IDMCDB_1923

ارائه چهارچوبی به منظور رفع ابهام معنی کلمات در زبان فارسی با حداقل نظارت

محمدرضا محمودوند[1]؛ مریم حورعلی[2]

چکیده

با پیشرفت روزافزون روشهای پردازش گفتار و زبان طبیعی، هم اکنون می‌توان در لبه دانش گام برداشت و به تولید برنامه‌های کاربردی پرداخت که پیش از این امکان پذیر نبود. برنامه هایی که به طور کامل با انسان در تعامل خواهند بود و ورودی‌های کامپیوتر را از حوزه محدود ابزارهای ورودی خارج می‌کند و آسان‌ترین فرمان ورودی یعنی زبان انسان را می‌پذیرد و درک می‌کند. ابهام زدایی معنی کلمات یکی از موضوعات مهم مطرح در پردازش زبان طبیعی می‌باشد. هدف اصلی زبان ارائه و نشان دادن یک مفهوم خاص به مخاطب می‌باشد. این مفهوم برگرفته از معانی کلمات موجود در آن زبان استخراج می‌شود. بدون شناسایی صحیح قصد و منظور کلمات در یک جمله نمی‌توان درک درستی از مفهوم و قصد گوینده جمله داشت. برای شناسایی صحیح مفاهیم موجود در متون نیز کامپیوتر باید نقش و معنی کلمات را شناسایی کند. این موضوع زمانی دشوارتر به نظر می‌رسد که کلماتی در زبان وجود دارند که با توجه به نقش و کلماتی که در همسایگی آن کلمه وجود دارد، معانی متفاوتی اتخاذ می کنند.

با توجه به گسترش برنامه‌های کاربردی متفاوت به زبان فارسی این نیاز احساس می‌شود که راه حلی برای ابهام زدایی کلمات در زبان فارسی نیز ارائه شود. کمبود داده‌های آموزشی بزرگترین چالش پیش‌روی ابهام زدایی معنی کلمات در زبان فارسی می‌باشد. به منظور روبه‌رو شدن با این مشکل در این پژوهش از رویکرد یادگیری ماشین با حداقل نظارت استفاده شده است. روش به‌کار گرفته شده در این پژوهش با توجه به ویژگی‌های تعریف شده از کلمات هدف و همچنین استفاده از روش یادگیری هماهنگ نسبت به رفع ابهام کلمات مبهم معنایی در زبان فارسی اقدام می‌شود. پیکره استخراج شده از اخبار منتشرشده توسط خبرگزاری‌ها به عنوان پیکره مرجع استفاده شده است. با ارزیابی سامانه توسط پیکره موجود بر روی سه کلمه مبهم درنظر گرفته شده، روش پیاده‌سازی شده قادر به شناسایی صحیح معنی ۵۳۶۸ سند و با معیار فراخوانی ٪۸۸ ، معیار دقت ٪۹۵ و معیار صحت ٪۹۳ می‌باشد.

کلمات کلیدی

پردازش زبان طبیعی، ابهام زدایی معنی کلمات، یادگیری نیمه نظارتی، متن کاوی، شناسایی آماری الگو

Introduction a framework for Persian word sense disambiguation based on semi-supervised method

Mohamadreza Mahmoodvand; Maryam Hourali

ABSTRACT

ambiguous words have found in every language and also in Persian. so there is a big requirement in Persian language to work on this field to determine right sense for Persian words. Because of shortage of manually sense-tagged data is an obstacle to supervised word sense disambiguation methods. In this thesis we introduce a framework based on semi-supervised approach to disambiguate Persian words. At first a crawler fetches document form official Persian news agencies that included ambiguous word then seed labeled data divided into three part as an input to hybrid method based on decision list and tri-training. finally when decision about sense made, a new training data added to data set and repeat that procedure until no document were unlabeled. Result for disambiguation of three Persian word with this framework was remarkable in compare of related works. In this thesis we achieve precision 95.31 % , recall 88.39% , accuracy 93.73 % in average of disambiguation of three words on Persian language.

KEYWORDS

Natural language processing, Word sense disambiguation , Semi-supervised learining , Textmining , Statistical pattern recognition

---

۱. کارشناس ارشد هوش مصنوعی ، دانشگاه صنعتی مالک اشتر mr.mahmoodvand@yahoo.com

۲. استادیار و عضو هیات علمی دانشگاه صنعتی مالک اشتر، مجتمع ICT hourali@mut.ac.ir

۱. مقدمه

ابهام زدایی معنایی کلمات به معنای مشخص کردن معنی صحیح یک کلمه می‌باشد که در نگاه اول مبهم بوده و بتوان چندین کلاس معنایی را به آن نسبت داد. به طور معمول کلماتی که تنها دارای یک معنی مفرد نمی‌باشند را می‌توان به عنوان کلمات هدف معرفی کرد که قصد رفع ابهام معنی آنها را داریم. انسان با توجه به مفهومی که از جمله استنباط می‌شود به راحتی می‌تواند معنی صحیح کلمه را تشخیص دهد اما کامپیوتر نیاز به اطلاعاتی فراتر از دریافت جمله ورودی را دارد تا نوع کلاس معنایی کلمه مبهم را مشخص نماید. با استفاده از روش‌های ابهام زدایی می‌توان مفاهیم جمله را استخراج کرد و برای تصمیم‌گیری در رابطه با کلاس معنایی کلمه هدف در اختیار کامپیوتر قرار داد. ابهام زدایی معنایی کلمات در زبان‌های رایج نظیر انگلیسی بسیار مورد توجه بوده و پژوهش‌های بسیاری در این حوزه برای رفع ابهام کلمات مبهم در زبان انگلیسی صورت گرفته است. کلمات مبهم در تمام زبان‌های مرسوم و زنده دنیا یافت می‌شوند. زبان فارسی نیز از این قائده مستثنی نبوده است و نیاز است که روش‌های ابهام زدایی معنایی بر روی کلمات مبهم زبان فارسی نیز اعمال گردند. با توجه به کاربرد فراوان سامانه‌های ابهام زدایی معنایی کلمات در دیگر برنامه‌های کاربردی پردازش زبان طبیعی لازم است که نسبت به حل این مساله اقدام گردد تا یکی از گلوگاه‌های موجود در توسعه برنامه‌های کاربردی از بین رفته و فضا برای فراهم آمدن پژوهش‌ها و برنامه‌های کاربردی قدرتمند و موثر در زبان فارسی ایجاد شود. بر این اساس برای رفع ابهام معنی کلمات مبهم فارسی باید راه‌حل مناسبی ارائه داد.

هدف از این پژوهش ارائه یک راه حل جامع و قابل توسعه به منظور ابهام زدایی معنایی کلمات می‌باشد که با استفاده از آن بتوان با صرف کمترین هزینه و زمان نسبت به رفع ابهام معنایی کلمات در زبان فارسی اقدام کرد. به همین منظور این راه حل جامع در قالب یک چهارچوب ارائه شده است که اجزای متفاوتی تشکیل دهنده آن می باشند. بر اساس ساختار کلی چهارچوب ارائه شده می توان با فراهم کردن اجزای متفاوتی که در ادامه مقاله شرح داده خواهند شد به حل مساله ابهام معنی کلمات پرداخت. سعی بر این بوده است که ساختار تشکیل دهنده این مدل به نحوی باشد که توانایی مقاومت در میان چالش های پیش روی ابهام زدایی معنایی کلمات در زبان فارسی باشد.

یکی چالش های عمده ابهام زدایی معنی کلمات، فراهم آوردن یک پیکره مناسب می باشد که قابل استفاده برای مقاصد آموزشی حل مساله ابهام زدایی باشد. به منظور آموزش یک سامانه رفع ابهام معنی کلمات، پیکره ای مناسب است که در آن تمام کلمات هدف مبهم مشخص شده باشند و متمایز از دیگر کلمات باشند و برچسب معنایی صحیح این کلمات از بین کلاس های معنایی مربوط به کلمه هدف تعیین شده باشد. همانطور که مشخص است فراهم آوردن پیکره ای جامع که شامل تمامی کلمات هدف مبهم و همچنین برچسب گذاری معنایی کلمات باشد نیاز به صرف زمان و هزینه بالایی خواهد داشت. ضمن آنکه دقت در تهیه پیکره نیز بسیار اهمیت دارد زیرا در نهایت این پیکره است که به عنوان مرجع آموزشی استفاده می شود از این روی باید استانداردهای لازم برای استفاده در کاربرد ابهام زدایی معنایی کلمات را داشته باشد. در اکثر زبان های موجود دنیا کمبود پیکره های مناسب برای ابهام زدایی مشهود می باشد اما این نقصان بیش از پیش در زبان فارسی جلوه می کند.

هر چه در زبان بیشتر از آرایه های ادبی و کنایات استفاده شده باشد تشخیص مقصود صحیح کلمه دشوارتر خواهد. زبان فارسی نیز یکی از زبان های غنی از آرایه های ادبی می باشد که همین موضوع تشخیص صحیح معنی کلمه را به امری سخت بدل می کند. ضمن آنکه به دلیل رسم الخط زبان فارسی و شیوه نوشتار در این زبان، تشخیص کلمات هدف و همچنین رخداد ویژگی ها در یک عبارت با مانع سختی رو به رواست که در دیگر زبان ها نیازی به رفع این موارد وجود نخواهد داشت. به همین منظور برای پیاده سازی یک سامانه جامع ابهام زدایی معنایی کلمات در زبان فارسی باید یک ساختار پیش پردازش بومی نیز تعبیه شود تا در ابتدای مراحل نسبت به غربالگری اسناد مناسب اقدام نماید.

