اولین کنفرانس ملی محاسبات نرم و فناوری اطلاعات
1st National Conference on Soft Computing and Information Technology
دانشگاه آزاد اسلامی واحد ماهشهر
۱۲و۱۳ اسفند ماه ۸۹

بازشناسی رابطه تقابل در زبان فارسی در حالت حذف نقش‌نمای گفتمان به کمک ماشین بردار پشتیبان

حبیب خدادادی [۱] – سعید راحتی قوچانی [۲] – اعظم استاجی [۳]
(۱) گروه کامپیوتر - دانشگاه آزاد اسلامی واحد مشهد
habibekhodadadi@gmail.com
(۲) گروه برق – دانشگاه آزاد اسلامی واحد مشهد
Rahati@Mshdiau.ac.ir
(۳) گروه زبانشناسی – دانشگاه فردوسی مشهد
estagi@ferdowsi.um.ac.ir

خلاصه: شناسائی رابطه تقابل در یک گفتمان می‌تواند به کمک نقش‌نماهای خاص رابطه تقابل مانند "اما" و "ولی" انجام شود؛ اما این نقش‌نماها در مواردی حذف می‌شوند و شناسائی رابطه را با مشکل مواجه می‌کنند. به همین علت از ویژگی‌هائی شامل زمان فعل‌ها، استفاده از جفت کلمات، وجود اعداد در دو قسمت متنی اطراف نقش‌نما و وجود فعل ها و کلمات منفی در یک طرف نقش‌نما و عدم وجود آن در طرف دیگر به منظور شناسائی این رابطه استفاده شد. در این مقاله از مجموعه داده پژوهشکده هوشمند علائم استفاده شده و ۵۰۰۰ نمونه رابطه تقابل و ۵۰۰۰ نمونه سایر روابط گردآوری شده، بردار ویژگی را به ازای هر نمونه شکل داده و دسته‌بند *SVM* با این ویژگی‌ها آموزش داده شد. میزان صحت دسته‌بند ۶۱.۰۲ در بهترین حالت می‌باشد.

کلمات کلیدی: گفتمان ، نقش‌نمای گفتمان ، شناسائی رابطه تقابل، ماشین بردار پشتیبان.

۱ – مقدمه

گفتمان *(Discourse)* به هر قطعه یا پاره‌ای از زبان گفته می شود که به قصد برقراری ارتباط به کار برده شده باشد [۱] و به عنوان واحد بالاتر از جمله در زبان‌شناسی شناخته می شود . پردازش کامپیوتری گفتمان یکی از شاخه های نسبتا جدید حوزه پردازش متن می باشد که در زبان فارسی به آن چندان پرداخته نشده است. یکی از زمینه های مطالعه در پردازش گفتمان، شناسائی و برچسب گذاری روابط بین واحدهای متنی (بند، جمله یا پاراگراف) در سطح یک گفتمان است . بعضی از این رابطه ها شامل علّی ، افزایشی ، استمراری ، تقابل ، زمانی و ... می باشد .

نقش نماهای گفتمان (*Discourse connectives*) عناصری در سطح گفتمان هستند که کارکرد آنها نشان دادن رابطه ای است که بین یک پاره گفتار و پاره گفتار قبلی وجود دارد [۱] . از جمله این عناصر در زبان انگلیسی می توان *"And"* ، *"Or"* ، *"But"* ، *"Well"* ، *"Because"* و ... و در زبان فارسی "اما" ، "ولی" ، "خب" ، "حالا" ، "پس" ، "بنابراین"، و ... را نام برد . در صورت وجود داشتن نقش نماها ، از آنها می توان به عنوان عامل شناسائی نوع رابطه استفاده کرد ولی در مواردی این عناصر وجود ندارند .

در تحقیقات گسترده ای که به خصوص در زبان انگلیسی صورت گرفته است ، با استخراج ویژگیهائی سعی در شناسائی این روابط در سطح گفتمان شده است ؛ اما در زبان فارسی در این مورد کاری انجام نشده است .

