رفع ابهام معنایی کلمات فارسی
با استفاده از رویکرد نظارت شده الگوریتم‌های IBL

منیر خیرمند پاریزی*
دانشجوی کارشناسی ارشدکامپیوتر، دانشگاه آزاداسلامی، واحدسیرجان، ایران
پست الکترونیکی: Monir_Kheirmand@yahoo.com

رضا نورمندی پور
استادیار، گروه کامپیوتر واحدسیرجان، دانشگاه آزاداسلامی، واحدسیرجان،ایران
پست الکترونیکی: noormandi_r@iausirjan.ac.ir

چکیده
در هر زبان کلماتی مبهم وجود دارند که دارای معانی متفاوتی هستند. مسئلهٔ یافتن معنای صحیح کلمهٔ دارای معانی متعدد، از مسایل جاری در حوزه پردازش زبان‌های طبیعی محسوب می‌شود. انتخاب معنی درست ممکن است برای انسان بسیار واضح و آسان باشد ولی تشخیص این‌که کدام یک از معانی موجود برای یک کلمه بایستی در جمله خاص انتخاب شود برای ماشین دشوار است. چهار رویکرد متفاوت برای رفع ابهام معنایی کلمات وجود دارد: رویکرد مبتنی بر دانش، رویکرد تحت نظارت، رویکرد نیمه نظارتی و رویکرد بدون نظارت هستند. در میان این روش‌ها، ما روش یادگیری مبتنی بر نمونه[1](IBL) را که بر اساس رویکرد نظارتی است برگزیدیم. در این مقاله دو دسته ویژگی مبتنی بر کلمات موضوعی معرفی کردیم. دسته اول وجود و عدم وجود کلمات موضوعی را وابسته به وزن هر کلمه موضوعی تعیین می‌کند و دسته دوم مجموع وزن کلمات موضوعی در هر جمله را مشخص می‌کند. بعد از انجام عملیات پیش‌پردازش روی نمونه‌های ذخیره شده برای هر کلمه، دو ماتریس ویژگی استخراج می‌شود. هر ماتریس ویژگی به‌عنوان ورودی نرم‌افزار وکا در نظر گرفته شده است. سپس با استفاده از طبقه‌بندی کننده مبتنی بر نمونه و با روش اعتبار سنجی متقابل 10 تایی نتایج را ارائه می‌دهیم. سپس نتایج حاصل از هر ماتریس ویژگی را برای یافتن ویژگی بهینه مورد بررسی قرار می‌دهیم. همچنین روش خود را روی دیگر الگوریتم‌های یادگیری ماشین مورد بررسی قرار داده و نتایج را ارائه می‌کنیم. میانگین صحت عملکرد روش پیشنهادی 88/31٪ می‌باشد.

کلمات کلیدی: رویکرد نظارت شده، رفع ابهام معنایی، روش مبتنی بر نمونه، طبقه‌بندی، کلمات موضوعی.

1. مقدمه
رفع ابهام معنایی (WSD)[2] اصطلاحی است که به استخراج معنی صحیح و مناسب از واژه‌هایی که چندمعنایی هستند اطلاق می‌شود و این کار غیر ممکن نیست چرا که اگر چه کلمه‌ها ممکن است معانی زیادی داشته باشند طبیعتا تنها یکی از آن‌ها مناسب بافتی است که کلمه مورد نظر در آن ظاهر می‌شود. ابهام معنایی در بیشتر برنامه‌های کاربردی پردازش زبان طبیعی همچون بازیابی اطلاعات، متن‌کاوی و همچنین در زمینه‌های تحقیقاتی جدید مانند وب معنایی بیش از پیش مورد توجه قرار گرفته است.

---

* نویسندهٔ مسئول

[1] Instance Base Learning
[2] Word-Sense Disambiguation

اهمیت این نوع ابهام زدایی در ترجمهٔ ماشینی بیش از برنامه‌های کاربردی دیگر است و در واقع ابهام زدایی معنایی واژگان یکی از مراحل اصلی در ترجمهٔ ماشینی به شمار می‌آید. به‌عنوان مثال کلمه انگلیسی paper دارای معانی مختلفی در زبان فارسی (کاغذ، روزنامه، مقاله) است و انتخاب درست یکی از معانی درواقع عمل ابهام زدایی در ترجمهٔ ماشینی است. چهار رویکرد مختلف برای رفع ابهام معنایی کلمات وجود دارد. این رویکردها شامل رویکردهای نظارت شده، مبتنی بر دانش، نیمه نظارت و رویکرد غیر نظارتی است [1].

لسک[3] یکی از اولین محققانی بود که برای رفع ابهام تعاریف فرهنگ‌لغت قابل خواندن ماشین (MRD)[4] با استفاده از الگوریتم‌ها تلاش کرد، الگوریتم او به خوبی در میان محققان WSD شناخته شده است. این الگوریتم مبتنی بر این فرض است که کلمات در یک همسایگی داده شده گرایش به یک موضوع مشترک خواهند داشت و بنابراین اهداف آن رفع ابهام کلمات در عبارات کوتاه است. آن الگوریتم لسک اساس بر پایه رویکرد مبتنی بر دانش می‌باشد[2].

از جمله روش‌های نظارت شده روش لیست تصمیم است که در [3] الگوریتمی بر پایه لیست‌های تصمیم‌گیری پیشنهاد شده است. با استفاده از این روش مجموعه کاملی از ویژگی‌های همسایگی، ریخت‌شناسی و نحوی استفاده شده و دقت سیستم 78/1٪ گزارش شده است.

