چسبیده‌اند. نقش اصلی مازول تشخیص ضمیر، جدا کردن این واژگان است.

در مرحله نخست، تابع تشخیص ضمیر از برچسب‌گذار ۱۰۰ (اجزای کلام) استفاده می‌کند. همان‌طور که در بخش ۱-۴-۲ یاد شد، منظور برچسب‌گذار اجزای کلام آموزش دیده با مجموعه صد برچسبی است. پیکره اصلی دکتر بی جن‌خان 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، به عنوان ضمایر متصل استخراج می‌شوند.

۴-۱-۵- تشخیص هسته گروه اسمی
تشخیص هسته گروه اسمی^۵ 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 به جز اعداد و ضمایر اشاره وابسته‌های اسم، پسین هستند. در این مژول نخستین واژه‌ای که در گروه اسمی (از راست به چپ) برچسب اسم داشته باشد، هسته گروه اسمی محسوب می‌شود. بنابراین بهدلیل سادگی قانون MORTH و POSStagger مورد استفاده عملکرد این مژول به عملکرد قطعه‌بند موردن است. (۹۴ و ۷۰ در معیار F)

۴-۱-۶- 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 به صورت قاعده‌مند و براساس تجزیه وابستگی تولید شده است. دقت این دو قطعه‌بند به ترتیب ۷۰ و ۷۶ است.

۴-۱-۶-۱- 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 به کاررفته 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^۶ 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 مژول با معیار F در حدود هفتاد است.

^۵ 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.

^۶ Conditional random fields

در پیکره متنی مواردی که در این پژوهش با نام ضمیر متصل شناخته می‌شوند (مان، تان، شان و ...) به عنوان پی‌چسب^۱ 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^۲ به واژه اختصاص می‌دهد.

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

۴-۱-۶-۲- تشخیص جانداری^۳

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 به صورت «بی‌جان، بی‌جان، جاندار» برچسب‌گذاری می‌شود؛ اما با توجه به اینکه واحدهای مورد استفاده در این پژوهش قطعات^۴ 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.

^۱ clitic

^۲ این برچسب به معنی N+clitic است. یعنی پی‌چسبی که به اسم چسبیده است.

^۳ animacy

^۴ chunk



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

قطعه‌بند ب، بر پایه پارسرا وابستگی کار می‌کند. از مالت پارسرا^۱ تعليمیافته بر روی مجموعه دادگان با برچسب‌های استاندارد^۲ برای تجزیه روابط وابستگی اجزای جمله استفاده می‌شود. داده ورودی پس از پیش‌پردازش‌های لازم وارد پارسرا وابستگی شده و سپس از روابط وابستگی موجود و برچسب نقش‌های دستوری (از قبیل فاعل، مفعول، گروه اسامی، 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 پارسرا وابستگی است؛ زیرا فرآیند شناسایی مرز آرگومان با استفاده از مدل دادگان در مالت پارسرا انجام شده است و عملکرد مالت پارسرا مطابق با تحقیقات آقای خلاش در پایان نامه 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.

۴-۱-۷- تشخیص موجودیت‌های نامدار

برچسب‌گذار موجودیت‌های نامدار موجود، با استفاده از پیکره اعلام تولید شده است [15]. دادگان مورد استفاده برای تولید پیکره موجودیت نامدار پیش از ۵۵ هزار توکن است. تعداد برچسب‌های موجودیت نامدار به کاررفته در پیکره سیزده موجودیت است. این موجودیت‌ها شامل: شخص، مکان، سازمان، رخداد، تاریخ، بازه، زمان، عدد اصلی، عدد ترتیبی، LNORP^۳، درصد، پول و اندازه^۴ است. در تولید ابزار تشخیص موجودیت نامدار از روش مبتنی بر یادگیری میدان تصادفی شرطی استفاده شده است. ابزار تشخیص

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 بیان گر اندازه

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^۵ بههمراه ویژگی‌ها بهدست می‌آید. توضیحات مربوط به ابزارهای پیش‌پردازشی مورد استفاده در جدول (۴) مشاهده می‌شود.