در [۳] با جمع آوری نمونه های زیادی از ۴ رابطه گفتمانی و سپس حذف نقش نماهای گفتمان و با این فرض که خود کلمات موجود در رابطه می توانند نوع رابطه را تعیین کنند، یک مدل احتمالاتی از جفت کلمات دو طرف رابطه را بدست آورده (هر جفت کلمه شامل یک کلمه قبل و یک کلمه بعد از نقش نما می باشد) و سپس از این مدل و با داشتن کلمات دو طرف هر رابطه گفتمانی و بدون استفاده از نقش نما ، احتمال تعلق به هر کدام از رابطه ها را بدست آورده و بیشترین احتمال نوع رابطه را تعیین می کند .

در [۴] که بر روی زبان ژاپنی کار شده است ، علاوه بر استفاده از جفت

کلمات از اطلاعات الگوهای گروهی *(Phrasal Pattern)* نیز استفاده شده است . این الگوهای عبارتی قسمت هائی از جمله هستند که در بعضی از روابط ظاهر می شوند . مثلا رابطه تضاد در زبان انگلیسی در مواردی به صورت *"... did ... should have done ... ." And "...* ظاهر می شود . با استفاده از این اطلاعات اقدام به شناسائی روابط شده است . در [۵] فقط به بررسی حالاتی که نقش نما وجود ندارد پرداخته شده است و بر روی *PDTB* ، که مجموعه برچسب خورده گفتمانی در زبان انگلیسی است ، نمونه های حذف نقش نما پیدا شده و ویژگی هائی مانند وجود قطب های مختلف کلمات ، کلاس های فعلی ، جفت کلمات و ... استخراج شده و با کمک این ویژگی ها دسته بندی صورت گرفته است .

در [۶] نیز برای حالتی که نقش نما وجود ندارد و بر روی *PDTB* ویژگی هائی استخراج شده است و با کمک این ویژگی ها نوع رابطه تشخیص داده شده است . که از ویژگی هائی مانند روابط قبل و بعد رابطه فعلی ، وجود وابستگی های رابطه ای ، ویژگی های استخراج شده از درخت نحو و ... در این کار استفاده شده است . در [۷] که از مجموعه داده *PDTB*۲ استفاده شده است ، ضمن استفاده از ویژگی های مقالات قبلی ویژگی جدیدی معرفی شده است ، که با استفاده از یک مدل زبانی ابتدا نقش‌نما پیش گوئی می شود (در حالت نبود نقش نما) و سپس از این نقش نمای پیش گوئی شده به عنوان یک ویژگی اضافی به منظور شناسائی نوع رابطه استفاده می شود .

۲ – رابطه‌های گفتمانی و رابطه تقابل

در سطح هر گفتمان روابطی بین واحدهای متنی وجود دارد که در حالت عادی نوع این روابط به کمک عناصری به نام نقش‌نماهای گفتمان تشخیص داده می‌شود . شناسائی رابطه های موجود در یک گفتمان کاربردهای گسترده ای دارد ؛ از جمله این که شناسائی این روابط به توانائی تولید و درک گفتمان کمک می کند و در سیستم های پرسش و پاسخ و خلاصه ساز و ... کاربرد دارند [۳]. مثال ۱ نمونه ای از یک رابطه گفتمانی در زبان انگلیسی می باشد که با وجود نقش نمای *But* نوع رابطه *Contrast* تشخیص داده می شود [۶].

مثال ۱ :

In any case, the brokerage firms are clearly moving faster to create new ads than they did in the fall of ۱۹۸۷. ***But*** *it remains to be seen whether their ads will be any more effective.*

تشخیص نوع رابطه به کمک نقش نما با سه مشکل مواجه است :

۱- بعضی از نقش نما ها دارای کارکرد غیر نقش نمائی هم می باشند؛ مانند مثال ۲ قسمت ۲*b* که کلمه *and* نقش نمای گفتمان نیست ولی در ۲*a* نقش نمای گفتمان می باشد [۹].

۲- بعضی از نقش نماها بین دو یا چند رابطه مشترک هستند و تشخیص رابطه با ابهام مواجه می شود ؛ مانند مثال ۳ که نقش نمای *since* هم برای رابطه زمامی (*temporal*) و هم برای رابطه علّی *(causal)* به کار می رود[۹].

۳- در خیلی از موارد نقش نماها حذف می شوند (مثال ۴)[۶] .

مثال ۲ :

(۲a) Selling picked up as previous buyers bailed out of their positions and aggressive short sellers– anticipating further declines–moved in.
(۲b) my favorite colors are blue and green.