در [4] از ترکیب رویکرد شبکه عصبی که رویکردی نظارت شده است و مفهوم اطلاعات هم‌رخدادی برای رفع ابهام کلمات استفاده کرده و نشان داده که رفع ابهام معنایی می‌تواند با ترکیب چندین نشانه بهبود داده شود. در [5] روش مبتنی بر حافظه دررفع ابهام معنایی پیشنهاد شد که دقت این سیستم 75/1٪ گزارش شده است.

روش‌های نظارت شدهٔ مبتنی بر نمونه، تحت طبقه‌بندی مبتنی بر حافظه است چون مثال‌های داده آموزشی در حافظه ذخیره می‌شوند. برای این روش یادگیری الگوریتم Knn[5] (نزدیک‌ترین همسایه) استفاده می‌شود چون طبقه‌بندی دادهٔ آزمایش مبتنی بر شبیه‌ترین معانی k مثال ذخیره شده است. به منظور به دست آوردن مجموعه‌ای از نزدیک‌ترین همسایه‌ها، هر ویژگی از دادهٔ آزمایش با ویژگی مربوطه از هر مجموعه دادهٔ آموزشی مقایسه شده و فاصله بین آن‌ها محاسبه می‌شود. در [5] روش مبتنی بر حافظه برای رفع ابهام معنایی پیشنهاد شد که دقت این سیستم 75/1 ٪ گزارش شده است.

از فعالیت‌هایی که در زمینه رفع ابهام کلمات فارسی صورت گرفته می توان به[6] اشاره کرد. در این مقاله روش مبتنی بر پیکره و یک فرهنگ لغت برای امتیاز دهی هر معنی کلمهٔ مبهم پیشنهاد شده است. دقت میانگین این روش برای 15 کلمه مبهم 91/46٪ گزارش شده است. در [7] یک روش بیزی برای رفع ابهام معنایی از کلمات فارسی ارائه شده است که به منظور استفاده برای ترجمهٔ ماشینی از زبان فارسی به انگلیسی بوده است. در [8] روش رفع ابهام معنایی به منظور استفاده در ترجمه ماشینی (انگلیسی به فارسی) ارائه شده است. در [9] از مدل غیر نظارتی تخصیص پنهان دریکله استفاده شده است.

به‌طور کلی اکثر تحقیقات انجام گرفته روی رفع ابهام کلمات انگلیسی بوده و برای رفع ابهام کلمات فارسی متاسفانه تحقیقات محدودی انجام شده است. علاوه بر این با توجه به تحقیقات مقایسه‌ای[10] انجام شده دراین زمینه، روش‌های یادگیری با ناظر از جمله روش‌های انتخاب کردن مناسب‌ترین مفهوم می‌باشد که علیرغم نتایج بسیار خوبی که دارد استقبال چندانی در رفع ابهام کلمات فارسی از این روش‌ها نشده است و آن هم به علت مشکل آموزش دادن این روش‌ها و کمبود دادهٔ آموزشی مناسب برای زبان فارسی است.

از این رو ما در این تحقیق برای رفع ابهام معنایی کلمات فارسی به سراغ روش‌های یادگیری با ناظر رفته و برای آموزش، از بانک منحصر به فردی که تعداد بیشماری صفحه با جملات تقطیع شده در زمینه موضوعات مختلف دارد به‌عنوان دادهٔ آموزشی استفاده می‌کنیم.

---

[3] Lesk

[4] Machine-Readable Dictionary

[5] K-nearest neighbor

در این مقاله برای انجام رفع ابهام با استفاده از رویکرد مبتنی بر نمونه به‌طور کلی مراحل زیر انجام می‌پذیرد:

1- برای هر کلمه مبهم به تعداد معانی آن، رده وجود دارد. ما تعدادی متون حاوی کلمه مبهم را که معنی درست آن کلمه در آن متن برچسب خورده باشد انتخاب می‌کنیم. در واقع برای هر کدام از معانی کلمه مبهم تعدادی از متون به عنوان نمونهٔ دادهٔ آموزشی آن رده ذخیره می‌شود.

2- استخراج ویژگی از متون نمونه و تبدیل هر متن نمونه به بردار ویژگی متناظر با آن متن. انتخاب مناسب این ویژگی‌ها بسیار مهم است و در کارایی این روش تاثیر بسزایی دارد. به همین جهت، انتخاب بهینه از ویژگی‌ها و خصوصیاتی که باید داشته باشند یکی از کارهای اساسی این تحقیق است.

3- انتخاب معیار شباهت برای اندازه‌گیری شباهت بین نمونهٔ آزمون و نمونه‌های آموزشی.

4- استفاده از الگوریتم Knn جهت طبقه‌بندی دادهٔ آزمون جدید در یکی از رده‌های داده شده و در نهایت دستیابی به معنای صحیح کلمه مبهم

5- ارزیابی کارایی این روش و بیان نتایج حاصل از تحقیق. برای ارزیابی عملکرد روش پیشنهادی همانند بسیاری از روش‌های به کار گرفته شده در یادگیری ماشینی و داده کاوی از روش اعتبارسنجی متقابل k تایی استفاده می‌کنیم. ما k را در اینجا 10 در نظر می‌گیریم که برای هر کلمه مبهم، تمام نمونه‌های مربوطه به 10 قسمت مساوی تقسیم شده که 9 قسمت برای آموزش و یک قسمت باقی مانده برای دادهٔ آزمون مورد استفاده قرار می‌گیرد. و با استفاده از معیارهای دقت و فراخوانی ارزیابی را انجام می‌دهیم.