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

روش	f1/معیار	ابزار
قاعده‌مند	بیش از ۹۰	نرم‌السازی، قطعه‌بندی ^۶ ، جداسازی
آماری	۹۴	برچسب‌گذار اجزای کلام
آماری و قاعده	۷۰	قطعه‌بند گروه اسامی
قاعده مند	۹۲	تشخیص ضمیر
آماری	۸۸	تشخیص موجودیت نامدار
قاعده مند	۸۰	تشخیص جانداری

۴-۲- تشخیص نامیده‌ها

در عملیات مرجع‌گزینی مجموعه‌ای از عبارات نامزد که در اینجا 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 اسامی که هسته آن اسم عام است (مانند: corporation) یا ضمیر [13]. برای تشخیص اسامی خاص از ابزار تشخیص موجودیت‌های نامدار، برای تشخیص گروه‌های اسامی از قطعه‌بند عبارت و برای تشخیص ضمایر از ابزار تشخیص ضمیر استفاده شده است.

ترتیب انتخاب این نامیده‌ها می‌تواند به دو صورت باشد: انتخاب بهتر ترتیب: در این روش نخست ضمایر انتخاب، از متن حذف، سپس موجودیت‌های نامدار انتخاب و از متن حذف و در انتهای گروه‌های اسامی از متن باقی مانده انتخاب می‌شوند. در این روش تعداد نامیده‌ها محدودتر از روش دوم است و پردازش سریع‌تر صورت می‌گیرد.

منظور از فایل متنی تمیز، فایل بدون خطأ و تقطیع شده است.
⁶ tokenization

- در پایان مرحله سوم مجموعه نامزدها و متن خام به صورت زیر است:
- مجموعه نامزدها: «ش – ما – ما – استقلال – مس – مهاجم تیم فوتبال – بازی دیروز – ورزشگاه مملو از تماشاگر – تمام توان – بردن»
 - متن باقیمانده^۲: «مهاجم تیم فوتبال استقلال گفت: در بازی دیروز در ورزشگاه مملو از تماشاگر، مس تمام توانش را برای برقن 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 انتخاب شده 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 رده مثبت یا منفی در نمونه وجود ندارد.

۱- 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 روشن شدن مطلب به

^۲ واژگانی که روی آنها خط کشیده شده است از متن حذف شده‌اند.

انتخاب توأم: در این روش هیچ یک از موارد انتخاب شده، از متن حذف نمی‌شوند. لذا ممکن است یک عبارت هم به عنوان موجودیت نامدار در نامیده‌ها حاضر شود و هم به عنوان 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 کنترل است. بازخوانی این روش بیشتر از روش قبل ولی دقت آن کمتر است. همچنین در خروجی نهایی ممکن است عبارات استخراج شده 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.» به ترتیب زیر انجام می‌شود.

حالت نخست: انتخاب به ترتیب:

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 به صورت زیر است:

• مجموعه نامزدها: «ش – ما – ما»

• متن باقیمانده^۱: «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 به صورت زیر است:

• مجموعه نامزدها: «ش – 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.

^۱ واژگانی که قرمز شده‌اند و روی آنها خط کشیده شده است از متن حذف شده‌اند.



به عنوان مجموعه نهایی مورد استفاده قرار گرفتند؛ همچنین با توجه به ویژگی‌های خاص زبان فارسی، برخی ویژگی‌ها مانند جنسیت، هرس شدن. مجموعه ویژگی مورد استفاده در این مقاله به صورت یک مجموعه ویژگی مرجع مبنای بیشتر ابزارهای هم‌مرجع‌بایی است [19].

۴-۳-۲- ویژگی‌ها

- فاصله دو نامیده: فاصله بین جملاتی که دو نامیده در آنها قرار دارند.
- در یک جمله‌بودن دو نامیده: در صورتی که 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 برابش یک است.
- مطابقت دو نامیده: در صورتی که دو نامیده به‌طور کامل مثل هم باشند (هر دو یک اسم یا یک ضمیر یا یک عبارت باشند). مثلاً «کشور ایران» و «کشور ایران»
- مطابقت هسته دو نامیده: در صورتی که هسته دو نامیده یکی باشد، مقدار یک است. برای استخراج این ویژگی نیازمند مازولی برای استخراج هسته گروه اسامی هستیم. این مازول براساس برچسب اجزای کلام و جایگاه واژه در عبارت، هسته را استخراج می‌کند. مثال: «مجلس ایران» و «مجلس شورای اسلامی»
- مطابقت واپسنه با هسته: در صورتی که هسته نامیده دوم در در واپسنه‌های 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 اسامی: در صورتیکه نامیده دوم با یکی از واژگان «این، همین، همان و...» شروع شود، این ویژگی یک می‌شود.
- بخشی از رشتة: در صورتی که 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: در صورتی که نوع موجودیت نامدار هر دو نامیده یکسان باشد، این ویژگی یک می‌شود.