چهارچوب ارائه شده یک مدل بومی برای حل مساله ابهام زدایی معنایی کلمات بوده که سعی دارد ضمن ارائه راه حل برای رفع ابهام معنی کلمات نسبت به ارائه راهکار به منظور مقابله چالش های موجود در این حوزه بپردازد. بستر ارائه شده آماده ی رفع ابهام معنی تمام کلمات مبهم در زبان فارسی است و تنها نیاز است که برای هر کلمه هدف ورودی‌های اولیه و مناسب به سامانه داده شود تا بدون تمایز مابین کلمات نسبت به رفع ابهام کلمه هدف در میان عبارات و جملات فارسی اقدام شود. یکی از ویژگی های مهم این چهارچوب توسعه پذیر بودن آن است که در آن صورت می توان بنابر نیاز همگام با پژوهش های نوین موجود در این حوزه، الگوریتم های جدیدی را در اجزای مختلف آن اعمال کرد تا خروجی های بهبود یافته و نتایج جدیدی فراهم آید. به این ترتیب امکان بررسی روش‌های مختلف ابهام زدایی معنایی کلمات در زبان فارسی فراهم می‌شود و می‌توان موثرترین روش را انتخاب کرد.

۲. پژوهش‌های پیشین

با وجود دشواریهای پیش روی پژوهشگران حوزه پردازش زبان طبیعی برای پژوهش در زمینه ابهام زدایی معنایی کلمات ، تحقیقات فراوانی با هدف پیدا کردن راه حل این مساله صورت پذیرفته است. بیشتر پژوهش‌های صورت گرفته در حوزه ابهام زدایی معنایی کلمات با روش نظارتی

عملیاتی می‌شوند. مهم‌ترین مسئله در پیاده‌سازی این نوع روش، فراهم آوردن یک دیتاست مناسب و برچسب‌گذاری شده براساس معنای صحیح کلمات مبهم و همچنین حجم انبوه داده‌های موجود در آن می‌باشد[۲] . ویژگی‌هایی که براساس کلماتی که در همسایگی کلمه هدف می‌باشند نیز بسیار اهمیت دارند. با استفاده از این نوع ویژگی‌ها می‌توان با دقت بیشتری معنی صحیح کلمه هدف را حدس زد. در [۲] مقایسه‌ای از روش‌های موجود ابهام زدایی نظارتی در زبان انگلیسی و با موضوع اطلاعات پزشکی صورت گرفته است. در اطلاعات پزشکی پنجره هدف به اندازه تمام پاراگراف دربرگیرنده کلمه هدف درنظر گرفته شده است اما برای متون ساده به زبان انگلیسی محدوده پنجره تنها شامل ۴ تا ۱۰ کلمه مجاور کلمه هدف را دربرمی‌گیرد. با استناد بر نتایج حاصله روش بیز ساده تا ۹۸ درصد در شناسایی کلمات مبهم موفق عمل کرده است. به دلیل چالش عمده موجود و کمبود داده‌های مناسب برای استفاده در روش‌های ابهام‌زدایی معنایی کلمات بیشتر پژوهش‌ها سعی در ایجاد الگوریتمی دارند که توسط آن بتوانند این داده مناسب را ایجاد کنند.

در [۳] نیز به منظور استفاده از روش‌های یادگیری نظارتی، یک نرم‌افزار جانبی طراحی شده است که نمونه اسناد مناسب را بصورت خودکار از بستر شبکه معنایی کلمات، فرهنگ‌های لغات و وب جمع‌آوری کند. روش بیز ساده نیز یکی از روش‌های متداول ابهام زدایی نظارتی محسوب می‌شود[۴]. این روش بر روی دو کلمه متدوال مبهم زبان انگلیسی line و interest اعمال شده است. در بیز ساده فرض می‌شود که هریک از بردارهای ویژگی موجود بصورت مستقل تنها به یکی از کلاس‌های معنایی اشاره دارند و یکتا می‌باشند. در پژوهش مذکور از بردارهای ویژگی دودویی استفاده شده است که تنها تعیین می‌کند که مجموعه کلمات ویژگی در بستر سند حضور خواهند داشت یا خیر. در واقع در روش بیز قصد داریم به کلاس معنایی اشاره داشته باشیم که بتواند بیشترین نرخ احتمال را در بین دیگر ویژگی‌ها برای کلمه هدف بدست آورد. همانند لیست تصمیم، در روش بیز ساده نیز ممکن است با فراوانی‌های صفر روبه‌رو شویم که برای رفع این مانع باید از ضریب هموارسازی استفاده نمود. در نهایت روش بیز ساده این توانایی را دارد که معنی صحیح این دو کلمه مبهم در زبان انگلیسی را به ترتیب با ۸۹ و ۸۸ درصد به درستی رفع ابهام نماید. نتایج بدست آمده از اعمال روش بیز ساده در زبان‌های مختلف قابل قبول بوده است[۷-۵].

مرز بین روش‌های ابهام زدایی نظارتی و نیمه نظارتی خود نیز دارای ابهام می‌باشد. زیرا در روش‌های نیمه نظارتی نیز مانند روش‌های نظارتی از مجموعه‌ای داده برچسب گذاری شده و همچنین روش‌های یادگیری ماشین استفاده می‌شود. در واقع بخشی از فرآیند روش‌ها مشترک می‌باشد. اما در روش نیمه نظارتی به دلیل کمبود منابع مناسب برای آموزش تنها به بخش کوچکی از داده‌های آموزشی دسترسی داریم و بخش عظیمی از داده‌های موجود بدون هیچگونه اطلاعات و برچسبی به عنوان ورودی در اختیار روش‌های نیمه نظارتی قرار می‌گیرد[۸]. با این وجود اساسی‌ترین هدف یک روش نیمه نظارتی تولید یک کلاس‌بند معنایی با حداقل داده‌های موجود می‌باشد. برای این منظور الگوریتم با حجم اندکی از نمونه‌ها فرآیند آموزش را آغاز می‌کند و پس از اتمام به ارزیابی داده‌های بدون برچسب می‌پردازد. پس از مشخص شدن کلاس صحیح معنایی، داده تازه تولید شده به مجموعه دادگان اولیه اضافه شده و فرآیند آموزش تکرار می‌گردد. برای تهیه مجموعه داده اولیه می‌توان از برچسب گذاری به کمک نیروی انسانی استفاده کرد [۹]و یا از انتخاب خودکار بر اساس معیارهای هیوریستیک بهره برد[۱۰].

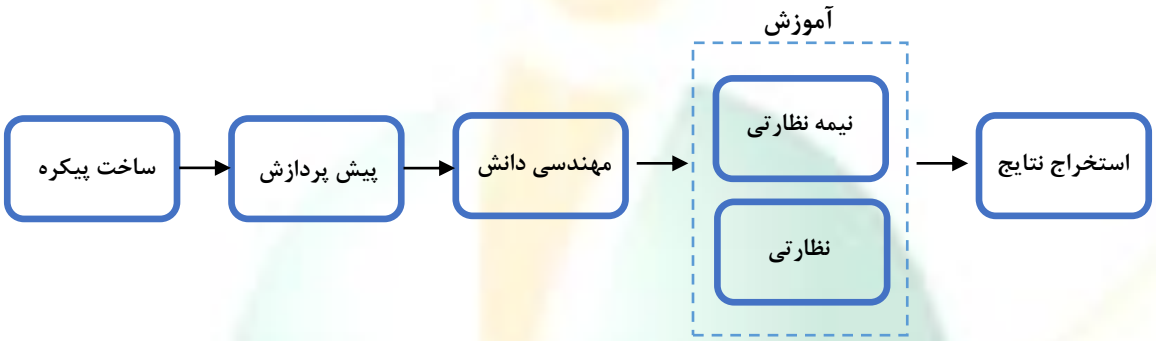
تصمیم‌گیری و آموزش نمونه‌ها نیز می‌تواند با روش‌های مختلفی صورت گیرد. ممکن است در روش اعمال شده یک کلاس‌بند مستقل به تصمیم‌گیری راجب نمونه‌ها بپردازد و یا در روشی دیگر از بازخورد موجود بین نظرات حادث شده توسط چند کلاس‌بند موازی و مستقل درباره کلاس نهایی کلمه هدف تصمیم‌گیری شود. [۱۱] نمونه ای از ترکیب چند واحد کلاس‌بند در تصمیم‌گیری نهایی می‌باشد. در روش مطرح شده از دو کلاس‌بند برای آموزش دونوع از ویژگی‌های تعریف شده استفاده شده است، درحالیکه این امکان وجود دارد که همه ویژگی‌ها در یک کلاس‌بند مورد آموزش قرار گیرند و پاسخ نهایی بصورت مستقل توسط آن اعمال شود.در [۱۲] نمونه‌های مختلفی از روش یادگیری نیمه نظارتی که براساس الگوریتم یاروفسکی بنا شده است، ارائه گردیده است. روش‌های نیمه نظارتی نیز با چالش‌های متفاوتی روبه‌رو هستند از جمله می‌توان به شرط توقف آموزش، تعداد تکرار هر دوره‌ی آموزش و همچنین نحوه برگزیدن و تصمیم‌گیری برای ارائه خروجی نهایی از اهمیت بالایی برخوردارد است[۱۳]. یکی دیگر از روش‌های نیمه نظارتی اعمال شده به منظور رفع ابهام معنی کلمات، روش آموزش سه‌گانه می‌باشد[۱۴].