ادامهٔ مقاله به این صورت بخش بندی شده است: بخش دوم و سوم پیاده‌سازی روش پیشنهادی را با توصیف روش مبتنی بر نمونه، الگوریتم‌های اندازه‌گیری شباهت و پارامترها و متغیرهای مورد نیاز و مدل پیشنهادی بیان می‌کند. بخش چهار آزمایش‌ها و نتایج به دست آمده را نشان می‌دهد. و در بخش پنج مقایسه عملکرد روش پیشنهادی آورده شده است.

2. روش پیشنهادی

در این تحقیق قصد داریم با استفاده از روش یادگیری مبتنی بر نمونه و متون موجود حاوی کلمات مبهم به گونه‌ای معنای مناسب کلمه مبهم را تشخیص دهیم که از بالاترین دقت برخوردار بوده و از طرفی با انتخاب ویژگی‌های مناسب، معایب این روش را کاهش داده و روش مبتنی بر نمونه را بهبود دهیم. همان‌طور که قبلا اشاره کردیم در واقع ما برای تعیین معنی درست کلمه مبهم، باید سند را به ماتریسی از ویژگی‌ها تبدیل کنیم. مهم‌ترین مسئله در تشخیص رفع ابهام انتخاب ویژگی است که در این تحقیق بر روی ویژگی کلمات موضوعی متمرکز می‌شویم. در این بخش ابتدا در مورد الگوریتم‌های مبتنی برنمونه توضیح داده شده، سپس مشخصه های اصلی این الگوریتم‌ها آورده شده و در ادامه نحوهٔ انتخاب کلمات موضوعی و استخراج ویژگی‌ها بیان شده است.

2-1 الگوریتم یادگیری مبتنی بر نمونه

در روش یادگیری مبتنی بر نمونه (IBL) مثال‌ها را ذخیره می‌کنیم و هر گونه تعمیم تا مشاهده مثال جدید به تعویق می‌افتد. به همین دلیل این روش گاهی روش تنبل یاLAZY نامیده می‌شود. یادگیرنده‌های مبتنی بر نمونه یک نمونه را با مقایسهٔ آن با پایگاه دادهٔ نمونه‌های از پیش طبقه‌بندی شده، طبقه‌بندی می‌کند. فرض اساسی آن این است که نمونه‌های مشابه، طبقه‌بندی مشابه خواهند داشت[11].

الگوریتم‌های IBL از دسته‌بندی‌کنندهٔ الگوی (NN)[6] گرفته شده‌اند که در عین حال به ذخیره و استفاده از نمونه‌های منتخب برای پیش‌بینی دسته‌بندی می‌پردازد. روش یادگیری مبتنی بر نمونه دارای سه مشخصه اصلی است:

- **انتخاب نمونه‌ها برای ذخیره.** در این الگوریتم سعی می‌شود نمونه‌هائی ذخیره شوند که عمومی‌تر باشند. تشخیص این‌که آیا یک نمونه عمومیت دارد یا خیر، می‌تواند کار مشکلی باشد.

[6] Neighbor Nearest

- **تابع شباهت/ فاصله.** مشخص می‌کند که دو نمونه چقدر نزدیک به هم هستند. انتخاب این تابع می‌تواند بسیار مشکل باشد معیار شباهت یا فاصله میان دو نقطهٔ داده، یک چالش و موضوع مهم در روش‌های داده کاوی و کشف دانشی که نیازمند محاسبه شباهت هستند، می‌باشد. میزان نزدیکی بر حسب یک معیار فاصله یا شباهت تعریف می‌گردد. موفقیت اغلب سیستم های یادگیری به یک تابع شباهت یا فاصله خوب بستگی دارد که آن‌ها استفاده می‌کنند. در این مقاله از توابع فاصلهٔ اقلیدوسی، منهتن، کانبرا، چبیشف استفاده شده است.
- **تابع دسته‌بندی کننده.** تابعی است که با مشاهده یک مثال دسته‌بندی آن را تعیین می‌کند. برای این کار فاصله تا دیگر نمونه‌های آموزشی محاسبه می‌شود. K تا از نزدیک‌ترین همسایگان شناسایی می‌شود. از برچسب‌های ردهٔ نزدیک‌ترین همسایگان برای تعیین برچسب ردهٔ نمونه ناشناخته استفاده می‌کند (به‌عنوان مثال با گرفتن رای اکثریت). الگوریتم‌های دسته‌بندی کننده‌ای که دراین مقاله استفاده خواهیم کرد الگوریتم‌های دسته‌بندی کنندهٔ مبتنی بر نمونه است که شامل الگوریتم‌هایLWL[7], KStar, IBK, IB1 می‌باشد [12]. انتخاب مقدار K یکی از مراحل اصلی الگوریتم‌های IBL می‌باشد، اگرK خیلی کوچک باشد، نسبت به نوفه حساس خواهد بود و اگر K خیلی بزرگ باشد ممکن است یک همسایگی نقاطی از سایررده‌ها را نیز در برگیرد. مقادیر خوب برایK می‌تواند با اعتبارسنجی متقابل روی داده آموزشی پیدا شود[13].