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

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

این متن شامل دو زنجیره هم‌مرجعی است. یکی H,D,{A,D,H} و دیگری {B,G}. در زنجیره نخست A,D و B,G با یکدیگر همسایه محسوب می‌شوند و دو نمونه 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} ، {آماده آنجا} نمونه‌های منفی: {او دیروز} ، {آماده ش} {ش آنجا} ، {ش آماده} {آنچا او}

پس از انتخاب نمونه‌های مثبت و منفی لازم است، ویژگی‌های مورد استفاده استخراج شوند. برای انتخاب ویژگی، ابتدا ویژگی‌های معرفی شده در مقالات و سامانه‌های موجود به‌صورت جامع بررسی شد؛ سپس از میان مجموعه موجود، انتخاب ویژگی هم بر اساس ادعای بهترین ویژگی‌های مشترک مطرح شده در مقالات (مقاله مرجع [19]) و هم با روش تجربی و از طریق گزینه Feature selection در ابزار وکا، متمایز‌کننده‌ترین ویژگی‌ها





- بهطور معمول فضای برداری متون تنک است. بهصورت تئوری و تجربی اثبات می شود که الگوریتم های افزایشی مثل SVM که بایاس استنتاجی دارند، عملکرد خوبی با ماتریس های تنک دارند.
- نشان داده شده است که برای اغلب متون، جداسازی ها با کرنل خطی بهصورت بهینه انجام می گیرد.

۴-۴-۲- دسته‌بندی‌کننده ماشین بردار پشتیبان

پس از استخراج نمونه های مثبت و منفی، یک دسته‌بندی‌کننده^۵ با استفاده از این ویژگی ها آموزش داده می شود. دسته‌بندی‌کننده مورد نظر در این ابزار، ماشین بردار پشتیبان است که با استفاده از ابزار SVMLight و داده استخراج شده از بخش قابلی آموزش داده شده است. برای آموزش ماشین بردار پشتیبان یکبار از مجموعه ویژگی بالا استفاده شد و در آزمایش بعدی واژگان موجود در نامیده ها نیز به عنوان ویژگی به ماشین بردار پشتیبان داده شد؛ اما به دلیل محدودیت های SVMLight برای استفاده از ویژگی واژگان باید واژگان را با استفاده از اعداد نمایش بدهیم. برای این کار احتیاج به یک دادگان از واژگان فارسی داریم. حجم فعلی این واژگان سی هزار واژه است. با توجه به اینکه هر واژه جدیدی که در داده ها مشاهده شود، در فایل های آموزش وجود نداشته و درنتیجه در فرایند یادگیری تأثیر نداشته است، این واژه نامه تنها از مجموعه واژگان داده آموزش تشکیل شده است. در ضمن با توجه به محدودیت های SVMLight از مجموعه واژگان بهصورت کیسه واژگان^۶ استفاده شده و ترتیب آنها در بردار ویژگی لحاظ نشده است.

۴-۵- تولید زنجیره خروجی

درنهایت خروجی های مرحله قبل براساس شماره زنجیره هم مرجعی، با یکدیگر ادغام و زنجیره های نهایی تولید می شوند. درواقع خروجی بخش قبل بهصورت دوتایی هایی از نامیده هاست که مشخص می شود هم مرجع هستند یا خیر. با خوشبندی خروجی، به زنجیره های هم مرجعی نهایی می رسیم.^۷

۵- ارزیابی و نتایج

پیکره بهصورت مجموعه ای از فایل های متند تهیه شده است. نمونه های از فایل پیکره برچسب گذاری شده در شکل (۵) نشان داده شده است. برای استفاده از پیکره از برنامه ای

⁵ classifier

⁶ bag of words

۴-۴- دسته‌بندی

۴-۴-۱- انتخاب دسته‌بندی‌کننده مناسب

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

دسته‌بندی‌کننده های ماشین بردار پشتیبان^۸، درخت تصمیم C4.5، درخت تصمیم J48، درخت تصادفی^۹ و جنگل تصادفی^{۱۰}، از مواردی هستند که در پیشینه مطالعات مرجع یابی از آنها استفاده شده است و نتایج مناسب تری داشته اند. در جدول (۶) نتایج آزمون با استفاده از شصت هزار واژه با هر یک از دسته‌بندی‌کننده های یادشده مشاهده می شود. به جز ماشین بردار پشتیبان بقیه تست ها با Weka انجام شده است.