کمبود مجموعه داده برچسب گذاری شده قابل استفاده برای ابهام زدایی معنی کلمات فارسی در [۱۵]مورد توجه قرار گرفته است. یک روش نیمه نظارتی در این پژوهش برای دسته‌بندی صحیح کلمات هم‌نوشت ارائه شده است. روش ارائه شده یک روش دانه درشت بوده، و تنها بر روی معانی نزدیک کلمات هم‌نوشت تمرکز دارد. برای یادگیری نظارتی از لیست تصمیم استفاده شده است و همچنین برای یادگیری نیمه نظارتی، روش آموزش سه‌گانه به کار گرفته شده است. برای شناسایی ۱۵۰۰ کلمه هدف در پیکره درنظرگرفته شده، روش مذکور برای رفع ابهام کلمه هدف "اعمال" ۹۳ درصد و برای کلمه "اشراف" ۸۴ درصد موفق عمل می‌کند.

در پاره‌ای از اوقات نیز اندک دادگان برچسب گذاری شده و مناسب ابهام زدایی مهیا نمی‌باشد. در این صورت تنها راه حل استفاده از روش‌های گنجینه‌ای و بدون نظارت می‌باشد. نمونه‌ای از این روش‌ها در[۱] به کار گرفته شده است و هدف نهایی استفاده از این روش در یک سامانه مترجم ماشینی می‌باشد. حتی روش استخراج معانی یک کلمه هدف نیز بصورت خودکار صورت می‌پذیرد. برای تعیین کلاس‌های معنی کلمات به یک فرهنگ لغت دوزبانه مراجعه می‌شود و گراف وابستگی معنایی کلمات براساس پیکره فارسی درنظر گرفته شده شکل می‌گیرد. در این پژوهش از نسخه دوم پیکره همشهری به عنوان پیکره مرجع استفاده شده است. این گراف شامل معانی مختلف کلمات مبهم جمله و وابستگی معنایی بین آنها است. همچنین روش جدیدی به منظور تقویت معیار وابستگی معنایی کلمات ارائه شده است. یکی از ویژگیهای این روش عدم وابستگی آن به زبانهای مبدا و مقصد میباشد و میتواند برای هر جفت زبان در ترجمه ماشینی مورد استفاده قرار گیرد.

۳. پیکره بندی چهارچوب

چهارچوب ارائه شده دارای اجزای مستقلی است که در عین استقلال بصورت هدفمند با یکدیگر همکاری دارند. پیکره بندی چهارچوب به صورتی طراحی شده است که با چالش‌های موجود در ابهام زدایی معنایی کلمات در زبان فارسی مقابله کند. در ابتدا پیکره مناسب برای استفاده در روش‌های یادگیری ماشین درنظر گرفته شده اتخاذ می‌شود سپس با انجام فعالیت‌های پیش پردازش بومی شده به غربال‌گری اسناد پرداخته می‌شود و خلوص داده‌های آموزشی بالا خواهد رفت. با استفاده از مهندسی دانش برچسب معنایی هر یک از اسناد آموزشی نسبت داده می‌شود و ویژگی‌های مرتبط با هر کلمه هدف تعریف می‌گردد. در ادامه یک روش یادگیری ماشین نظارتی انتخاب می‌شود که با استفاده از مجموعه داده‌های اولیه [۲]برچسب گذاری شده آموزش می‌بینند و در داخل یک ساختار یادگیری نیمه نظارتی جای می‌گیرد. در نهایت عبارات شامل کلمات مبهم هدف که در هیچ یک از مراحل چهارچوب استفاده نشده‌اند و فاقد برچسب نیز هستند برای آزمون در اختیار سامانه قرار می‌گیرد و برچسب نهایی معنی هر یک از این جملات ارائه می‌گردد. ساختار کلی اجزای تشکیل دهنده چهارچوب در شکل ۱ نمایش داده شده است.



شکل (۱) ساختار کلی چهارچوب ابهام زدایی معنایی کلمات در زبان فارسی

۱،۳. استخراج پیکره

با وجود اینکه دو پیکره متداول زبان فارسی یعنی پیکره دکتر بیجن خان[۱۶] و پیکره همشهری [۱۷] کاربرد فراوانی در اکثر پژوهش‌های مرتبط با پردازش زبان طبیعی فارسی و استخراج اطلاعات در زبان فارسی دارند، اما برای برخی کاربردهای خاص نظیر ابهام زدایی معنایی کلمات مناسب نمی‌باشند. در روش‌های ابهام زدایی معنایی کلمات نیاز به اسنادی داریم که در آنها کلمه هدف مبهم حضور داشته باشد. تعداد این اسناد باید تعداد قابل قبولی باشد تا قابلیت استفاده از روش‌های یادگیری ماشین فراهم باشد. ضمن اینکه طولانی بودن این اسناد در این وظیفه کاربردی نیستند. بر اساس روش‌های موجود تنها کلمه هدف و محدوده خاصی از کلمات که در اطراف کلمه هدف حضور دارند، مورد استفاده قرار می‌گیرند. این موضوع باعث می‌گردد که حجم انبوهی از اطلاعات بدون استفاده باقی بمانند.

برای مقابله با چالش کمبود داده‌های آموزشی مناسب برای ابهام زدایی معنایی کلمات فارسی، از روش خزش[۳] با هدف ساخت پیکره استفاده شده است. با استفاده از موتور جستجوی گوگل می‌توان پیوند صفحات مورد نظر را که در آنها کلمه هدف موجود است را جستجو کرده و آنها را یافت. پس از مشخص شدن آدرس صفحات، آنها را به ربات خزشگر ارجاع داده و در این حالت ربات خزشگر در بین صفحات پیوند داده شده به دنبال اسنادی می‌گردد که در آنها کلمه هدف حضور داشته باشد. در روشهای ابهام زدایی معنایی کلمات تنها کلمه هدف برای استفاده در روش‌های یادگیری ماشین کافی نیست بلکه باید پنجره‌ای به اندازه کافی و مشخص از کلمات همسایه کلمه هدف نیز استخراج شود. به همین منظور ربات خزش طوری برنامه‌ریزی شده است که بتواند جملاتی که کلمه هدف را دارا می‌باشند به طور کامل استخراج کند. به این ترتیب هر یک اسناد بدست آمده نهایی جملاتی هستند که دربرگیرنده کلمه هدف می‌باشند.

به منظور پیاده سازی بستر ربات خزش و دریافت پیوند صفحات خبرگزاری‌ها یک رابطه کاربری تحت وب طراحی شده است که توسط این رابط، کاربر در ابتدا نوع کلمه هدف مبهم مورد نظر خود را انتخاب می‌نماید. این کلمه هدف می‌تواند یکی از سه کلمه هدف کاندید در این پژوهش یعنی "شیر" ، "راست" ، "تار" و ... باشد. البته ذکر این نکته ضروری است که کلمات هدف با توجه به نظر کاربران قابل تعریف می‌باشد و با فراهم آوردن مجموعه ویژگی‌های هر کلمه هدف دلخواه می‌توان اسناد مربوط به آن را استخراج کرد و پس از آن به رفع ابهام معنایی آن اسناد اقدام کرد. پس از آن می‌توان آدرس صفحات مورد نظر را که با کمک موتور جستجوی گوگل بدست آمده را در بخش‌های در نظر گرفته شده وارد کرد. سپس آدرس ها برای ربات ارسال شده و با خزش در آدرس‌های مقصد در نهایت جملاتی که کلمه هدف را دربرگیرند به پایگاه داده اضافه می‌گردد و نتایج و تعداد اسناد اضافه شده برای کاربر به نمایش درمی‌آید. به این ترتیب با صرف زمانی اندک می‌توان یک پیکره مناسب برای آموزش یک سامانه ابهام زدایی معنایی کلمات در زبان فارسی تولید کرد که اسناد این پیکره می‌تواند در برگیرنده عبارت و جملاتی باشد که شامل هر کلمه مبهم هدف در زبان فارسی باشد.

---

[2] Seed
[3] Crawl

۲,۳. پیش پردازش

با توجه به اینکه پیکره توسط جستجوی ربات خزشگر و بدون دخالت انسان بدست آمده است بدیهی است که نیاز است که پیکره مورد بررسی قرار گیرد تا از این بابت که پیکره نهایی عاری از عبارات و حروف زائد است مطمئن باشیم. زیرا در صورتی که اسناد شامل موارد و اطلاعات اضافی باشند، روند آموزش سامانه با مشکل رو به رو خواهد شد و باعث تحمیل سربار محاسباتی بالا و نامفیدی در روند یادگیری خواهد شد. مراحل پیش‌پردازش داده‌ها که در پژوهش‌های حوزه پردازش زبان طبیعی می‌باشد، شامل حذف حروف ربط، حروف توقف، حذف زائده‌های نوشتاری از جمله رسمالخط عربی و ... می‌باشد. روش‌های مرسوم پیش‌پردازش نیز بر همین پایه بناگذاری شده اند. نبود یک پیکره که برای اعمال و ارزیابی روش ابهام زدایی معنایی کلمات مناسب باشد، باعث شد که در پژوهش از متون استخراج شده از بستر وب استفاده شود. همین موضوع یک بخش به فاز پیش‌پردازش اضافه خواهد کرد. در بستر وب، متون به صورت زبان زبان نشانه‌گذاری ابرمتنی ارائه می‌شوند. همچنین از زبان شیوه نمایش آبشاری نیز برای طرح و نقش دادن به متون استفاده می‌شود. این موضوع باعث می‌شود که علاوه بر متون استخراج شده در بستر وب که شامل کلمه هدف نیز می‌باشند، محتویات تگ‌های این زبان‌ها نیز در نمونه‌های استخراج شده وجود داشته باشند. برای ایجاد یک داده مناسب باید هر یک از متون استخراج شده پالایش شده و تگ‌های نامربوط از آنها حذف گردند. این موضوع به عنوان یک مرحله از پیش‌پردازش در این پژوهش اعمال شده است.