2-2 انتخاب کلمات موضوعی

کلمات موضوعی کلماتی هستند که در یک سند بیش از سایر کلمات تکرار شده اند که به احتمال زیاد این کلمات منعکس کنندهٔ موضوع اصلی متن ورودی هستند[14].بعد از انجام عملیات پیش‌پردازش روی متن، هر کلمه با میزان تکرار آن در سند گروه بندی می‌شود و سپس به ترتیب نزولی مرتب می‌شوند. تعداد کلمات موضوعی انتخاب شده در این تحقیق 5 در صد از کل کلمات مبهم در هر حوزه معنایی خود است.

2-3 نحوه استخراج ویژگی‌ها

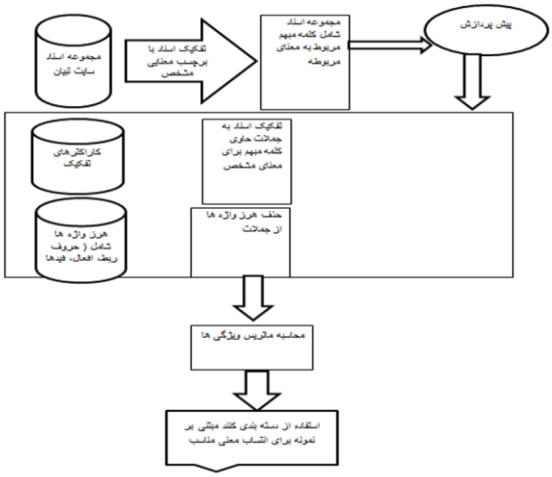
ما برای انجام آزمایشات خود دو دسته ویژگی استخراج کردیم که به صورت زیر است: دسته اول ویژگی‌ها حضور و عدم حضور کلمات موضوعی با در نظر گرفتن وزن آن‌ها است. برای محاسبه این ویژگی به این صورت عمل می‌کند که ابتدا کلمات موجود در هر جمله با هر یک از کلمات موضوعی ذخیره شده در فایل متنی مقایسه شده و در صورت وجود حوزهٔ آرایه آن کلمه موضوعی صفر می‌شود و در غیر این‌صورت عدد 1 برای آن کلمه موضوعی قرارمی‌گیرد. سپس با استفاده از الگوریتم بهره اطلاعاتی[8] در وکا ویژگی‌ها که در اینجا همان کلمات موضوعی است رتبه بندی شدند. الگوریتم بهره اطلاعاتی به این صورت عمل می‌کند که رتبه بندی با اختصاص وزن بین صفر تا 1 داده می‌شود که ویژگی با وزن بالاتر رتبه بالاتری را نشان می‌دهد. و هر چه عدد به سمت صفر میل می‌کند رتبه ویژگی به سمت آخر می‌رود. مقدار هر ویژگی با توجه به حضور و عدم حضور کلمه موضوعی (0،1) در وزن مربوطه ضرب شده و آن عدد به مقدار ویژگی منتسب می‌شود. بنابراین، اگر کلمه موضوعی حضور نداشته باشد که همان مقدار صفر به آن منتسب می‌شود ولی اگر کلمه موضوعی وجود داشته باشد مقدار وزن آن به آن ویژگی منتسب می‌شود. به عبارت دیگر، متن ما که حاوی جملات کلمه مبهم است به ماتریسی تبدیل می‌شود که سطرهای آن جملات و ستون‌ها حضور یا عدم حضور کلمات موضوعی انتخاب شده با در نظر گرفتن وزن در آن جمله را نشان می‌دهد. آخرین ستون در ماتریس نوع برچسب کلاسی را مشخص می‌کند. در مورد ویژگی دوم به این صورت است که این ویژگی مشابه ویژگی اول می‌باشد با این تفاوت که بعد از به دست آمدن ضرب وزن هر ویژگی در حضور و عدم حضور آن ویژگی مجموع ضرب آن‌ها برای هر سطر (جمله) محاسبه می‌شود و به مقدار آن ویژگی اختصاص داده می‌شود. بنابراین ماتریس ویژگی مجموع وزن کلمات موضوعی شامل دو ستون است که یکی مربوط به مجموع ضرب مقدار ویژگی در وزن مربوطه است و دیگری مربوط به برچسب رده‌ای جمله است.

---

[7] Locally Weighted Learning
[8] Information Gain

3. مدل پیشنهادی

روش پیشنهادی ما به این صورت عمل می‌کند که ابتدا اسناد و متون حاوی کلمه مبهم مورد نظر را از مجموعهٔ داده خود استخراج می‌کنیم. برای هر معنی کلمه مبهم مورد نظر، اسناد استخراج شده به‌طور جداگانه در فایل‌های مربوط به آن معنی قرار می‌گیرند. در مرحلهٔ بعد برنامه ما با خواندن هر فایل مربوط به معنی کلمه مبهم مورد نظر، فقط جملاتی را که شامل آن کلمهٔ مبهم می‌شوند استخراج کرده و استخراج ویژگی‌ها از روی این جملات انجام می‌گیرد. بنابراین ما برای هر برچسب معنایی کلمه مبهم، یک فایل حاوی بردارهای ویژگی مربوط به آن معنی را داریم. مجموع این فایل‌های ویژگی به عنوان داده ورودی در قالب فایل arff وارد نرم‌افزار وکا می‌شود. در نهایت با انتخاب پارامترهای مربوط به الگوریتم مبتنی بر نمونه، طبقه‌بندی و ارزیابی انجام می‌شود. ساختار کلی سیستم پیشنهادی به صورت شکل شماره 1 می‌باشد. در ادامه به توضیح هر یک از قسمت‌های شکل 1 می‌پردازیم.



