

نگرشی جدید به تحلیل عبارت های اسمی هم مرجع

مریم نظری دوست^۱، بهروز مینایی بیدگلی^۲

^۱ کارشناسی ارشد، مهندسی فناوری اطلاعات، دانشگاه قم nazaridousm@gmail.com

^۲ عضو هیئت علمی و استادیار دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران bminaei@gmail.com

چکیده

پردازش زبان طبیعی شامل وظایفی همچون استخراج اطلاعات، خلاصه سازی متن، پرسش و پاسخ می باشد که همگی نیاز دارند تا تمام اطلاعاتی که در مورد یک موجودیت در متن وجود دارد را شناسایی نمایند. بنابراین وجود سیستمی که بتواند موضوع تحلیل عبارت های اسمی هم مرجع را بررسی نماید، کمک شایانی به انجام موفقیت آمیز این وظایف خواهد نمود. ما در این مقاله، سعی داریم تا به طور دقیق، فرآیند تحلیل مرجع مشترک را بررسی نمائیم. در همین راستا نیز فرآیند مشابه دیگری تحت عنوان تحلیل پیشایند را نیز مطالعه خواهیم نمود. امید داریم تا با بررسی این فرآیند و مقایسه ی شباهت ها و تفاوت های آنها، به نگرشی جدید در مورد تحلیل عبارت های اسمی هم مرجع برسیم.

کلمات کلیدی

پردازش زبان طبیعی، استخراج اطلاعات، تحلیل مرجع مشترک، تحلیل پیشایند

موجودیت های نامدار^{۱۵}، کشف اشاره^{۱۶} و تحلیل مرجع مشترک^{۱۷} تشکیل شود. برای شروع، مثال (۱) را در نظر بگیرید^۱:

۱- مقدمه

یکی از ویژگی های خاص گفتمان این است که می توان در یک متن آزادانه در مورد یک یا چند موجودیت صحبت کرد و برای اشاره به هر موجودیت^۱ از انواع مختلف عبارت ها مانند ضمیر (او)، اسم عام^۲ (دانشمند)، اسم خاص^۳ (لطفعلی عسگر زاده) و یا یک عبارت اسمی (بنیان گذار منطق فازی) بهره برد تا به این ترتیب از تکرار عبارت ها کاسته و شیوایی مطلب نیز افزایش یابد. همین ویژگی موجب می شود که زنجیره های بالقوه ای از تمام عبارت های اسمی که به یک موجودیت واحد در متن اشاره دارند، ایجاد گردد. (مانند: او، دانشمند، لطفعلی عسگر زاده، بنیان گذار منطق فازی که به شخص پرفسور زاده اشاره دارند). یکی از اهداف مهم استخراج اطلاعات، شناسایی این زنجیره ها در متن است که تحت فرآیندی با عنوان تحلیل مرجع مشترک^۴ انجام می پذیرد.

استخراج اطلاعات^۵، به عنوان یکی از مهم ترین کاربردهای پردازش زبان طبیعی^۶ محسوب می شود که در دو دهه اخیر بسیار مورد توجه پژوهشگران قرار گرفته است. روال کار در استخراج اطلاعات به این ترتیب است که ابتدا متن های خام به عنوان ورودی دریافت شده و نهایتاً داده های پردازش شده به عنوان خروجی برگشت داده می شوند. قالب بدست آمده از استخراج اطلاعات می تواند برای کارهایی مانند داده کاوی^۷، پرسش و پاسخ^۸، درک زبان^۹، خلاصه سازی^{۱۰} و بازیابی اطلاعات^{۱۱} بسیار مفید واقع شود. هر سیستم استخراج اطلاعات می تواند به ترتیب از ماژول هایی مانند نشانه گذاری^{۱۲}، برچسب گذاری ادات سخن^{۱۳}، تجزیه گر^{۱۴}، تشخیص

مثال ۱: (سیستم آبیاری گلاب) Ant₁ در روز ۳ شنبه رونمایی شد. (این سیستم) Ana₁ محصول اندیشه ی (دکتر سارا شکری) Ant₂ است. (او) ۲، ۱ (محقق) Ana در (شرکت آبیاری لاله) ۳ است.