زبان مرجع برای ابهام زدایی معنایی کلمات در این پژوهش زبان فارسی در نظر گرفته شده است. همین موضوع باعث ایجاد چالش‌های جدید فراتر از زبان‌های دیگر نظیر زبان‌های لاتین می‌باشد. یکی از این چالش‌ها طرز نوشتار زبان فارسی است. در زبان فارسی، نوشتار کلمات و حروف می‌تواند به صورت چسبیده و در اصطلاح "سرهم" نوشته شوند. همین موضوع باعث می‌شود که شناسایی کلمه هدف در متون فارسی دشوار باشد. به طور مثال اگر کلمه هدف "شیر" باشد، متن ورودی نیز شامل عبارت "شیرها" شود. این موضوع باعث می‌شود که موقعیت مکانی کلمه هدف درست تشخیص داده نشود و امکان اعمال ویژگی‌های ترتیبی وجود نداشته باشد. به همین منظور روش قطعه بندی‌کلمات هدف در این پژوهش اعمال شده است. به همین منظور در هنگام استخراج نمونه متون شامل کلمه هدف، هر یک از کلمات هدف پالایش شده تا در هیچ نوشتاری به صورت چسبیده به دیگر کلمات و عبارات و یا علائم نوشتاری حضور نداشته باشد. به این ترتیب تمام کلمات هدف موجود در متن بصورت منفرد بدون پیشوند و پسوند خاصی ارائه می‌شوند. همین موضوع باعث افزایش درصد شناسایی صحیح کلمات هدف در متون خواهد شد و درصد صحت نتایج را افزایش خواهد داد.

۳,۳. مهندسی دانش

بخش مهندسی دانش این چهارچوب تنها بخشی است که نیاز به دریافت اطلاعات از کاربر را دارد. در واقع این بخش به برچسب گذاری اسناد آموزشی و تعریف پایگاه داده ویژگی‌های کلمه هدف اختصاص دارد. پس از اتمام فاز پیش پردازش، داده ها آماده هستند تا در فرآیند آموزش سامانه ابهام زدایی کلمات شرکت کنند اما پیش از آن نیاز به برچسب گذاری قسمتی از این داده‌ها وجود دارد. تعریف ویژگی‌ها به دو روش عمده صورت می‌گیرد. در نوع اول به دنبال ویژگی‌هایی هستیم که در فاصله مکانی خاصی از کلمه هدف حضور داشته باشند و وابسته به موقعیت مکانی خود در جمله هستند. در نوع دوم از ویژگی‌ها برداری تعریف می‌شود که بیانگر حضور و یا عدم حضور یک مجموعه از کلمات مرتبط با یک معنی خاص کلمه هدف در جمله می‌باشد. برای بدست آوردن نتایج قابل قبول در این پژوهش سعی شده است که از مزیت‌های هر دو نمونه از ویژگی‌های مطرح شده بیشترین بهره برده شود. برای این منظور ساختاری برای تعریف ویژگی‌ها ایجاد شده است.

بر اساس این ساختار ویژگی‌ها به دو نوع منظم و نامنظم تقسیم بندی شده‌اند. در هر ویژگی کلمه هدف مشخص شده است و همچنین عبارتی به عنوان نشانه ویژگی تعیین شده است. نشانه ویژگی، بخشی است که الگوریتم استخراج ویژگی قصد دارد موجودیت آن را در میان نمونه اسناد ورودی پیدا کند. اگر ویژگی از نوع منظم باشد، موقعیت مکانی قبل و بعد نسبت به کلمه هدف ارائه شده است که به این معنا خواهد بود که نشانه ویژگی در فاصله‌های تعیین شده باید پیدا گردد. اگر ویژگی از نوع نامنظم باشد این فواصل اهمیتی پیدا نمی‌کند. در نهایت این چهارچوب شامل معنی صحیح کلمه هدف نیز خواهد بود که هر ویژگی به یک معنی منحصربه‌فرد اشاره خواهد کرد. با ارائه این چهارچوب هریک از دو نوع ویژگی اصلی در قالب یک جدول ویژگی‌ها قابل ارائه و استفاده در متون هستند. ویژگی‌های منظم به منظور پیدا کردن عبارات در محدوده معینی فعالیت می‌کنند و ویژگی‌ها نامنظم تنها به دنبال حضور یک عبارت بصورت نامحدود در بستر جمله و متن هستند.

نکته مثبت دیگری که در نوع استفاده از این ساختار تعیین شده برای ویژگی‌ها نسبت به دیگر ویژگی‌ها وجود دارد امکان استفاده از N-gram به عنوان عبارت نشانه ویژگی‌های نامنظم می‌باشد. به دلیل خصوصیات متفاوت زبان فارسی نسبت به دیگر زبان‌ها، استخراج و انتخاب ویژگی‌های مناسب ابهام زدایی معنایی کلمات نیز دشوار خواهد بود. در زبان فارسی مجموعه‌ای از عبارات به صورت N-gram وجود دارند که حضور آنها در یک جمله دلالت بر معنی خاصی از کلمه هدف می‌باشد. امکان استفاده از این N-gram ها در ویژگی‌های منظم وجود ندارد. زیرا فاصله بین کلمات در این ویژگی‌ها برابر با همان فاصله نوشتاری[۴] در زبان می‌باشد و در صورت اینکه مرزبندی بین کلمات به وضوح مشخص نشده باشد، فرآیند شناسایی ویژگی‌ها در متن با مشکل روبه‌رو خواهد شد. بنابراین برای تعریف یک ویژگی ۳-gram باید سه کلمه به صورت آبشاری[۵] و پی‌درپی به عنوان ویژگی مطرح شوند که هزینه محاسباتی و زمانی بالایی خواهد داشت. به

[4] Space
[5] Cascade

طور مثال با استفاده از این روش می‌توان ویژگی‌هایی نظیر "چپ به راست"، " رک و راست"، "رو راست" و ... را برای کلمه هدف "راست" تعریف کرد. تعریف این ویژگی‌ها علاوه بر اینکه بارمحاسباتی کمتری نسبت به ویژگی‌های منظم ساده دارند، باعث دستیابی سریعتر به جواب نهایی خواهند شد. نمونه ای از ویژگی‌های تعریف شده در جدول شماره (۱) مشخص شده است.

جدول (۱) مثالی از ویژگی‌های تعریف شده برای کلمات هدف

کلمه هدف	کلمه ویژگی	نوع ویژگی	برچسب معنایی	جمله
شیر	صنایع	منظم	MILK	در حال حاضر همچنان یارانه صنایع شیر پرداخت نشده است.
راست	افراطی	منظم	POLITICAL	انتقاد از حزب حاکم ژاپن به دلیل ارتباط با گروههای راست افراطی
تار	عنکبوت	نامنظم	STRING	تارهای تنیده شده توسط عنکبوت در عین سبکی بسیار مقاوم می‌باشد.

۴,۳. آموزش

با مشخص شدن مجموعه اسناد برچسب گذاری شده اولیه و پایگاه داده ویژگی‌های مربوط به کلمه هدف، حال تمامی موارد مورد نیاز برای استفاده از روش‌های یادگیری ماشین با هدف آموزش سامانه رفع ابهام فراهم شده است. برای مواجه شدن با کمبود داده‌های آموزشی یک روش یادگیری نیمه نظارتی[6] در این چهارچوب به کار گرفته شده است تا بتوان با چالش کمبود داده‌های آموزشی در زبان فارسی روبه‌رو شد. در روش نیمه نظارتی نیازی به مجموعه انبوه داده‌ها برای آموزش وجود ندارد. فرآیند آموزش با یک مجموعه اولیه کوچک آغاز می‌شود. پس از اتمام فرآیند آموزش، کلاس بند با داده‌های ورودی خام روبه‌رو می‌شود و در مورد برچسب معنایی آنها تصمیم گیری می‌کند. پس از مشخص شدن برچسب نهایی سند ورودی، این سند به مجموعه داده‌های اولیه افزوده می‌شود و بار دیگر روند آموزش با مجموعه جدید و بهبود یافته برقرار می‌گردد. این روند تا زمان اتمام تمام داده‌های خام بدون برچسب ادامه می‌یابد. برای کلاس بندی داده‌ها از یک روش یادگیری نظارتی نظیر روش بیز ساده، لیست تصمیم، نزدیکترین همسایگی و ... استفاده می‌شود. در این پژوهش از لیست تصمیم[7] [۱۸] در فاز یادگیری نظارتی استفاده شده است.