شکل 1: ساختار کلی روش پیشنهادی

3-1 انتخاب اسناد از مجموعهٔ داده. برای انتخاب اسناد از وبگاه تبیان استفاده می‌کنیم. در این وبگاه انبوهی از مقالات در زمینه‌های موضوعی (دین، زندگی، فرهنگ، جامعه) وجود دارد. در این تحقیق تعدادی از کلمات مبهم که بیشترین تکرار را در اسناد این وبگاه دارند استخراج می‌کنیم. سپس مجموعه مقالات حاوی کلمه مبهم را استخراج کرده، و آن‌ها را در فایل‌های مجزا برای هر معنی کلمه مبهم جمع‌آوری می‌کنیم. برای مثال برای کلمه مبهم شیر که دارای سه معنی مختلف(Milk، Lion، Valve) است، ما سه فایل برای هر معنی ایجاد کردیم که در هر فایل متون حاوی کلمه شیر با آن معنی مربوطه گردآوری شده است.

3-2 تفکیک جملات بعد از این که متن به مجموعه‌ای از پاراگراف‌ها تفکیک شد، به سراغ تفکیک جملات داخل هر پاراگراف می‌رویم. جملات از طریق حائل‌هایی مثل(".", "؛","!","؟", " ؟"). جدا می‌شوند. چالشی که در این نوع جداسازی وجود دارد این است که برخی از واژه‌ها ممکن است یک جمله به حساب آیند در صورتی که این‌طور نیست. برای مثال، نقطه بین حروف مانند (ه.ق) و یا نقطه بین اعداد (1.2) یا (...) که برای نشان دادن سایر موارد به کار می‌رود هم به صورت یک جمله در نظر گرفته می‌شود. برای مثال، سه نقطه برای اشاره به عبارت و غیره به صورت سه جمله محاسبه می‌شود که کاملا نادرست است و اگر تعداد این عبارات در متن زیاد باشد برای شناسایی ویژگی‌ها مشکل ایجاد می‌کند. سیستم ما قادر به تشخیص نقاط پایانی جمله با نقاط بین حروف و اعداد و غیره است که این کار با کد برنامه نویسی#c انجام گرفته است. دراین سیستم عنوان‌ها و زیرعنوان‌های موجود در متن هم به صورت یک جملهٔ مجزا در نظر گرفته شده است.

3-3 تفکیک کلمات تفکیک کلمات در سیستم ما با علائمی مثل فضای خالی بین دو کلمه، علامت خط جدید، ",", ";"، "/"، ".", "؟" صورت می‌گیرد که البته برای کلمات مرکّب هم در این زمینه چالش‌هایی وجود دارد. برای مثال واژهٔ «سیستم‌عامل»

اگر با فاصله نوشته شود سیستم به عنوان یک کلمه و عامل به عنوان یک کلمه دیگر شناخته می‌شود که جالب نیست. برای بهبود روش باید سیستم قادر به شناسایی کلمات مرکب نیز باشد.

4-3 حذف هرزواژه‌ها برخی واژه‌ها در همه متون با تکرار زیاد وجود دارد. در واقع این واژه‌ها، واژه‌هایی مثل ضمایر، قیود، حروف اضافه و ربط هستند که تاثیری در ارزش محتوایی متن ندارند. به این واژه‌ها هرزواژه گفته می‌شود. بعد از تعیین حدود جملات باید هرزواژه‌ها را در جمله شناسایی و آن‌ها را حذف کنیم که ما فهرست کاملی از هرزواژه‌ها را با جستجو در سایر مقالات و متون جمع‌آوری کرده و از آن برای حذف هرزواژه‌ها در متن استفاده می‌کنیم. هرزواژه‌ها در فایل متنی ذخیره شده است که با خواندن تک تک کلمات ورودی و مقایسه آن با هرزواژه‌های داخل فایل اگر کلمه‌ای در ورودی وجود داشت که جزو هرزواژه‌ها در فایل بود آن کلمه با یک رشته خالی جایگزین می‌شود. بدین ترتیب کلیه هرزواژه‌ها در متن ورودی حذف خواهند شد.

بعد از تفکیک متن به مجموعه‌ای از جملات، نوبت به استخراج ویژگی‌ها از جملات متن می‌رسد که در واقع بخش اصلی کار پردازش است و با توجه به توضیحات بخش قبل ماتریس ویژگی‌ها را ایجاد می‌کنیم. سپس ماتریس ویژگی‌ها به عنوان ورودی برنامهٔ طبقه‌بندی مبتنی بر نمونه دریافت شده و عملیات طبقه‌بندی و ارزیابی با ابزار وکا انجام می‌گیرد.

4.آزمایش‌ها

در این قسمت ابتدا اطلاعات مربوط به جمع‌آوری داده بعد از عملیات مربوط به پیش‌پردازش متن نشان داده شده است و سپس ماتریس‌های ویژگی را به دست آورده، فایل‌های ویژگی به‌دست آمده را به قالب arff تبدیل کرده و به‌عنوان ورودی برنامهٔ وکا جهت اجرای الگوریتم طبقه‌بندی مربوطه به کار می‌بریم. همان‌طور که گفته شد ما در این تحقیق دو فایل ویژگی را برای هر کلمه مبهم استخراج کردیم. این فایل‌های ویژگی هر کدام به‌طور جداگانه وارد برنامه می‌شوند. طبقه‌بندی شده و نتایج مورد ارزیابی قرار می‌گیرند.