مثال ۳ :

(۳a) there have been more than ۱۰۰ mergers and acquisitions Within the European paper industry since the most recent wave of friendly takeovers was completed in the U.S. in ۱۹۸۶.
(۳b) It was a far safer deal for lenders since NWA had a healthier cash flow and more collateral on hand.

مثال ۴ :

"A lot of investor confidence comes from the fact that they can speak to us,"he says.
"To maintain that dialogue is absolutely crucial."

بنابراین در حالتی که هیچ یک از ۳ مشکل بالا برای یک نقش نما خاص اتفاق نیفتد ، صرفا با نقش نما می توان نوع رابطه را شناسائی کرد؛ اما در حالتی که یکی از مشکلات بالا برای یک نقش نمای خاص اتفاق می افتد باید از ابزارهای واژگانی، نحوی و... دیگری برای این منظور استفاده کرد.

یکی از روابط موجود در گفتمان در زبان فارسی رابطه تقابل است . از نظر مناسبات معنائی در رابطه تقابل در سطح گفتمان بین بخش اول و بخش دوم به نوعی نتیجه خلافی ، عکس یا مغایرت وجود دارد . این رابطه در سطح گفتمان به وسیله نقش نما های گفتمان مخصوص رابطه تقابلی شناخته می شوند از قبیل اما ، ولی ، لکن ، با این همه ، با این حال ، علی رغم و .... این نقش نماها می توانند در مواردی حذف شوند که خواننده (شنونده) از روی معنا به نوع رابطه پی می برد . در مثال ۵ و ۶ دو نمونه از رابطه تقابلی از پیکره متنی بی جن خان *(http://ece.ut.ac.ir/dbrg/bijankhan)* نشان داده شده است که اولی دارای نقش نمای تقابلی "ولی" می باشد ، اما دومی فاقد نقش نما می باشد.

مثال ۵ :

البته تمام اقلام غذائی ذکرشده برای همه مبتلایان به میگرن اثر نامطلوب ندارند ***ولی*** *در تعداد زیادی از بیماران موثر هستند .*

مثال ۶ :

این دوربین ابتدا برای تشخیص پرتوهای کیهانی طراحی شده بود ، در حال حاضر با تغییراتی که در آن ایجاد شده است ، مصارف پزشکی نیز پیدا کرده است .

بنابراین و با توجه به مشکل حذف نقش نمای تقابلی، از روش‌های

دیگری به جز استفاده از نقش نما برای شناسائی این رابطه استفاده می کنیم تا در موقع حذف نیز مشکلی برای شناسائی وجود نداشته باشد.

۳ – مشکلات بازشناسی روابط گفتمان در زبان فارسی

یکی از بزرگترین موانع پیش روی پردازش کامپیوتری گفتمان، نبود داده برچسب خورده خاص گفتمان می باشد . در زبان انگلیسی داده برچسب خورده وجود دارد . یکی از این داده ها ، داده *Penn Discourse Treebank (PDTB)* [۸] می باشد؛ که ضمن برچسب گذاری هر رابطه ، در حالت حذف نقش نما نیز ، بهترین نقش نمائی که می تواند در آنجا قرار گیرد معرفی شده است . و برای هر نقش نما هم دو بخش قبل و بعد آن با *Arg*۱ و *Arg*۲ نشان داده شده است . مثال ۷ و ۸ دو نمونه از این داده ها را نشان می دهد .

مثال ۷:

***Arg*۱:** *In any case, the brokerage firms are clearly moving faster to create new ads than they did in the fall of* ۱۹۸۷.
***Arg*۲: But** *it remains to be seen whether their ads will be any more effective.*

(Contrast)

مثال ۸:

***Arg*۱:** *"A lot of investor confidence comes from the fact that they can speak to us," he says.*
***Arg*۲: [so]** *"To maintain that dialogue is absolutely crucial."*

(Cause)

علاوه بر عدم وجود داده برچسب خورده ، فقدان برخی از ابزارهای پردازش زبان طبیعی مانند *wordnet* و... نیز از مشکلات مهم می باشد. مشکلات طبیعی رسم الخط زبان فارسی از قبیل مشکل کلمات مرکب و راست به چپ نویسی و ... نیز وجود دارد.