هدف لیست تصمیم این خواهد بود که در انتها معین کند داده ورودی یا به عبارت دیگر کلمه‌ی ورودی دارای کدام معنی صحیح می‌باشد. برای اینکه این امر میسر شود، لیست تصمیم از الگوهای موجود در متن دربرگیرنده کلمه هدف استفاده می‌کند. برای شناسایی الگوهای موجود نیز از ویژگی‌های استخراج شده موجود در بانک استفاده می‌شود. در نهایت برای تصمیم‌گیری نهایی از تصمیمی استفاده می‌شود که در بین لیست موجود بیشترین امتیاز را داشته باشد و همچنین در متن دربرگیرنده کلمه هدف نیز ویژگی‌های مورد نظر آن تصمیم وجود داشته باشد برای شناسایی معنی صحیح کلمات باید توزیع هریک از ویژگی‌ها را در میان داده‌های آموزشی مورد ارزیابی قرار داد. برای این منظور یک الگوریتم جستجو بر روی تمام داده‌های آموزشی اعمال می‌شود که طی آن حضور هر یک از ویژگی‌های ترتیب متنظم کلمات بررسی می‌گردد. طی این فرآیند تعداد رخداد هر یک از ویژگی‌ها در جملات به همراه معنی صحیح کلمه هدف بدست می‌آید. در واقع هدف اصلی لیست تصمیم، یافتن مجموعه‌ایی بزرگ و پرتکرار از ویژگی‌ها می‌باشد که بیشترین و نزدیکترین تخمین به معنی اصلی کلمه را به ارمغان آورد. در این مرحله از یادگیری یک نوع غربال گری صورت گرفته و ویژگی‌هایی که ارزش بالاتری دارند و تعداد رخداد بیشتری را تجربه کرده‌اند، معرفی می‌شوند. پس از فراهم آوردن جمعیت و فراوانی داده‌ها و ویژگی‌های موجود، حال نوبت به محاسبه نرخ احتمال می‌رسد. محاسبه نرخ احتمال، کمک می‌کند تا درک عمیقی از قدرت یک ویژگی در میان دیگر ویژگی‌ها بدست آورد و بتوان براساس آن ویژگی ها را رتبه بندی کرد. رابطه محاسبه نرخ احتمال به صورت زیر تعریف می‌شود:

$$\text{Abs}\left(\text{Log}\left(\frac{P(Sense_1 \mid Collocation_i)}{P(Sense_2 \mid Collocation_i)}\right)\right)$$

بر اساس رابطه ارائه شده نسبت احتمال رخداد یک ویژگی‌خاص به شرطی که جمله هدف شامل معنی صحیح شماره یک باشد به احتمال رخداد همان ویژگی به شرطی که معنی صحیح غیر از معنی صحیح شماره یک را داشته باشد، محاسبه می‌شود. همچنین لگاریتم این رابطه محاسبه شده تا درک بهتری از مقادیر بدست آمده ارائه شود و تنها مقادیر مثبت خروجی این رابطه خواهد بود. بر اساس این رابطه می‌توان تخمین درستی از قدرت یک ویژگی ارائه داد که براین اساس ویژگی‌هایی که محتمل تر باشند نسبت به سایر ویژگی‌ها امتیاز بیشتری کسب کرده و مقدار نرخ احتمال بالاتری را حادث می‌شوند و به همین ترتیب ویژگی‌های کم اهمیت تر و نادر، مقادیر نرخ احتمال پایینی را کسب خواهند کرد که در نهایت از این مقادیر برای تصمیم‌گیری صحیح استفاده خواهد شد.

---

[6] Semi-Supervised

[7] Decision List

برای رهایی از مواجهه با احتمالات صفر از روش ضریب هموارساز[8] در این پژوهش استفاده شده است. این ضریب به عنوان پیش فرض احتمال رخداد هریک از ویژگی‌ها درنظر گرفته می‌شود. در صورتی که تطابق یک ویژگی با کلاس معنی خاصی بدون هیچ عضوی باشد. احتمال مربوط به آن طبق ضریب هموارساز درنظر گرفته می‌شود. به این صورت در تمام محاسبات تطابق ویژگی‌ها و کلاسهای معنی، ضریب هموارساز به رابطه‌ها ارائه شده اضافه می‌گردد تا امر محاسبه نرخ احتمالات را آسان سازد.

بر اساس این ساختار ویژگی‌ها به دو نوع منظم و نامنظم تقسیم بندی شده‌اند. در هر ویژگی کلمه هدف مشخص شده است و همچنین عبارتی به عنوان نشانه ویژگی تعیین شده است. نشانه ویژگی، بخشی است که الگوریتم استخراج ویژگی قصد دارد موجودیت آن را در میان نمونه اسناد ورودی پیدا کند. اگر ویژگی از نوع منظم باشد، موقعیت مکانی قبل و بعد نسبت به کلمه هدف ارائه شده است که به این معنا خواهد بود که نشانه ویژگی در فاصله‌های تعیین شده باید پیدا گردد. اگر ویژگی از نوع نامنظم باشد این فواصل اهمیتی پیدا نمی‌کند. در نهایت این چهارچوب شامل معنی صحیح کلمه هدف نیز خواهد بود که هر ویژگی به یک معنی منحصربه‌فرد اشاره خواهد کرد. با ارائه این چهارچوب هریک از دو نوع ویژگی اصلی در قالب یک جدول ویژگی‌ها قابل ارائه و استفاده در متون هستند. ویژگی‌های منظم به منظور پیدا کردن عبارات در محدوده معینی فعالیت می‌کنند و ویژگی‌ها نامنظم تنها به دنبال حضور یک عبارت بصورت نامحدود در بستر جمله و متن هستند.

در مرحله پایانی لیست تصمیم مناسب ساخته شده است و آماده استفاده برای شناسایی معنی صحیح کلمات هدف می‌باشد. بنابر مفهوم بنیادی لیست تصمیم، ویژگی‌هایی که در بالای این لیست جای دارند نمونه‌های قابل اعتمادتری برای ابهام زدایی معنی کلمه هدف می‌باشند. به عنوان ورودی لیست تصمیم از یک رشته کاراکتری مانند جمله یا عبارت استفاده می‌شود که شامل کلمه هدف نیز می‌باشد. پس از دریافت ورودی، از ابتدای لیست تصمیم و بالاترین ویژگی‌ها فرآیند شناسایی تطابق ویژگی بر روی ورودی صورت می‌گیرد. درصورتی که ویژگی مورد نظر در محتوای دریافتی ورودی یافت شد، معنی آن کلاس ویژگی به جمله ورودی نسبت داده می‌شود. در غیر اینصورت ویژگی بعدی موجود در لیست که دارای نرخ احتمال کمتری نسبت به ویژگی اول است مورد بررسی قرار می‌گیرد. این روند به همین منوال ادامه دارد تا زمانی که یک ویژگی با محتوای موجود در ورودی مطابقت داشته باشد. اگر تمام ویژگی‌های موجود در لیست تصمیم بررسی شود و هیچ کدام با عبارت ورودی تطابق نداشته باشند، نمونه ورودی غیر قابل تشخیص معرفی خواهد شد. این روند کمک می‌کند که بارمحاسباتی کمتری صرف شود و ویژگی‌های متداول در زبان زودتر از ویژگی‌های دیگر مورد بررسی قرار گیرند تا نتیجه با بارمحاسباتی کمتر و همچنین سریعتر حاصل شود.

مطابق شکل ۳-۲ ترسیم شده از نمای کلی یک سامانه یادگیری با روش نیمه نظارتی، پس از آنکه بخش نظارتی وظیفه خود را به خوبی انجام داد و نوع صحیح معنی کلمه را مشخص کرد، باید در مورد خروجی مسئله تصمیم گیری کرد. . استفاده از لیست تصمیم به طور مجرد، نمی‌تواند عدالت کامل را بین تمام ویژگی‌ها اعمال کند. همچنین به منظور ایجاد ساختار نیمه نظارتی از روش سه‌گانه آموزش [۳۹]در این پژوهش استفاده شده است. در این روش برای اینکه از خاصیت مجرد بودن پیکره کاسته شود، در ابتدا پیکره اصلی را به سه بخش مساوی تقسیم بندی می‌کند. این سه بخش به طور کاملا تصادفی ایجاد می‌گردند. هدف از این کار ایجاد پیکره‌هایی با پراکندگی طبیعی نسبت به تمام متون موجود در دنیای حقیقی می‌باشد. در پاره‌ای از اوقات این امکان وجود خواهد داشت که یک پیکره بیشتر از دو پیکره دیگر شامل اسناد باشد که به منظور توزیع متوازن نمونه‌ها این امر اجتناب ناپذیر است. هر یک از پیکره‌های تولید شده را با نمادهای S_1 ، S_2 و S_3 نام‌گذاری می‌شودکه اشاره به نمونه‌های تولید شده از مجموعه پیکره برچسب خوره با نماد L می‌باشد.

در روش یادگیری سه گانه، علاوه بر سه پیکره موجود، به سه عدد کلاس‌بند نیز نیاز است که هر سه مورد این کلاس‌بندها از یک روش یادگیری همراه با نظارت بر روی پیکره‌های موجود استفاده خواهند کرد. همانطور که پیش تر ذکر شد، در این پژوهش از روش یادگیری ماشین همراه با نظارت لیست تصمیم برای هر یک از این کلاس‌بندها استفاده خواهد شد. در نهایت سه نمونه کلاس‌بند که با روش لیست تصمیم آموزش داده شده‌اند نیز ایجاد می‌گردد. این نمونه‌ها با نمادهای C_1 ، C_2 و C_3 معرفی می‌شوند. برای آموزش هر یک از این کلاس‌بندها به ترتیب از پیکره‌های شماره یک، دو و سه استفاده خواهد شد. به عبارت دیگر پیکره آموزشی هر یک منحصر به فرد خواهد بود و شامل اسناد تکراری نخواهد بود. فرآیند آموزش لیست تصمیم به طور کامل برای هر سه کلاس‌بند موجود اتفاق می‌افتد و هریک تولید خروجی می‌کنند.