4-1 آماده سازی مجموعهٔ داده

مجموعهٔ کلمات مبهم همراه با تعداد تکرار که از وبگاه تبیان استخراج کرده‌ایم به صورت جدول 1 می‌باشد:

جدول 1: جزئیات کلمات مبهم مورد استفاده در آزمایش‌ها

کلمه مبهم	معنی	تعدادجملات حاوی کلمه مبهم	تعداد کلمه مبهم
شیر	Milk	422	460
	Lion	219	244
	Valve	115	121
مهر	Love	175	250
	Punch	159	191
	Solar Month	232	334
نفس	Ego	545	564
	Air	156	202
سیر	Garlic	292	485
	Journey	126	173
	Opposite Of Hungry	128	156

و به‌عنوان نمونه مجموعهٔ کلمات موضوعی در حوزه معنایی مهر به صورت جدول 2 می‌باشد.

جدول 2: کلمات موضوعی برای واژه مبهم «مهر»

مهربابرچسب معنایی 1	تعداد تکرار	مهربابرچسب معنایی 2	تعداد تکرار	مهر با برچسب معنایی 3	تعداد تکرار
مسکن	122	نامه	37	ماه	167
محبت	59	موم	34	روز	62
خبرگزاری	50	معنای	30	سال	43
عشق	28	علی	26	اول	40
افزایش	26	امضا	25	بهمن	39

همان‌طور که گفته شد در این مقاله دو فایل ویژگی را برای هر کلمه مبهم استخراج کردیم و 5 در صد کلّ کلمات در هر حوزهٔ معنایی را به عنوان تعداد کلمات موضوعی که براساس آن ماتریس ویژگی‌ها ساخته می‌شود در نظر می‌گیریم. به‌عنوان مثال، کلمه مبهم مهر در سه حوزه معنایی 1 و 2و 3 به ترتیب 250، 191 و 334 مرتبه آمده است که بعد از محاسبه 5 درصد آن‌ها به ترتیب 12، 9، و 16 کلمه برای سه حوزه معنایی می‌باشد. بنابراین ما برای کلمه «مهر» به‌طور کلی 37 کلمه موضوعی انتخاب کردیم.

بنابراین ماتریس ویژگی‌های ما در این نمونه، ماتریسی است که 38 ستون دارد که 37 ستون آن مربوط به کلمات موضوعی واژه مهر است و یک عدد مربوط به برچسب رده است و سطرها تعداد نمونه‌ها و جملات را مشخص می‌کند که در این تحقیق در حوزه معنایی مهر 556 نمونه را انتخاب کردیم.

بعد از آماده شدن ماتریس‌های ویژگی نوبت به طبقه‌بندی آن‌ها می‌رسد. همان‌طور که گفته شد در این مقاله از طبقه‌بندی کننده مبتنی بر نمونه برای رفع ابهام معنایی استفاده می‌کنیم. بنابراین در قسمت طبقه‌بندی کنندهٔ برنامه وکا، الگوریتم LAZY را انتخاب می‌کنیم. در دستهٔ LAZY الگوریتم‌هایIBK، Kstar و LWL قرار دارند. ما هر کدام از این الگوریتم‌ها را برای دو دسته ویژگی انتخاب شده روی هر کلمه مبهم آزمایش کرده‌ایم. هر کدام از این الگوریتم‌ها پارامترهای خاص خود را دارند که باید قبل از انجام آزمایش، آن‌ها را تنظیم کنیم. ما برای هر کلمه تنظیمات پارامترهای مختلف این الگوریتم‌ها را انجام دادیم و نتایج را برای هر کدام مشاهده کردیم. نتیجه مشاهدات را در مورد چند پارامتر مهم بدین صورت بیان می‌کنیم که برای الگوریتم IBK، ما آزمایش‌ها را با مقادیر Knn 1، 3 و 5 با هر کدام از الگوریتم‌های جستجوی نزدیک‌ترین همسایه موجود در وکا و هر کدام از توابع فاصله موجود روی هر کدام از کلمات مبهم انجام دادیم. مشاهدات نشان داد که توابع فاصله و الگوریتم‌های جستجوی نزدیک‌ترین همسایه تاثیری در نتایج روش پیشنهادی ما ندارد. ولی نتایج با Knn با مقدار 1 کمی بهتر از سایر مقادیر عمل می‌کند. همچنین برای الگوریتم Kstar تنها تعیین مقادیر مختلف برای پارامتر global Blend در نتایج تاثیر می‌گذارد که انتخاب مقدار 1 بهترین نتایج را داشته است.

4-2 نتایج

نتایج به دست آمده بر روی دو دسته از ویژگی‌ها با سه الگوریتم طبقه‌بندی‌کنندهٔ مبتنی بر نمونه به شرح جدول 3 و 4 می‌باشد.

جدول 3: نتایج طبقه‌بندی الگوریتم‌های LAZY روی کلمات مبهم(ورودی ماتریس ویژگی نوع اول)

کلمه مبهم الگوریتم	مهر	شیر	نفس	سیر	میانگین
IBK	83.21	84.19	96.51	86.14	87.51
KStar	83.71	85.32	97.45	86.78	88.31
LWL	83.21	84.19	96.51	86.14	87.51

جدول4: نتایج طبقه‌بندی الگوریتم‌های LAZY روی کلمات مبهم(ورودی ماتریس ویژگی نوع دوم)

کلمه مبهم الگوریتم	مهر	شیر	نفس	سیر	میانگین
IBK	66.24	69.02	75.18	72.13	70.64
KStar	68.37	69.87	76.75	73.20	72.04
LWL	66.25	69.02	75.18	72.13	70.64

نتایج به دست آمده از آزمایش‌ها نشان می دهد ویژگی اول دقت بالاتری نسبت به ویژگی نوع دوم داشته است. بنابراین ما همان ویژگی اول را انتخاب می‌کنیم. از طرف دیگر، در میان سه الگوریتم از دسته الگوریتم‌های مبتنی بر نمونه، نتایج حاصل از الگوریتم kstar با توجه به تنظیم پارامتر global blend کمی بهتر از دو الگوریتم دیگر عمل کرده است. بنابراین ما الگوریتم kstar و دسته ویژگی نوع اول را معیار برای روش خود قرار داده و نتایج حاصل از آن را با روش‌های دیگر مقایسه می‌کنیم.