۴ – استخراج ویژگی برای سیستم شناسائی رابطه تقابل در گفتمان های فارسی

در این مقاله ما به منظور شناسائی رابطه تقابل در زبان فارسی از ۴ ویژگی استفاده کرده ایم . در همه جا منظور از *Arg*۲ عبارت و جملاتی است که نقش نما به طور ساختاری در آن طرف قرار دارد و طرف دیگر به عنوان *Arg*۱ شناخته می شود . این ویژگی ها شامل :

۱- از آنجا که در عبارات تقابلی ممکن است کلمات و فعل های مثبت و منفی در دو طرف نقش نما ظاهر شوند ، اولین ویژگی عبارت است از وجود فعل ها و کلمه های منفی در یک طرف نقش نما و عدم وجود آن در طرف دیگر .

۲- اگر هر دو *Arg*۱ و *Arg*۲ شامل عدد ، درصد و یا مقدار عددی باشد، احتمالا یک رابطه تقابل وجود دارد[۵]. بنابراین دومین ویژگی شامل وجود عدد، درصد و یا مقدار عددی در هر دوی *Arg*۱ و *Arg*۲ می باشد.

۳- هم زمان بودن فعل های اصلی *Arg*۱ و *Arg*۲ . از آنجا که در بعضی از رابطه ها مانند رابطه زمانی *(Temporal)* ، زمان فعل‌های *Arg*۱ و *Arg*۲ ممکن است با هم تفاوت داشته باشد؛ هم زمان بودن فعل قبل از نقش نما و آخرین فعل بعد از نقش نما به عنوان ویژگی در نظر گرفته شده است.

۴- وجود جفت کلمات پرکاربرد رابطه تقابلی.

اگر رابطه تقابل بین دو قسمت متنی برقرار باشد و *W*۱ و *W*۲ به تمام کلمات این دو قسمت اشاره کند، فرض ما این است که جفت کلمات $(w_i,w_j) \in W1 \times W2$ (حاصل ضرب برداری دو طرف رابطه) می تواند به ما وجود رابطه تقابل را نشان دهد . در این راستا و به منظور شناسائی این نوع جفت کلمات کارهای زیر انجام شد :

۱- از دو پیکره بی جن خان و پیکره‌ی متنی زبان فارسی که توسط پژوهشکده پردازش هوشمند علائم (*http:// www.rcisp.com*) تهیه شده است ، تمام نمونه های رابطه تقابل که دارای یکی از نقش نماهای "ولی"، "اما "، "لکن "، "بااین‌وجود"، "بااین‌حال"، "باوجوداین" و "لیکن" هستند ، استخراج شدند که کلمات قبل از آن به عنوان *Arg*۱ و کلمات بعد از آن به عنوان *Arg*۲ برچسب زده شد . مثال‌هائی از هر کدام از این نقش‌نماها را در جدول ۱ مشاهده می کنید .

۲- این نقش نماها و هر کلمه ای که نقشی بجز فعل، اسم،صفت یا قید دارد حذف شدند . همچنین فعل است هم که معنی خاصی را به همراه ندارد نیز حذف شد.

۳- هر کلمه از *pos* آن جدا شد .

۴- ضرب برداری کلمات باقی مانده در دو طرف رابطه انجام شد بدین معنی که هر کلمه در *Arg*۱ را با تک تک کلمات *Arg*۲ به صورت یک جفت مجزا در نظر گرفته و در تمام نمونه ها تعداد این جفت ها شمرده می شود.

۵- تمامی جفت کلماتی که زیر ۵۰ بار تکرار شده است را حذف و باقی جفت کلمات را ذخیره می کنیم. بنابراین وجود داشتن این جفت کلمات در هر رابطه را به عنوان یک ویژگی در نظر گرفته ایم .

جدول (۱): مثال‌هائی برای نقش‌نماهای تقابل استفاده شده .

نقش‌نمای تقابل	مثال
اما	روایت داستان آنقدر کشش داشت که مخاطب را به سمت خود بکشد و در هر قسمت او را وادار کند پای تلویزیون بنشیند. اما ضعف آن مانند اکثر