۴,۳. استخراج نتایج

اساس روش آموزش سه‌گانه ایجاد سه نمونه متفاوت پیکره منحصر به فرد و در پی آن ایجاد سه نمونه کلاس‌بند می‌باشد. بر طبق این روش هر یک از این سه عامل تصمیم‌گیرنده می‌توانند نظر متفاوتی را نسبت به یک داده ورودی داشته باشند و خروجی‌هایی متفاوت ایجاد کنند. از مستقل بودن خروجی‌های کلاس‌بند برای نتیجه گیری نهایی استفاده خواهد شد. اساس تصمیم‌گیری نهایی روش آموزش سه‌گانه رای و نظر اکثریت است. پس از دریافت ورودی که عبارت است از یک عبارت به زبان فارسی که دربرگیرنده کلمه هدف معین شده می‌باشد، هر یک سه لیست تصمیم موجود به صورت مستقل و بنابر لیست منحصر به فردشان راجع به معنی صحیح کلمه هدف نظر خود را اعمال می‌کنند. به این ترتیب اگر دو نظر یا بیشتر از سه نظر موجود مشابه هم بود، آنگاه رای اکثریت به عنوان معنی صحیح کلمه هدف درنظر گرفته خواهد شد. در صورتیکه هر سه نظر کلاس‌بند ها متفاوت از یکدیگر باشد، در آن صورت روش‌های متفاوتی را می‌توان به کار گرفت. می‌توان کلمه هدف را غیرقابل تصمیم‌گیری و شناسایی معرفی کرد و یا بررسی کرد که نرخ احتمال کدام یک از نظرات داده شده برای کلمه هدف بیشتر است و متعلق به کدام کلاس‌بند خاص می‌باشد. در این صورت آن ویژگی که دارای نرخ احتمال بیشتری باشد در موارد خاصی که نظر هر سه کلاس‌بند با یکدیگر متفاوت است به عنوان نظر نهایی روش آموزش سه‌گانه ارائه می‌گردد.

[8] Smoothing ratio

پس از اتمام عملیات آموزش سه‌گانه و دریافت خروجی، یک داده جدید که دارای برچسب صحیح معنی کلمه نیز می‌باشد به مجموعه دادگان اولیه اضافه شده است. این داده جدید را به پیکره اولیه اضافه کرده و بار دیگر تمام مراحل آموزش سه‌گانه تکرار خواهند شد. به عبارت دیگر دوباره پیکره به سه قسمت مساوی و تصادفی تقسیم شده و سه لیست تصمیم توسط این پیکره‌ها مورد آموزش قرار خواهند گرفت و در نهایت با اعمال یک عبارت ورودی شامل کلمه هدف که برچسب گذاری نشده است، نظر نهایی خود را درباره معنی صحیح کلمه هدف اعلام خواهد شد. به این ترتیب یک داده برچسب خورده معنایی دیگر به مجموعه دادگان اضافه خواهد شد. همین روند آموزش و تولید داده‌های جدید ادامه خواهد داشت تا زمانی‌که دیگر داده‌ی خام که بدون برچسب معنی صحیح کلمه هدف باشد، یافت نشود و تمام داده‌ها به طور کلی دسته بندی شوند. الگوریتم آموزش سه گانه در شکل (۲) ارائه شده است.

```
1: for i ∈ {1..3} do
2:       S_i ← bootstrap_sample(L)
3:       c_i ← train_classifier(S_i)
4: end for
5: repeat
6:       for i ∈ {1..3} do
7:          for x ∈ U do
8:             L_i ← ∅
9:             if c_j(x) = c_k(x)(j, k ≠ i)then
10:                  L_i ← L_i ∪ {(x, c_j(x)}
11:            end if
12:         end for
13:         c_i ← train_classifier(L ∪ L_i)
14:      end for
15: until none of c_i changes
16: apply majority vote over c_i
```

شکل (۲) الگوریتم آموزش سه گانه [۱۴]

۴. نتایج

چهارچوب ارائه شده توسط سه کلمه هدف مورد ارزیابی قرار گرفته است. نقش اساسی را در مسئله ابهام زدایی معنایی کلمات را انتخاب کلمات کاندید مبهم بازی می‌کند. این کلمات مبهم در تمام زبان‌های طبیعی زنده جهان موجود هستند. این امکان وجود دارد که در بعضی از زبان‌ها اشتراکی نیز میان این کلمات وجود داشته باشد. با توجه به زبان هدف این پژوهش که زبان فارسی می‌باشد، از سه کلمه هدف برای آموزش و ارزیابی روش‌های مطرح شده استفاده شده است. این کلمات از پراکندگی معنایی مناسبی برخوردارند و همچنین در متون فارسی بسیار متداول می‌باشند و از کلمات شاخص مبهم در زبان فارسی می‌باشند. این کلمات عبارتند از : شیر ، راست و تار. کلمه هدف "شیر" نیز یکی از پرتکرارترین کلمات مبهم زبان فارسی است.

جدول (۲) اسناد استخراج شده به تفکیک کلمات هدف و خبرگزاری‌های مرجع

خبرگزاری	شیر	راست	تار	مجموع
ایسنا	۴۹۲	۵۰۸	۸۰۲	۱۸۰۲
خبرآنلاین	۱۵۶۲	۵۴۵	۱۳۵	۲۲۴۲
مهر	۵۵	۷۶۲	۵۰۷	۱۳۲۴
مجموع	۲۱۰۹	۱۸۱۵	۱۴۴۴	۵۳۶۸

جدول (۳) کلاس‌های معنی تعریف شده برای کلمات هدف

کلمه هدف	برچسب کلاس‌های معنی
شیر	MILK – VALVE – LION
تار	MUSIC INSTRUMENT –AMBIGUOUS - STRING
راست	RIGHTSIDE – POLITICAL – TRUTH - FIRM

ویژگی‌های مرتبط با هر کلمه هدف در قالب یک ساختار از پیش تعیین شده ارائه شده است که یافتن و تشخیص آنها در میان اسناد به راحتی صورت پذیرد. دو نوع ویژگی منظم و نامنظم در این چهارچوب ارائه شده است که در صورت منظم بودن ویژگی جستجو تنها در مرز تعیین شده برای ویژگی انجام خواهد شد. اگر ویژگی در فاصله معین نسبت به کلمه هدف یافت شود، آنگاه آن جمله دلالت به معنی شاخص تعریف شده توسط ویژگی خواهد کرد. هر ویژگی تنها می‌تواند بر یک معنی شاخص منفرد اشاره داشته باشد. در جدول ۴ تعداد ویژگی‌های تعریف شده برای هر یک از کلمات هدف مشخص شده است.

جدول (۴) پراکندگی آماری ویژگی‌ها منظم و نامنظم تعریف شده به تفکیک کلمات هدف

کلمه هدف	ویژگی منظم	ویژگی نامنظم	مجموع
شیر	۳۴	۱۰۳	۱۳۷
راست	۳۲	۶۲	۹۴
تار	۷	۵۵	۶۲
مجموع	۷۳	۲۲۰	۲۹۳

ارزیابی الگوریتم‌های پیاده‌سازی شده توسط روش یادگیری نیمه نظارتی بسیار اهمیت دارد و با دیگر روش‌های یادگیری تفاوت دارد. یک نوع ارزیابی این روش‌ها توسط افراد خبره صورت می‌گیرد. به این معنی که داده‌های برچسب گذاری شده توسط کلاس‌بند نیمه‌نظارتی توسط یک فرد خبره مورد بررسی دقیق قرار گرفته و در نهایت میزان صحت عملکرد این سامانه توسط قضاوت فرد خبره تعیین می‌گردد. روش مرسوم و متداول دیگری که برای ارزیابی کلاس‌بندی‌های نیمه نظارتی به‌کار گرفته می‌شود، استفاده از داده پشتیبان [۹] می‌باشد .

جدول (۵) آمار اسناد و ویژگی‌های ذخیره شده در پایگاه‌داده به تفکیک کلمات هدف

کلمه هدف	تعداد کل اسناد	برچسب گذاری شده	بدون برچسب	تعداد کل ویژگی‌ها
شیر	۲۱۰۹	۶۹۱	۱۴۱۹	۱۳۷
راست	۱۸۱۵	۷۲۰	۱۰۹۵	۹۴
تار	۱۴۴۴	۷۲۲	۷۲۲	۶۲
مجموع	۵۳۶۸	۲۱۳۳	۳۲۳۶	۲۹۳

همانطور که پیش‌تر مطرح شد به منظور بررسی و ارزیابی روش پیاده سازی شده ابهام زدایی معنایی کلمات، در این پژوهش از مجموعه داده پشتیبان استفاده شده است. اسناد تشکیل دهنده این مجموعه از میان اسناد برچسب‌گذاری شده بصورت تصادفی انتخاب می‌شوند. بیست درصد از مجموعه اسناد دارای

[9] Heldout Data

برچسب معنایی به عنوان مجموعه پشتیبان درنظر گرفته می‌شوند. با استفاده از این روش ارزیابی، می‌توان درک درستی از سطح عملکرد واقعی سامانه طراحی و پیاده‌سازی شده ابهام زدایی معنایی کلمات بدست آورد. در این پژوهش برای ارزیابی از معیارهای صحت، دقت و بازیابی استفاده شده است. محاسبه هریک از این معیارها براساس یک رابطه از پیش تعریف شده صورت می‌پذیرد. البته ذکر این نکته ضروری است که معیارهای مطرح شده بصورت پیش‌فرض برای ارزیابی کلاس‌بندهای دودویی کاربرد دارند و برای مجموعه دادگانی که بیش از دو کلاس را به عنوان کلاس‌های نهایی خود می‌پذیرند نیاز به تغییر در روابط محاسباتی دارند .