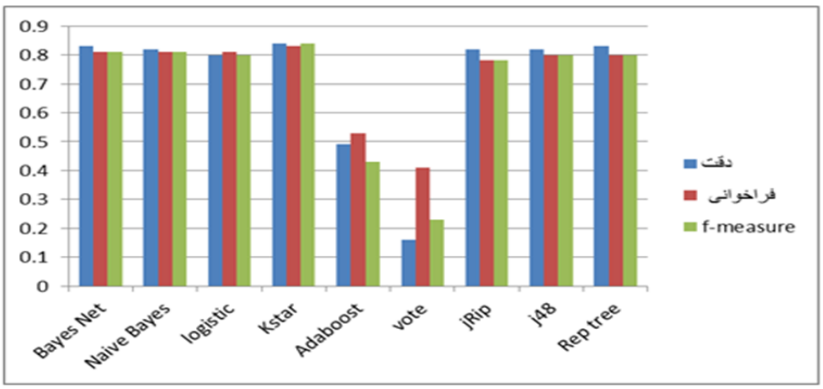
5. مقایسه عملکرد روش پیشنهادی

ما در این قسمت عملکرد روش پیشنهادی خود را روی دیگر الگوریتم‌های طبقه‌بندی بررسی می‌کنیم. در قسمت طبقه‌بندی‌کنندهٔ نرم‌افزار وکا از هر دسته طبقه‌بندی‌کننده (مانند bayes، functions، meta، rules،trees) یک یا دو الگوریتم را به‌عنوان نماینده آن دسته انتخاب کردیم و روش پیشنهادی را روی آن‌ها پیاده‌سازی کردیم. و برای همه الگوریتم‌ها از روش10 fold–cross Validation در ارزیابی استفاده شده است.

نتایج ارزیابی در جدول 5 نشان داده شده است. در شکل 2 نمودار میله‌ای بر حسب معیارهای ارزیابی دقت، فراخوانی و F-measure برای کلمه «مهر» نشان داده شده است.

جدول 5: نتایج عملکرد روش پیشنهادی روی روش‌های طبقه‌بندی دیگر (بر اساس پارامتر ارزیابی درصد طبقه‌بندی درست)

کلمه مبهم الگوریتم	مهر	شیر	نفس	سیر	میانگین
BayesNet	81.27	84.07	85.65	80.78	82.94
NaiveBayes	80.91	84.07	85.35	80.75	82.77
Logistic	80.21	84.07	85.43	82.12	82.95
AdaboostM1	53.18	75.12	78.19	73.38	69.96
Vote	40.98	75.12	76.۲۳	68.65	65.99
JRip	78.97	81.88	83.20	81.41	81.36
J48	81.27	82.99	86.۷۲	82.۶۴	83.40
Rep Tree	80.91	82.58	86.۰۹	82.09	82.91



شکل 2: نمودار میله ای برای مقایسه روش پیشنهادی روی الگوریتم‌های طبقه‌بندی مختلف مربوط به کلمه «مهر»

همان‌طور که شکل2 نشان می‌دهد سه معیار دقت، فراخوانی و f-measure در الگوریتم مبتنی بر نمونه kstar نسبت به سایر روش‌های طبقه‌بندی بالاتر است. بررسی نتایج حاصل از الگوریتم‌های طبقه‌بندی مختلف در شکل 2 و جدول 5 نشان می دهد که الگوریتم‌های vote و آدابوست بدترین دقت را داشته اند که هر دو این الگوریتم‌ها مربوط به دستهٔ meta"" می باشند و الگوریتم‌های IBK، kstar، و LWL بهترین نتایج را به دست آورده‌اند که هر سه این الگوریتم‌ها مربوط به دستهٔ مبتننی بر نمونه یا LAZY می باشند. و در میان سه الگوریتم، دقت طبقه‌بندی kstar به‌طور اندکی بهتر از دو الگوریتم دیگر در این دسته می‌باشد.