اگر فرض کنیم که ماژول های نشانه گذاری تا کشف اشاره به عنوان پیش پردازش هایی بر روی متن اجرا شوند، با اجرای این ماژول ها، انواع عبارت های اسمی موجود در متن (سیستم آبیاری، این سیستم، دکتر سارا شکری، او، یک محقق و شرکت آبیاری لاله) تعیین و برچسب گذاری می شوند. سپس با اجرای ماژول تشخیص مرجع مشترک، ارتباطات میان این عبارات و اطلاعات نهفته در مورد موجودیت های شرکت کننده در متن آشکار می شود. به عنوان نمونه، می دانیم «او» و «سارا شکری» (با اندیس ۲) به یک فرد مشخص و همچنین «سیستم آبیاری گلاب» و «این سیستم» (با اندیس ۱) نیز به یک سیستم مشخص اشاره می کنند.

استفاده از اصطلاح موجودیت در تحلیل مرجع مشترک، این سؤال را مطرح می کند که چه چیزهای موجودیت محسوب می شوند؟ تاکنون گروه بندی های متعددی برای انواع موجودیت ها ارائه شده است، به عنوان

^۱ در این مقاله دو نوع اندیس برای هر عبارت اسمی در نظر گرفته ایم، اندیس شماره برای تشخیص مرجع مشترک می باشد، به عنوان مثال، سیستم آبیاری گلاب و این سیستم هر دو به یک موجودیت اشاره دارند و اندیس این موجودیت ۱ می باشد. همچنین اندیس Ant نشان گر مقدم و اندیس Ana نشان گر تالی است که برای بررسی ارتباط پیشایندی در نظر گرفته شده است.

می‌شناسند به طوری که در بسیاری از پژوهش‌ها این دو اصطلاح را به جای هم استفاده می‌کنند. [3,4,5]

نخستین بار ارتباط میان عبارت‌های اسمی توسط کلرک در سال ۱۹۷۵ مطرح شد که آن‌را تحت عنوان پل زنی پیشنهاد می‌کرد. [6] پس از آن هیرست در سال ۱۹۸۱، عبارت پیشنهادی را بیشتر به منظور تفسیر مخفف‌ها در متن بکار برد. هیرست معتقد بود پیشنهاد یک ابزار ساخت، در گفتمان است که با استفاده از آن یک مخفف به یک یا چند عبارت غیر مخفف که پیش از آن در متن آمده‌اند، اشاره می‌کند تا به این ترتیب با استفاده از عبارت‌های پیشین، عبارت مخفف تفسیر شود. مفهوم پیشنهاد با یک جفت عبارت همراه است که عموماً عبارت ارجاع‌دهنده را با عنوان «تالی^{۲۲}» و عبارت مفسر را با عنوان «مقدم^{۲۳}» یا «مرجع» می‌شناسیم. هیرست این دو عبارت ارجاع‌دهنده و مرجع را با هم «هم-ارجاع^{۲۴}» و فرآیند شناسایی یک مقدم برای یک تالی، را «تحلیل^{۲۵}» می‌نامد. [7]

برای اولین بار هیرسکمن و چینکور اصطلاح هم‌مرجع را کنفرانس MUC^{۲۶} استفاده کردند [8,4] آنها و بسیاری از پژوهشگران دیگر، زمانی دو عبارت را با یکدیگر هم‌مرجع در نظر می‌گیرند که هر دو عبارت دقیقاً به یک موجودیت در دنیای واقعی اشاره داشته باشند. البته این پژوهشگران تفاوتی میان دو رابطه‌ی هم‌مرجعی و پیشابندی در نظر نگرفته بودند، تا اینکه این مسئله با انتقاد ون و کبیل مواجه شد به طوری که آن‌ها این دو اصطلاح را از منظری دیگر مورد بررسی قرار دادند. [9]

۲-۱- ارتباط هم‌مرجع:

ارتباط میان دو عبارت اسمی مانند a_1 و a_2 زمانی به عنوان هم‌مرجع در نظر گرفته می‌شود که هر دو عبارت اسمی به عنوان ارجاع‌هایی مشخص، یک موجودیت فرا زبانشناسی واحد را تفسیر نمایند. به واسطه‌ی این تعریف نیاز است تا به طور جداگانه مرجع هر کدام از عبارت اسمی $Reference(a_i)$ شناسایی شود. در نهایت یک ارتباط هم‌مرجعی به صورت ذیل خواهد بود:

(۱) عبارت‌های اسمی a_1 و a_2 با یکدیگر هم‌مرجع هستند اگر و تنها اگر $Reference(a_1) = Reference(a_2)$ باشد

مثال ۲: (باراک اوباما) Ant_1 ، به سوریه سفر کرد. (او) Ana_1 گفت...

۲-۲- ارتباط پیشابندی:

یک ارتباط پیشابندی^{۲۷}، از یک جفت عبارت اسمی (a_2, a_1) تشکیل می‌شود به طوری که عبارت اول نقش مفسر را برای عبارت دوم ایفا نماید. بنابراین یک ارتباط پیشابندی به صورت ذیل تعریف می‌شود:

(۲) عبارت اسمی اول (a_1) به عنوان مقدم برای عبارت اسمی دوم (a_2) در نظر گرفته می‌شود اگر و تنها اگر a_2 برای تفسیرش به a_1 وابسته باشد

مثال ۳: هنگامی که پسر وارد (اتاق) Ant شد، (درب) Ana به طور خودکار بسته شد.

نمونه ACE^{18} ، یک گروه هفت موجودیتی برای موجودیت‌ها (شخص، سازمان، مکان، سیاسی، تسهیلات، سلاح و خودرو) و تعداد زیادی زیرگروه (به عنوان مثال شخص: فرد، گروه) و کلاس برای هر موجودیت پیشنهاد کرده است [1] که اغلب پژوهشگران همه این موجودیت‌ها و یا گاهی برخی از آن‌ها را مورد بررسی قرار می‌دهند.

یکی از ویژگی‌های تحلیل مرجع مشترک این است که علاوه بر انواع موجودیت‌های رایج، می‌توانیم در حوزه‌های متفاوت از تعاریف پیش فرض خود نیز برای موجودیت‌ها نیز استفاده نماییم. همین ویژگی موجب شده است تا برخی از پژوهشگران مانند [2] به تحلیل مرجع مشترک در متون پزشکی پرداخته و بررسی موجودیت‌هایی مانند انواع دارو، بیماری، ژن و ... را هدف پژوهش خود قرار دهند.

با توجه به آنچه تا کنون گفته شد، انتظار می‌رود که با بررسی مراجع مشترک در مثال ۱، عبارت «یک محقق» نیز به همراه «او» و «سارا شکری» در یک زنجیره واحد قرار گیرد، اما خروجی ماژول تحلیل مرجع مشترک چنین نیست. هر چند از نظر ما این ارتباط کاملاً بدیهی است اما واقعیت این است که عبارت «یک محقق» به عنوان ارجاع به موجودیت شخص (در مثال ۱: سارا شکری) که در دنیای واقعی زندگی می‌کند در نظر گرفته نمی‌شود چون منظور از «یک محقق» می‌تواند هر شخص دیگری نیز باشد. در این حالت فرآیند دیگری تحت عنوان تحلیل پیشنهادی می‌تواند ارتباط میان «یک محقق» و «سارا شکری» را تشخیص دهد.

همان طور که مشاهده شد، تحلیل مرجع مشترک و تحلیل پیشنهادی دو مفهوم نزدیک به هم می‌باشند به طوری که عموماً به موازات تحلیل مرجع مشترک، با تحلیل پیشنهادی روبرو می‌شویم و حتی برخی به اشتباه این دو عبارت را معادل یکدیگر می‌پندارند. با وجود اینکه این دو ماژول از بسیاری از جهات با یکدیگر مشابه هستند، اما از جهاتی نیز با یکدیگر تفاوت دارند، و عدم توجه به این مسئله موجب سردرگمی و ایجاد ابهام در تحلیل متن می‌گردد.