اسناد موجود در این پژوهش نیز از این قاعده مستثنی نیستند و برای هر کلمه هدف موجود در سند بین سه و چهار کلاس هدف تعریف شده هستند. یک روش متداول محاسبه شاخص‌های ارزیابی درصورت وجود چندین کلاس هدف استفاده از میانگین‌گیری می‌باشد. به این منظور با هر یک از کلاس‌های موجود بصورت مستقل برخورد خواهد شد. به عبارت دیگر با هرکلاس بصورت دودویی برخورد می‌شود که داده‌ی حاصل بررسی می‌شود که عضو کلاس مطرح شده می‌باشد و یا خیر. داده‌های موجود در مجموعه پشتیبان در سامانه ابهام زدایی معنی کلمات اعمال می‌شوند و پس از اعلام نظر توسط سامانه، بر اساس هر کلاس موجود در کلمه هدف بصورت جداگانه شاخص‌های ارزیابی محاسبه می‌گردد. پس از بدست آمدن شاخص‌های منحصر به هر کلاس، می‌توان از آنها میان‌گیری کرد و حاصل را به عنوان شاخص‌های نهایی ارزیابی سامانه ابهام زدایی معنی کلمات در کلمه هدف تعیین شده ارائه داد .

یک راه‌حل بهبود یافته برای محاسبه شاخص‌های ارزیابی استفاده از میانگین وزن‌دار می‌باشد. به این ترتیب می‌توان درک درست‌تری نبست به میانگین‌گیری ساده در رابطه با عملکرد سامانه ابهام زدایی معنی کلمات را بدست آورد. براساس فراوانی هریک از کلاس‌های موجود در مجموعه داده پشتیبان وزنی برای هر کلاس تعریف می‌شود که در میانگین‌گیری نهایی از شاخص‌های هر کلاس این وزن در مقادیر شاخص‌ها ضرب شده و سپس عمل میانگین‌گیری صورت می‌پذیرد.

جدول (۶) شاخص‌های معیار ارائه شده برای ابهام‌زدایی نیمه نظارتی کلمه هدف "شیر"

سهم داده پشتیبان	سهم داده آموزش	دقت	بازیابی	صحت
۱۰٪	۹۰٪	۹۹,۲۸	۹۵,۷۱	۹۶,۲۴
۲۰٪	۸۰٪	۹۸,۷۲	۹۵,۶۸	۹۶,۵۶
۳۰٪	۷۰٪	۹۸,۴۳	۹۲,۳۰	۹۵,۶۸

جدول (۷) شاخص‌های معیار ارائه شده برای ابهام‌زدایی نیمه نظارتی کلمه هدف "راست"

سهم داده پشتیبان	سهم داده آموزش	دقت	بازیابی	صحت
۱۰٪	۹۰٪	۹۸,۳۹	۸۶,۱۱	۹۳,۶۱
۲۰٪	۸۰٪	۹۲,۸۴	۷۸,۴۷	۹۰,۹۳
۳۰٪	۷۰٪	۹۵,۲۸	۸۱,۴۸	۹۱,۰۹

جدول (۸) شاخص‌های معیار ارائه شده برای ابهام‌زدایی نیمه نظارتی کلمه هدف "تار"

سهم داده پشتیبان	سهم داده آموزش	دقت	بازیابی	صحت
۱۰٪	۹۰٪	۹۸,۶	۹۴,۵۲	۹۶,۷۹
۲۰٪	۸۰٪	۹۴,۰۳	۹۱,۰۳	۹۳,۷
۳۰٪	۷۰٪	۹۵,۰۷	۸۷,۰۹	۹۲,۴۷

در نهایت هرکلاسی که در مجموعه پشتیبان فراوانی بیشتری داشته باشد، وزن بیشتری خواهد داشت و در شکل‌گیری شاخص نهایی نقش بیشتری ایفا خواهد کرد. این روش محاسبه شاخص‌های ارزیابی با توجه به زبان طبیعی درنظر گرفته شده در این پژوهش بسیار مناسب می‌باشد. زیرا در بسیاری از کلمات

هدف، یک کلاس معنایی خاص متداول‌تر و پرتکرارتر از سایر کلاس‌های معنایی می‌باشد که همین موضوع انتخاب روش میانگین‌گیری وزنی را برای این پژوهش توجیه می‌کند. مقادیر شاخص‌های ارزیابی که با روش مطرح شده محاسبه گردیده‌اند به تفکیک کلمه هدف و مقدار داده پشتیبان در جداول ۶، ۷ و ۸ ارائه شده‌اند.

از ۵۳۶۸ سند استخراج شده از خبرگزاری‌های مهر، ایسنا و خبرآنلاین تعداد ۱۴۴۴ سند شامل کلمه هدف "تار"، ۱۸۱۵ سند شامل کلمه هدف "راست" و ۲۱۰۹ سند کلمه هدف "شیر" را در خود جای داده‌اند. داده‌های موجود برای اینکه قابلیت استفاده به عنوان داده‌های آموزش و ارزیابی را داشته باشند برچسب‌گذاری شده‌اند. در نهایت ۲۱۳۳ سند با موفقیت برچسب‌گذاری شده‌اند که تعداد ۶۹۱ سند مربوط به کلمه هدف "تار"، ۷۲۰ سند مربوط به کلمه هدف "راست" و ۷۲۲ سند برچسب‌گذاری شده مربوط به کلاس‌های معنایی کلمه هدف "شیر" می‌باشد. اسناد بدست آمده از میان ۱۲۹۴ صفحات خبری مربوط به خبرگزاری‌ها از بستر وب استخراج شده اند. به طور میانگین از هر خبر تعداد ۴ سند شامل کلمه هدف بدست آمده‌است.

سهم داده‌های برچسب‌گذاری شده نسبت به داده‌ها خام برای کلمات هدف "تار"، "راست" و "شیر" به ترتیب برابر است با ۴۷٪ ، ۴۰٪ و ۳۴٪ می‌باشد. برای شناسایی صحیح این کلمات هدف در میان اسناد تعداد ۲۹۳ ویژگی تعریف شده است که ۲۲۰ ویژگی از نوع نامنظم و ۷۳ ویژگی نیز از نوع منظم می‌باشند. ۶۲ ویژگی برای شناسایی کلمه هدف "تار"، ۹۴ ویژگی برای شناسایی کلمه هدف "راست" و ۱۳۷ویژگی به منظور یافتن معنی صحیح کلمه هدف "شیر" تعریف شده‌اند. در میان داده‌های برچسب‌گذاری شده برای کلمه هدف "تار" تعداد ۴۳۶ مورد دارای برچسبmusic، 240 مورد برچسب string و ۴۶ مورد دارای برچسب hide می‌باشند. داده‌های برچسب‌گذاری شده کلمه هدف "راست" دارای ۴ کلاس معنایی می‌باشند که از این میان تعداد ۳۵۴ مورد دارای برچسب political، تعداد ۲۴۴ مورد دارای برچسب rightside، 101 مورد دارای برچسب truth و ۲۱ مورد برچسب firm را اتخاذ کرده‌اند. کلمه هدف شیر نیز دارای سه کلاس معنایی تعریف شده می‌باشد که در این میان ۶۱۲ نمونه دارای برچسب milk، 74 نمونه دارای برچسب lion و در نهایت ۵ نمونه دارای برچسب valve می‌باشند. متداول‌ترین کلاس معنی در کلمه هدف "تار" ، music با سهم ۶۰٪ از کل نمونه‌های موجود می‌باشد. پرتکرارترین کلاس معنی در کلمه هدف "راست"، political با سهم ۴۹٪ از کل نمونه‌های موجود می‌باشد و بیشترین تکرار کلاس معنی در بین نمونه‌های کلمه هدف "شیر" متعلق به milk با سهم ۸۸٪ از تمام نمونه‌های موجود برچسب گذاری شده این کلمه هدف می‌باشد.

در روند آزمایش میزان سهم داده‌های پشتیبان متغیر انتخاب شده است تا درک صحیحی از عملکرد سامانه ارائه شود. با افزایش میزان سهم داده‌های پشتیبان اندکی از معیارهای شاخص کاسته می‌شود اما این میزان به هیچ وجه قابل توجه نبوده و این نشان از عملکرد پایدار سامانه طراحی شده در مواجه با حجم بالای داده‌های آزمایشی می‌باشد. معیار سهم بندی در این پژوهش ۲۰٪ مربوط به داده‌های پشتیبان و ۸۰٪ داده‌های آموزشی می‌باشد که بر این اساس معیارهای شاخص دقت، بازیابی و صحت به طور میانگین به ترتیب برابر ۹۵،۳۱۱ ، ۸۸،۳۹۶ و ۹۳،۷۳۶ می‌باشد. به عبارت دیگر ۹۵ درصد نمونه‌هایی که برچسب‌گذاری شده اند به کلاس معنایی صحیح کلمه هدف در جمله اشاره دارند.