6. نتیجه‌گیری

در این مقاله روش نظارتی را با استفاده از یادگیری مبتنی بر نمونه برای رفع ابهام معنایی کلمات فارسی ارائه دادیم و برای آموزش از بانک منحصر به فردی که جملات تقطیع شده در زمینه موضوعات مختلف دارد به عنوان دادهٔ آموزشی استفاده کردیم. تعداد کلمات موضوعی انتخابی ما 5 درصد از تعداد کل کلمات مبهم در آن حوزهٔ معنایی است. ما اگر مقدارکمتر از 5 درصد را انتخاب می‌کردیم، سبب می‌شد کلمات موضوعی کمتری انتخاب شوند و چون مقادیر ویژگی‌های ما بر اساس وجود کلمات موضوعی در جمله است این امر سبب می‌شد که مقدار ویژگی‌ها اغلب صفر باشد، به دلیل این‌که کلمات موضوعی را محدود کرده بودیم و این مسئله در قدرت تمایز نمونه‌های رده و در نتیجه در دقت الگوریتم تاثیر منفی می‌گذارد. برعکس اگر بیشتر از 5 درصد برای کلمات موضوعی انتخاب کنیم، تعداد ویژگی‌ها افزایش می‌یابد و برعکس مورد قبل، این بار افزایش ویژگی است که تاثیر منفی در الگوریتم می‌گذارد. به این صورت که با افزایش کلمات موضوعی، کلماتی که تعداد تکرار کمتری دارند و یا کلماتی که بین معانی مختلف یک واژه مشترک هستند نیز انتخاب می شوند که این کلمات سبب می‌شود دقت الگوریتم پایین آید. از طرف دیگر، افزایش ویژگی سبب بزرگ‌تر شدن ماتریس ویژگی‌ها می‌شود و این امر منجر می‌شود که زمان مصرفی برای پردازش و طبقه‌بندی افزایش یابد. ما با استخراج دو دسته ویژگی از جملات حاوی کلمات مبهم، مدلی جهت رفع ابهام معنایی ارائه کردیم. دسته اول ماتریس ویژگی است که سطرها جملات و ستون‌ها کلمات موضوعی را نشان می‌دهند. مقادیر هر درایه وجود یا عدم وجود کلمات موضوعی وابسته به وزن را در جمله با صفر و وزن آن کلمهٔ موضوعی نشان می‌دهد. دسته دوم ویژگی‌ها، مربوط به مجموع وزن کلمات موضوعی موجود در آن جمله است. سطرهای این ماتریس ویژگی جملات هستند و دارای یک ستون است که مقدار آن ستون برای هر جمله برابر با مجموع وزن کلمات موضوعی است که در آن جمله موجود است. نتایج آزمایش‌ها نشان می‌دهد در روش پیشنهادی ما الگوریتم‌های مبتنی بر نمونه نسبت به سایر الگوریتم‌ها بهتر عمل کرده است. پیشنهاد می‌شود که در آینده ویژگی POS کلمات به همراه ویژگی کلمات موضوعی و همچنین بهبود مراحل پیش‌پردازش و الگوریتم یافتن کلمات مرکب، با استفاده از یک شبکه واژگان فارسی مورد بررسی قرار گیرد.

مراجع

[1] Abhishek, Fulmari, وManoj B Chandak ,2013,. "A Survey on Supervised Learning for Word Sense Disambiguation." International Journal of Advanced Research in Computer and Communication Engineering,vol.2.

[2] M.Lesk , 1987, "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone." Proc. 5th Annu. Int. Conf. Syst. Doc.24-26

[3] yarowsky, D,2000,. "Hierarchical Decision Lists for Word Sense Disambiguation." Computers and the Humanities.pp.179-186.

[4] You-Jin Chung, et al ,2001, . "Word Sense Disambiguation Using Neural Networks with Concept Co-occurrence Information." Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium,. Tokyo, Japan : National Center of Sciences.

[5] J. Veenstra, et al, "Memory-Based Word Sense Disambiguation," Computers and the Humanities, vol.34 ,ppt.171-177, 2000.

[6] R. Makki and M. M. Homayounpour, “Word Sense Disambiguation of Farsi Homographs using Thesaurus and Corpus,” Presented at the proceeding of the 6 th international conference on Advances in Natural Language Processing, Gothenburg , Sweden, 2008.

[7] مسعودی, بابک, سعید راحتی قوچانی, و اعظم استاجی. 1389، یک روش بیزی برای رفع ابهام معنایی کلمات در زبان فارسی با تاکید بر ویژگی‌های محلی، اولین کنفرانس نرم محاسبات نرم و فناوری اطلاعات. ایران: دانشگاه آزاد اسلامی واحد ماهشهر, اسفند، شماره صفحه.1-5

[8] تدین, محمد علی, منصور جهرمی, و مصطفی فخراحمد. 1390،روش جدید رفع ابهام معنایی در ترجمه ماشین مبتنی بر مجموعه متون، همایش ملی علوم و مهندسی کامپیوتر. ایران: دانشگاه آزاد اسلامی واحد نجف آباد, اسفند. 1-8.

[9] مسعودی، بابک، سعید راحتی, و اعظم استاجی. 1389، یک مدل بیشینه بی‌نظمی جهت رفع ابهام معنایی کلمات فارسی به کمک ویژگی‌های مدل‌سازی موضوع، شانزدهمین کنفرانس ملی سالانه انجمن کامپیوتر. تهران: دانشگاه صنعتی شریف , شماره صفحه 1-5.

[10] LiuisMµarquez , 2006, Machine Learning Techniques for Word Sense Disambiguation /Villodre, German RigauClaramu. Thesis For the obtention of the PhD Degree, Barcelona: UniversitatPolitµecnica de Catalunya.

[11] David W.Aha, Dennis Kibler, Marc K.Albert,1991, . “Instance-Based Learning Algorithms.” Machine Learning,vol.6,pp. 37-66

[12] Kambe, M, و J Han,2001, Data Mining: Concepts and Techniques. San Diego Academic Press.

[13] Seishi Okamoto, Nobuhiro Yugami,2003, . “Effects of domain characteristics on IBL algorithms.” Theoretical Computer Science,pp.207-233.

[14] Ping Chen, Wei Ding, Max Choly, Chris Brows, “Word Sense Disambigation with Automatically Acquired Knowledge”, Volume 24. Intelligent System, IEEE, 2012