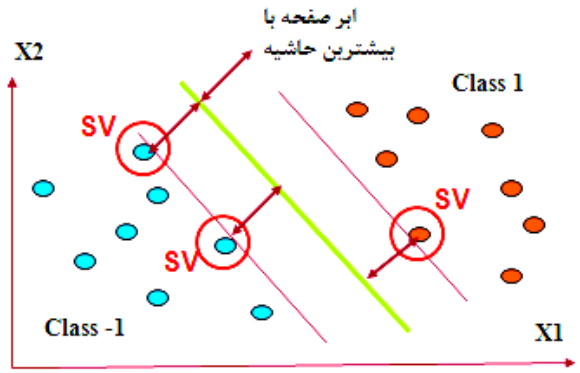
	سریالهای ایرانی پایان آن است که به راحتی در چند دیالوگ خلاصه می‌شود.
ولی	برآورد مبالغ غیر قابل وصول و اندازه گیری درآمد احتمال برابری میان مبالغ مطالباتی که هرساله غیر قابل وصول می‌گردد و مبلغی که به عنوان ذخیره برای همین رویداد منظور می‌شود فوق العاده ضعیف است. ولی تفاوتهای میان این دو رقم اگر به نسبت اندک باشد، قابل چشم‌پوشی است.
لکن	خانواده مادری یزید، از اعراب مسیحی بودند که بعدا اسلام آوردند. لکن اسلام آنان هیچ گاه آداب و ارزشها و فرهنگ مسیحی آنان را دگرگون نکرده بود.
لیکن	یک بار دیگر به بهانه مذاکره با کاسترو درباره خلیج خوکها لباس آغشته به سم را به او هدیه کردند، لیکن کاسترو موضوع را با دیده شک نگریست.
بااین‌حال	در کانادا و آمریکا تنبیه بدنی کودکان توسط والدین آنها تنها در شرایط خاصی جرم محسوب نمی‌شود. بااین حال چنین استثنائی نیز به تدریج در حال رنگ باختن است و گروههای حمایت از کودکان در صدد هستند این موارد را نیز غیرقانونی اعلام کنند.
باوجوداین	بر روی هر سکه یک مهر خاص ضرب شده بود که نشان ضمانت بهای سکه بود. با وجود این، به نوشته روزنامه ایندیپندنت، پس از آن که بر اداره کنندگان کشور روشن شد که از این نوع سکه ها در نقاط دوردست تر پادشاهی استقبال نمی‌شود، صنعت گران لیدیه به ساخت نخستین کوره ذوب و استخلاص طلا همت گماشتند.
بااین وجود	سم علف کش اترازین به علت قدرت بسیار زیاد در نابودی علف های هرز و همچنین قیمت پائین آن بسیار پرمصرف است. با این وجود پژوهشگران می گویند باید قوانینی برای محدود ساختن مصرف آن ایجاد کنند.

۵ – ماشین بردار پشتیبان

ماشین بردار پشتیبان *(Support Vector Machine (SVM))* یکی از انواع روش‌های دسته‌بندی می‌باشد که در هر لحظه توانائی دسته‌بندی دو کلاس را دارد ، یا به عبارت دیگر یک نوع دسته‌بند دودوئی می‌باشد. در *SVM* هدف پیدا کردن ابرصفحه‌ای برای جداکردن داده‌ها در دو کلاس جداگانه است. رویکرد *SVM* به این صورت است که در فاز آموزش، سعی می شود که مرز تصمیم گیری *(Decision Boundry)* به‌گونه‌ای انتخاب گردد که حداقل فاصله آن با هر یک از دسته‌های مورد نظر ماکزیمم گردد. این نوع انتخاب باعث می شود که تصمیم گیری ما در عمل، شرایط نویزی را به خوبی تحمل کند و پاسخ دهی خوبی داشته باشد. این نحوه انتخاب مرز بر اساس نقاطی به نام بردارهای پشتیبان انجام می شود. در *SVM* با فرض اینکه دسته‌ها به‌صورت خطی جداپذیر باشند، ابرصفحه‌هائی با حداکثر حاشیه *(Maximum Margin)* را بدست می‌آورد که دسته‌ها را جدا کنند . برای داده‌های جداپذیر به صورت خطی تابع تفکیک کنندگی به شکل زیر نمایش داده می‌شود:

$$g(x) = sign(w.\theta(x) + b) \qquad (۱)$$

در (۱) علامت مثبت تعلق به دسته اول و علامت منفی تعلق به دسته دوم را نمایش می‌دهد، x بردار نمونه‌های آموزشی (ویژگیها)، w بردار ابرصفحه و b بایاس آن می‌باشد که فاصله بین ابرصفحه تصمیم گیری و مبدا فضای ویژگی را نمایش می‌دهد. به طور کلی ابرصفحه‌های زیادی برای دسته‌بندی وجود دارد که در بین آنها *SVM* ابرصفحه‌ای را پیدا می‌کند که فاصله بیشتری را بین دو کلاس ایجاد کند؛ نزدیکترین داده‌های آموزشی به ابرصفحه‌های جداکننده بردار پشتیبان *(Support Vector(SV))* نامیده می‌شوند. این موضوع در شکل ۱ نشان داده شده است .