جدول (۹) مقایسه شاخص‌های معیار کلمات هدف با یکدیگر

کلمه هدف	سهم داده پشتیبان	سهم داده آموزش	دقت	بازیابی	صحت
شیر	۲۰٪	۸۰٪	۹۸,۷۲۴	۹۵,۶۸۳	۹۶,۵۶۹
راست	۲۰٪	۸۰٪	۹۲,۸۴	۷۸,۴۷۲	۹۰,۹۳۸
تار	۲۰٪	۸۰٪	۹۴,۰۳۷	۹۱,۰۳۴	۹۳,۷۰۳

جدول (۱۰) میانگین شاخص‌های معیار بدست آمده به منظور ابهام زدایی کلمات در زبان فارسی

سهم داده پشتیبان	سهم داده آموزش	دقت	بازیابی	صحت
۱۰٪	۹۰٪	۹۸,۷۶	۹۲,۱۱۵	۹۵,۵۵
۲۰٪	۸۰٪	۹۵,۳۱۱	۸۸,۳۹۶	۹۳,۷۳۶
۳۰٪	۷۰٪	۹۶,۲۶۶	۸۶,۹۶۲	۹۳,۸۴

در میان کلمات هدف انتخاب شده، بهترین نتیجه در اسناد شامل کلمه هدف "شیر" بدست آمده است. این اسناد با معیار صحت ۹۶,۵۶۹٪ بهترین عملکرد را از آن خود کرده‌اند. دلیل اصلی این موضوع محدود بودن کلمه هدف "شیر" به تنها سه کلاس معنایی بوده و هچنین نکته قابل توجه دیگر این کلمه اینست که ۸۸٪ از اسناد دارای یک کلاس معنایی خاص بوده‌اند که به تشخیص صحیح اسناد کمک بسزایی دارد. رتبه بعدی مربوط به کلمه هدف "تار" می‌باشد. کلاس‌های معنایی این کلمه از پراکندگی متعادل‌تری نسبت به کلمه هدف "شیر" برخوردارند. ۹۳,۷۳۶ درصد کلاس معنایی اسناد شامل کلمه هدف "تار" به درستی تشخیص داده شده‌اند. کم‌ترین نتایج این پژوهش در ابهام زدایی کلمه هدف "راست" بدست آمده است. علت اصلی این موضوع فراوانی کلاس‌های هدف کلمه "راست" می‌باشد. این کلمه دارای ۴ کلاس معنایی می‌باشد. علاوه بر متعدد بودن کلاس‌های معنایی درنظر گرفته شده برای این کلمه هدف، پراکندگی نمونه‌ها در این کلاس‌ها نیز متعادل می‌باشد که این موضوع تشخیص و رفع ابهام معنی کلمات را با مشکل روبه‌رو می‌کنند. در نهایت ٪ ۹۰،۹۳۸۱ از اسناد شامل کلمه هدف "راست" به درستی ابهام زدایی می‌شوند. در پایان عواملی که اثر مخرب در عملکرد صحیح سامانه ابهام زدایی معنی کلمات در زبان فارسی عبارتند از:

- کاراکترهای زائد موجود در اسناد که مانع از تشخیص صحیح کلمات هدف و استخراج ویژگی‌ها از اسناد خواهد شد
- عدم تعریف ویژگی‌های مناسب که تمام کلاس‌های معنایی کلمات هدف را پوشش دهند
- متعدد بودن کلاس‌های معنی کلمات هدف
- فراوانی متعادل کلاس‌های معنی کلمات هدف

در [۱۲]برای ابهام زدایی معنایی کلمات فارسی "اعمال" و "اشراف" از روش نیمه نظارتی استفاده شده است. نتایج بدست آمده در این پژوهش حاکی از این است که نمونه‌های مبهم کلمه هدف "اعمال" با معیار صحت ۸۴,۸۴٪ و بازیابی ۸۳,۸۳٪ ابهام زدایی شده‌اند. کلمه هدف "اشراف" نیز معیار صحت ۹۵,۸۷٪ و معیار بازیابی ۹۳,۸۱٪ را به خود اختصاص داده است. به منظور آموزش کلاس‌بندها در این پژوهش از ۱۵۰۰ نمونه استفاده شده است. پژوهش دیگری نیز بر روی ابهام زدایی معنایی کلمات فارسی صورت گرفته است که بر روی روش‌های نظارتی تمرکز دارد [۳۱]. در این پژوهش روش‌های بیز ساده و نزدیکترین همسایگی KNN برای ابهام زدایی ۴ کلمه هدف مبهم فارسی استفاده شده است. کلمات مبهم در نظر گرفته شده عبارتند از "تار" ، "شیر" ، "کرم" و "مهر" . معیار صحت به طور میانگین برای رفع ابهام این ۴ کلمه مبهم در روش بیز ساده برابر با ٪ ۵۷,۹ بوده است و همچنین در روش نزدیکترین همسایگی معیار صحت برابر با ۶۹,۱۴٪ بوده است. این نتایج بر روی پیکره ساده بدست آمده است اما در حالتی دیگر این روش‌ها بر روی یک پیکره تجزیه شده به ریشه کلمات اعمال شده است که در این حالت نیز معیار صحت در روش بیز ساده برابر با ۶۵,۱۳٪ بوده است و در روش نزدیکترین همسایگی معیار صحت مقدار ۷۷,۶۴٪ را به خود اختصاص داده است. با توجه به مقادیر کسب شده در روش ارائه شده در این پژوهش، معیارهای شاخص دقت، بازیابی و صحت به طور میانگین به ترتیب برابر ۹۵،۳۱ ، ۸۸،۳۹ و ۹۳،۷۳ می‌باشد که در مقایسه با دیگر روش‌های معدود انجام گرفته بر روی زبان فارسی قابل قبول می‌باشد. ذکر این نکته نیز ضروری است که پیکره استفاده شده و همچنین تعریف ویژگی‌ها در تعیین نتایج نهایی بسیار موثر است.

با توجه به اهمیت حوزه پردازش زبان طبیعی و رشد روز افزون پژوهشهای دخیل در این حوزه، این نیاز احساس می‌شود که توجه بیشتری باید به پژوهش های دخیل در حوزه پردازش زبان طبیعی فارسی صورت گیرد. از این رو برای ایجاد بستری مناسب برای پیشرفت و توسعه پردازش زبان طبیعی فارسی، ابتدا باید به از بین بردن مسائل و گلوگاه‌های موجود در این حوزه اقدام کرد. به این صورت می‌توان امکان فعالیت برای پژوهشگران پردازش زبان طبیعی فارسی را آسان تر کرد و این امکان را به وجود آورد که برنامه‌های کاربردی مناسبی با استفاده از روش‌های پردازش زبان طبیعی بوجود آید و دیگر حوزه‌های تحقیقاتی را بهره مند سازد.

از این رو پرداختن به مساله ابهام‌زدایی معنی کلمات می‌تواند نقش موثری در توسعه حال و آینده پردازش زبان طبیعی داشته باشد. پژوهش در این حوزه می‌تواند راهگشای تحقیقات بعدی باشد و راه حل مناسبی را برای یکی از گلوگاه های اصلی پردازش زبان طبیعی را فراهم آورد.

پیشنهادهای ارائه شده برای پژوهش‌های آینده در این حوزه به شرح زیر ارائه می‌گردد:

- فراهم آوردن مجموعه دادگان مناسب برای آموزش سامانه‌های ابهام زدایی معنی کلمات فارسی
- ارائه ویژگی‌های کلمات مناسب توسط زبان شناسان برای رفع ابهام کلمات مبهم فارسی
- ارائه یک سامانه پیش‌پردازش مناسب برای قطعه بندی کلمات و بهبود استخراج ویژگی‌ها
- استفاده از ریشه کلمات در تعریف ویژگی‌ها و روند ابهام زدایی معنی کلمات
- استفاده از ویژگی‌های پیشرفته نظیر دخیل کردن نقش کلمات در جمله به منظور بهبود تشخیص صحیح معنی کلمات هدف
- استفاده از شبکه‌های کلمات برای تعریف کلاس‌های معنی کلمات هدف و همچنین ویژگی‌های تمایز دهنده هر کلاس

۵. مراجع

[۱] فیلی هشام، سلطانی محمود، **" ابهام زدایی معانی کلمات با استفاده از روش‌های آماری"** ، پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران، مرکز توسعه فناوری نیرو،۱۳۸۸

[2] Liu H.; Teller V.; Friedman C., *"A multi-aspect comparison study of supervised word sense disambiguation"* , Journal of American Medical informatics association, 2004

[3] Fernández A.; Valdés C.; Claramunt R.; Batalla A., Jordi Tormo, "*Automatic acquisition of sense examples using Exretriever"* , 2004

[4] Pedersen T., "*A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation"* , 2000

[5] Borah, P.P.; Talukdar, G.; Baruah, A., "*Assamese Word Sense Disambiguation using Supervised Learning*", Contemporary Computing and Informatics (IC3I), 2014

[6] Fan D.; Lu Z.; Zhang R.; Li X., "*Word Sense Disambiguation method based on probability model improved by information gain"* , Intelligent Control and Automation, 2008

[7] Wai P., "*Myanmar to English verb translation disambiguation approach based on Naïve Bayesian classifier"*, Computer Research and Development (ICCRD), 2011

[8] Navigli R., "*Word sense disambiguation: A survey* ", ACM Computing Surveys (CSUR), 2009

[9] HEARST M., "*Noun homograph disambiguation using local context in large text corpora"***,** In Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora 1991

[10] YAROWSKY D,. "*Unsupervised word sense disambiguation rivaling supervised methods*", In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 1995

[11] MIHALCEA R,. "*Co-training and self-training for word sense disambiguation*", In Proceedings of the 8th Conference on Computational Natural Language Learning, 2004

[12] ABNEY S,. "*Understanding the Yarowsky algorithm*". *Computat* 2004.

[13] NG V,; CARDIE C., "*Weakly supervised natural language learning without redundant views"***,** In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics , 2003

[14] Li M.; Zhou Z.H., "*Tri-training: exploiting unlabeled data using three classifiers"*, IEEE Transactions on Knowledge and Data Engineering, 2005

[15] Riahi, N.; Sedghi, F., "*A Semi-Supervised method for Persian homograph Disambiguation"* , Electrical Engineering (ICEE), 2012 20th Iranian Conference on , vol., no., pp.748,751, 15-17 May 2012

[16] BijanKhan, M., "*The role of the corpus in writing a grammar: an introduction to a soft-ware"*, Iranian Journal of Linguistics,Vol. 19, No. 2. 2004

[17] AleAhmad A.; Amiri H.; Darrudi E.; Rahgozar M.; Oroumchian, F., "*Hamshahri: A standard Persian text collection"*, Journal of Knowledge-Based Systems, Vol. 22 No.5, p.382-387, Elsevier, July 2009

[18] Yarowsky D., "*Homograph disambiguation in speech synthesis"*, ESCAIIEEE Workshop on speech synthesis, 1994.