(جدول-۶): نتایج آزمون بر روی دسته‌بندی‌کننده های

مختلف در Weka

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

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

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

۱- برای ماشین بردار پشتیبان ابزارهای متن باز بیشتر، دقیق تر و کارآثری وجود دارند.

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

همچنین از نظر منطقی، دلیل این که SVM (از میان طبقه‌بندهای پایه) به عنوان طبقه‌بندی‌کننده انتخاب شده است، می توان به اجمال به صورت زیر بیان کرد:

- اگر تعداد ویژگی ها زیاد باشد، چون SVM دارای حفاظت در برابر با بیش برآژش^{۱۱} است، با کاهش ویژگی ها عملکرد SVM تحت تأثیر قرار نمی گیرد.

¹ Support vector machine(SVM)

² Random Tree

³ Random Forest

⁴ overfitting

اعلام می کند؟ ارزیابی بر روی داده آزمایش اوپسالا [23] انجام گرفته است. این پیکره در مجموع شامل ۶۱۴ جمله^۵ و ۱۶۲۷۴ واژه (متوسط ۲۶/۵ واژه در هر جمله) است که به طور کامل قابل مقایسه با بخش آزمون و توسعه پیکره های 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).

این پیکره طبق شیوه‌نامه مرجع‌گزینی پیکره واژگان هم‌مرجع، برچسب‌گذاری شده و دارای برچسب‌های دقیق است. پیکره اوپسالا به صورت 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 شانزده هزار توکن پیکره اوپسالا مشاهده می‌شود.

(جدول-۸): نتایج ارزیابی بر روی شانزده هزار توکن پیکره اوپسالا
 (Table-8): Evaluation results on 16,000 Uppsala corpus tokens

معيار	muc	b3	ceafe	ceafm	connl
١ فايل	٧٧/٢٧	٥٤/٩	٦٤/٥٦	٥٦/٧٨	٦١/٦٢
٢ فايل	٦٣/٣	٥٦/٨١	٥٩/١٧	٥٨/٧٨	٥٩/٥١
٣ فايل	٧٧/٠١	٥١/١٦	٥٢/٦٤	٤٦/٠٧	٥٦/٧٢
٤ فايل	٦٦/٤٢	٥٥/٩٣	٥٩/٨٨	٥٩/٣٦	٦٠/٣٩
ميانگين	٦٩/٢٥	٥٤/٧	٥٩/٠٦	٥٥/٢٤	٥٩/٥٦
k-fold	٦٣/٤٢	٥٦/٨	٥٧/٣١	٥٥/٩١	٥٨/٢٨

گفتنی است، برای هر یک از معیارهای ارزیابی، نتایج به صورت 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 ایجاد شده و با استفاده از ابزار Conll ارزیابی انجام شد.

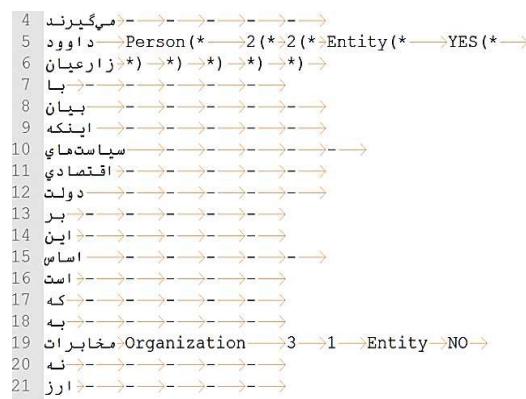
۶- نتیجہ گیری

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

⁴ Conll = (muc+b³+ceafm+ceafe)/4

^۵ لازم به ذکر است که ۱۴ جمله از بخش آموزش به بخش آزمون اضافه شده است تا متن یکباره باقی بماند.

که براساس الگوریتم توضیح داده شده در بخش ۴ تهییه شده است، می‌توان استفاده کرد. پیکره به گونه‌ای طراحی شده است که اضافه کردن برچسب جدید و استخراج اطلاعات از پیکره به آسانی صورت می‌گیرد. داده‌های پیکره و برنامه تهییه شده به ترتیب در بستر پارسیگان^۱ و سایت گیتهاب^۲ در دسترس عموم قرار دارند.