شکل ۱: دسته‌بند *SVM* برای جدا کردن یک نوع داده خاص . در این شکل بردارهای پشتیبان و ابرصفحه جداکننده را مشاهده می‌کنید.

در بسیاری از کاربردهای واقعی ابرصفحه‌ای که بتواند به صورت خطی جداسازی را در دو کلاس انجام دهد وجود ندارد ؛ برای حل این مشکل یک راه حل این است که داده‌ها از فضای اصلی به فضای دیگری (فضای با بعد بالاتر) انتقال دهیم و بعد ابرصفحه‌ای را تعیین نمائیم که بتواند به صورت خطی این جداسازی را انجام دهد که در این حالت بار محاسباتی زیاد می‌شود و هزینه محاسبات بالا می‌رود چرا که به طور کلی ابعاد فضائی که در آن فضا داده‌ها به طور خطی جداپذیر خواهند بود بی‌نهایت است. از این رو فرآیند نگاشت را توسط توابع هسته

(Kernel Function) به طور ضمنی انجام خواهیم داد. *SVM* در داده‌هائی که بصورت خطی جداپذیر نمی‌باشند وابسته به هسته *(Kernel)* می‌باشد. هسته تابعی است که الگوها را از یک فضای غیرخطی به یک فضای خطی نگاشت می‌کند تا *SVM* بتواند با استفاده از یک ابرصفحه الگوها را دسته‌بندی کند. در صورت استفاده از توابع هسته ، تابع تصمیم‌گیری یک *SVM* بر پایه هسته به شکل زیر تبدیل می‌گردد:

$$g(x) = sign(\sum_{i=1}^{N_s} \alpha_i y_i k(x_i, x) + b) \qquad (۲)$$

در (۲) $x_i \in R^d$ ،$i=1,...,N$ ، d ابعاد ویژگیها ، N تعداد نمونه‌های آموزشی مربوط به دو دسته، $\alpha i, yi \in \{1,-1\}$ ضرایبی که از حل مساله بهینه‌سازی به دست می‌آیند و N_S تعداد بردارهای پشتیبان می‌باشد که مهمترین داده‌های آموزشی هستند چرا که آنها به تنهائی برای تعریف حاشیه تفکیک پذیری کفایت می‌کنند. از معروفترین و متداولترین هسته‌هایی که در *SVM* به کار برده می‌شوند می‌توان به ۳و ۴ اشاره کرد:

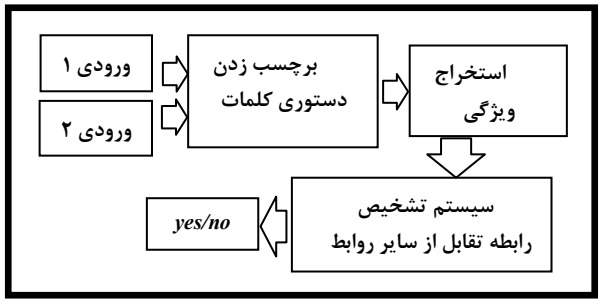
$$k(X, X') = exp(-\frac{1}{2\sigma^2} || X - X' ||^2) \qquad (۳)$$

$$k(X, X') = (X^T X' + 1)^p \qquad (۴)$$

که (۳) هسته *RBF* و (۴) هسته چند جمله‌ای می‌باشد.

۶ – سیستم شناسائی رابطه تقابل در گفتمان‌های زبان فارسی

سیستمی که ما قصد طراحی آن را داریم در شکل ۲ نشان داده شده است. در شکل۲ منظور ما از ورودی ۱ و ورودی ۲ جمله یا بند با پاراگراف قبل و بعد نقش نما با دو عبارت متنی می باشد که پس از برچسب زدن دستوری کلمات که در خود پیکره انجام شده است، به استخراج ویژگی اقدام می‌کنیم و اگر دو ورودی نسبت به هم رابطه تقابل داشته باشند، خروجی *yes* و اگرنه خروجی *no* تولید می شود.



شکل۲- شکل کلی سیستم شناسائی رابطه تقابل

به منظور نشان دادن کارائی ویژگی های گفته شده در شناسائی رابطه تقابل به کمک دسته‌بند *svm* آزمایشاتی صورت گرفت . در ابتدا ۵۰۰۰ نمونه رابطه تقابل که صراحتا با نقش نماهای خاص رابطه تقابل مانند اما، ولی، لکن، لیکن و ... وجود داشتند، استخراج شد و ۴۰۰۰ نمونه از سایر روابط به تعدادی که در جدول ۲ مشاهده می شود، استخراج گردید؛ همچنین تعداد ۱۰۰۰ نمونه جفت جملاتی را که ما فرض می کنیم رابطه ای بین آنها نیست را استخراج می کنیم . این جفت جملات همجوار نیستند و هر کدام را به طور تصادفی از متون پیکره استخراج می کنیم و سپس جمله اول را به عنوان *Arg*۱ و جمله دوم را به عنوان *Arg*۲ در نظر می‌گیریم و مانند یک رابطه به استخراج ویژگی می پردازیم .

جدول (۲): تعداد و نوع نقش نماهای استفاده شده.

نقش‌نما	نوع رابطه	تعداد
اما	تقابل	۲۹۰۰
ولی	تقابل	۱۵۰۰
لکن	تقابل	۵۰
لیکن	تقابل	۲۰۰
بااین‌حال	تقابل	۲۵۰
باوجوداین	تقابل	۵۰
بااین وجود	تقابل	۵۰
زیرا	علّی	۶۰۰
یعنی	افزایشی	۱۰۰۰
پس	زمانی	۴۰۰
و	افزایشی	۲۰۰
همچنین	افزایشی	۵۰۰
دراین میان	زمانی	۵۰
دراین حال	زمانی	۱۰۰
چون	علّی	۱۵۰
بنابراین	علّی	۲۰۰
به علاوه	افزایشی	۵۰
برای این که	علّی	۵۰
بعد	افزایشی [۲]	۴۰۰
آنگاه	زمانی	۱۵۰
علاوه بر این	افزایشی	۱۵۰

تمامی این نمونه ها از پیکره پردازش هوشمند علائم استخراج شده اند . در این پیکره کلمات یک جمله با توجه به جایگاه نحوی آن کلمه به چند گروه تقسیم شده و برای هر گروه یکسری ویژگی هایی مختص آن گروه بیان شده است برای مثال در این متون برای افعال ویژگی هایی مانند زمان فعل ،مرکب یا ساده بودن فعل ،شخص فعل، مثبت یا منفی بودن فعل و ... ذکر شده است و یا برای اسامی ویژگی

هایی مانند خاص یا عام بودن اسم ،جمع یا مفرد بودن ، اسم مکان ذکر گردیده است.

نمونه های استخراج شده را به صورت { [ARG۱] نقش‌نمای گفتمان [ARG۲] } مرتب می کنیم . ([ARG۱] و [ARG۲] دو قسمت گفتمانی در دوطرف نقش‌نما هستند که می‌تواند بند ، جمله یا پاراگراف باشد). در مورد ۱۰۰۰ نمونه بدون رابطه [ARG۱] را جمله اول و [ARG۲] را جمله دوم در نظر می‌گیریم. در ادامه نقش‌نماها را حذف کرده و صرفا از دو قسمت [ARG۱] و [ARG۲] به عنوان منابعی برای تصمیم گیری استفاده می کنیم . سپس بردار ویژگی را برای [ARG۱] و [ARG۲] های ایجاد شده برای هر نمونه تشکیل می دهیم و نمونه ها را به دو کلاس تقسیم می کنیم که تمام نمونه های رابطه تقابل را یک کلاس در نظر گرفته و بقیه را کلاس دیگری در نظر می گیریم.

در ادامه تمام بردارهای ویژگی که برچسب کلاس مربوطه را دارند را به نرم افزار MATLAB ۷.۸ منتقل کرده داده ها را به داده های آموزش و آزمایش تقسیم کرده و به منظور آموزش دسته بند svm استفاده می کنیم که مراحل آموزش و آزمایش را با توابع هسته مختلف امتحان می کنیم که نتایج و میزان صحت را در جدول ۳ مشاهده می کنید.