(شکل-۵): نمونه‌ای از داده‌های پر چسب خورده

(Figure-5): Example of labelled data

در تولید مدل‌های دسته‌بندی کننده ماشین بردار پشتیبان از ۹۰٪ داده کل پیکره استفاده شده است. و ۱۰٪ از داده نیز به عنوان داده آزمون مورد استفاده قرار گرفته است. داده آموزش نهصد هزار 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 بازتابی ۳ دارد. در جدول (۷) نتایج را برای این دو مدل مشاهده ممکن است.

٧- (حدوٰ) نتائج مدار SVM

(Table-7): SVM test results

(Table-7). SVM test results			
F1	Recall	Precision	نام مدل
۰/۵۵	۰/۳۸	۰/۹۹	بدون استفاده از ویژگی واژگان
۰/۷۱	۰/۶۴	۰/۸۰	با استفاده از ویژگی واژگان

در ارزیابی کلی سامانه مرجع گزینی ارزیابی با استفاده از ابزار ارزیابی استاندارد ارائه شده در مسابقات CoNLL صورت گرفته است. این ابزار نتایج را برای معیارهای MUC,B3,CEAF محاسبه می‌کند و براساس استاندارد CoNLL ارائه شده میانگین این معیارها را به عنوان معیار

¹ <http://parsigan.ir/datasources/Coref/4>

² <https://github.com/rcdat-rcisp/RCDAT-Coreference>

³ recall

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 ارائه شده در رقابت CONLL بر روی بخش آزمون از پیکره درخت وابستگی اوپسالا سنجیده شده است. نامیده‌ها در این ارزیابی به صورت gold در نظر گرفته شده‌اند و نتیجه measure با معیار CoNLL عدد ۴۸.۳٪ است. در مقایسه بین این دو روش شاهد بهبود نتیجه در سامانه حاضر نسبت به کار ارائه شده 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 به عمل آوریم.

۷- مراجع

- [۱] طباطبایی، ش. شکفته، ی. «سامانه پایه مرجع یابی گروههای اسمی در زبان فارسی با استفاده از قوانین ساده». اولین همایش جویشگر بومی، تهران، ۱۳۹۴.
- [۲] Sh. Tabatabaei 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.
- [۳] 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 .
- [۴] 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.
- [۵] Ch. Chen and Ng.Vincent, "Combining the best of two worlds: A hybrid approach to multilingual coreference resolution," Joint Conference on

رفته‌اند که دقت هر یک تأثیری بر دقت کل ابزار می‌گذارد. مهم‌ترین و اصلی‌ترین ابزاری که تأثیر بسیار زیادی بر نتیجه نهایی دارد قطعه‌بند است؛ زیرا قطعه‌بند هسته اصلی بخش تشخیص نامیده‌هاست. تشخیص نامیده‌ها مهم‌ترین بخش ابزار مرجع‌گزینی است؛ زیرا اگر نامیده‌ای در این بخش وارد رواں نشود باعث کم شدن بازخوانی می‌شود و هیچ راهی برای بازیابی آن نامیده وجود ندارد. و این مسئله بر دقت نهایی خیلی تأثیرگذار خواهد بود؛ لذا تولید یک ابزار دقیق قطعه‌بند برای مسئله هم‌مرجع یابی 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 بالا خواهی 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 تولید شده در این پژوهش از نظر حجمی بیش از پانصد برابر پیکره‌های قبلی زبان فارسی است و در عین حال به طور کامل قابل مقایسه با پیکره‌های مطرح ACE و Ontonotes است (ر.ک. جدول ۱). همچنین از نظر نامیده‌های برچسب‌گذاری شده این پیکره سه نوع نامیده را مشخص می‌کند؛ در حالی که پیکره‌های دیگر زبان فارسی، تنها یک نوع نامیده را مشخص می‌کنند.

سامانه مرجع‌گزینی تولید شده براساس معیار استاندارد CoNLL دارای f-measure نزدیک به شصت است. این در حالی است که سایر پژوهش‌های محدود انجام شده در فارسی دقت‌های متفاوتی از چهل تا نود درصد ارائه کرده‌اند که به دلیل این که سنجه دقت این پژوهش‌ها با معیارهای استاندارد رایج در مرجع‌گزینی صورت نگرفته است، این

- 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

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 رایانمه ایشان عبارت است از:

shadi.hn@gmail.com

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 رایانمه ایشان عبارت است از:

rahimi-z@redat.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