جدول ۳: درصد صحت روش پیشنهادی با دسته‌بند svm و توابع هسته مختلف

نوع هسته	درصد صحت
خطی	۶۱.۰۲
درجه ۲	۶۰.۵۹
درجه ۳	۵۹.۷۴
RBF	۶۰.۹۶

۷- نتیجه گیری و کارهای آینده

دراین مقاله روشی برای شناسائی رابطه تقابل در گفتمان های زبان فارسی ارائه شد که اولین کار در حوزه پردازش کامپیوتری گفتمان می باشد. از آنجا که در رابطه تقابل نیز ممکن است در مواردی نقش‌نما حذف شود ، بنابراین سیستمی مستقل از نقش‌نما ارائه شد که با استخراج ویژگی هائی سعی در شناسائی رابطه تقابل دارد. نتایج دسته-بند svm بر روی ویژگی های استخراج شده ، امیدوارکننده بود و در بهترین حالت بیشتر از ۶۰ درصد صحت ایجاد شد. به عنوان پیشنهاد کار برای آینده می‌توان ویژگی های بهتری را استخراج کرد مانند پیدا کردن کلمات متضاد در دو طرف رابطه و یا پیدا کردن الگوها و عباراتی که نقش نما نیستند اما در رابطه تقابل ممکن است به کار گرفته شوند. به علت این که تحقیقات زیادی در پردازش‌های کامپیوتری گفتمان در زبان فارسی انجام نشده است، کارهای زیادی را می توان در این حوزه انجام داد. به عنوان مثال شناسائی سایر روابط موجود در یک گفتمان می تواند انجام شود و ساخت یک پارسر برای روابط موجود در یک گفتمان و ....

۸- مراجع

[۱] ذوقدارمقدم، رضا و دبیرمقدم، محمد، **"نقش‌نماهای گفتمان مقایسه نقش‌نمای *but* در زبان انگلیسی با نقش‌نمای امّا در زبان فارسی"** ، پژوهش زبان‌های خارجی، ۱۲ ، ۵۵-۷۶ ، ۱۳۸۱.

[۲] مقدم‌کیا، رضا، **"بعد، نقش‌نمای گفتمان در زبان فارسی"**،نامه فرهنگستان ، ۲۳، ۸۱-۹۸، ۱۳۸۴.

[۳] D. Marcu and A. Echihabi. ۲۰۰۱. An unsupervised approach to recognizing discourse relations. In Proceedings of the ۴۰th Annual Meeting on Association for Computational Linguistics, pages ۳۶۸–۳۷۵.

[۴] Manami Saito, Kazuhide Yamamoto, and Satoshi Sekine. ۲۰۰۶. Using phrasal patterns to identify discourse relations. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL ۲۰۰۶), pages ۱۳۳– ۱۳۶, New York, USA, June.

[۵] Emily Pitler, Annie Louis, and Ani Nenkova. ۲۰۰۹. Automatic sense prediction for implicit discourse relations in text. To appear in Proceedings of the Joint Conference of the ۴۷th Annual Meeting of the Association for Computational Linguistics and the ۴th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP ۲۰۰۹) , pages ۶۸۳– ۶۹۱.

[۶] Z.H. Lin, M.Y. Kan and H.T. Ng. ۲۰۰۹. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. Proceedings of the ۲۰۰۹ Conference on Empirical Methods in Natural Language Processing (ACL and AFNLP ۲۰۰۹), pages ۳۴۳– ۳۵۱.

[۷] Zhou Z, M Lan , X Yu, Z Niu, J Su and C L Tan, Predicting discourse connectives for implicit discourse relation recognition, ۲۳rd International Conference on Computational Linguistics (COLING ۲۰۱۰), August ۲۳-۲۷, ۲۰۱۰, Bejing, China.

[۸] Rashmi Prasad, Nikhil Dinesh, Alan Lee, EleniMiltsakaki, Livio Robaldo, Aravind Joshi, and BonnieWebber. ۲۰۰۸. The Penn Discourse Treebank ۲.۰. In Proceedings of the ۶th International Conferenceon Language Resources and Evaluation (LREC۲۰۰۸).

[۹] E. Pitler and A. Nenkova. ۲۰۰۹. Using Syntax to Disambiguate Explicit Discourse Connectives in Text.Proceedings of the ACL-IJCNLP ۲۰۰۹ Conference Short Papers.