در این مقاله، هدف ما بررسی هر کدام از این فرآیندها و مطالعه برخی از شباهت‌ها و تفاوت‌های میان این دو فرآیند می‌باشد. در بخش ۲، هر کدام از اصطلاح‌های هم‌مرجع و پیشنهادی^{۱۹} را مورد بررسی قرار می‌دهیم و حوزه‌های اشتراک و افتراق آنها را مشخص می‌نماییم. در بخش ۳ و ۴، مروری کوتاه بر تحلیل مرجع مشترک^{۲۰} و تحلیل پیشنهادی^{۲۱} خواهیم داشت. بخش ۵ نیز از منظری دیگر، لزوم اجرای این دو فرآیند را در کنار یکدیگر بررسی می‌نماید و نهایتاً در آخر برخی از پژوهش‌های انجام شده در این راستا را معرفی می‌نماییم.

۲- هم‌مرجع در مقابل پیشنهادی

برای اینکه بخواهیم یک سیستم تحلیل مرجع مشترک و یا یک سیستم تحلیل پیشنهادی را انتخاب کنیم لازم است ابتدا دو پدیده زبان‌شناسی پیشنهادی و هم‌مرجع را به خوبی بشناسیم. شباهت‌ها و تفاوت‌های میان آن‌ها را درک کنیم. همان طور که پیش‌تر نیز اشاره شد، اغلب دو عبارت پیشنهادی و هم‌مرجع را به عنوان دو عبارت نزدیک یا گاهی مترادف

جدول ۱: مقایسه ویژگی های دو ارتباط هم مرجع و پیشایندی

پیشایندی	هم مرجعی
۱) یک رابطه ی پیشایندی ، متقارن نیست. به این معنا که اگر a_1 به عنوان مقدم برای a_2 در نظر گرفته شود عبارت اسمی a_2 نمی تواند متقابلاً نقش مقدم را برای عبارت اسمی a_1 ایفا نماید.	۱) یک رابطه هم مرجعی یک رابطه ی هم /وزی است به عنوان نمونه «او» و «باراک اوباما» هر دو به یک موجودیت اشاره دارند ۲) رابطه هم مرجعی یک رابطه ی متقارن است، به این معنا که اگر «او» با «باراک اوباما» هم مرجع است «اوباما» نیز با «او» هم مرجع می باشد
۲) رابطه ی پیشایندی یک جفت از عبارت های اسمی را در نظر می گیرد به طوری که در این رابطه یک عبارت نقش مفسر و دیگری نقش ارجاع دهنده را دارند . بنابراین به طور ذاتی این رابطه غیر متعدي می باشد	۳) ارتباط میان عبارت های اسمی هم مرجع، متعدي است به این معنا که اگر «او» با «باراک اوباما» هم مرجع است و «او» نیز با «رئیس جمهور آمریکا» هم مرجع باشد بنابراین حتماً «باراک اوباما» نیز با «رئیس جمهور آمریکا» هم مرجع می باشد.
۳) این رابطه به محتوای متن حساس است به این دلیل که تالی یا عبارت ارجاع دهنده برای اینکه مفهوم باشد نیاز به مفسر یا تفسیر دارد	۴) ارتباط هم مرجعی یک ارتباط فرا متنی است به این معنا که تفسیر هیچ کدام از عبارت های اسمی هم مرجع به محتوای متن وابسته نمی باشد. به عبارت دیگر، دو عبارت هم مرجع لزوماً برای تفسیر به یکدیگر وابسته نیستند
۴) معمولاً تالی یا عبارت ارجاع دهنده در یک رابطه ی پیشایندی، به دیگر اجزای تشکیل دهنده در همان جمله و یا نزدیک ترین اظهارات ممکن در گفتمان اشاره دارد	۵) رابطه ی هم مرجعی خالص می تواند در تحلیل مرجع مشترک در متون متقاطع بکار رود
۵) یک تالی تنها نیاز به یک مفسر دارد تا آن را تفسیر نماید بنابراین اگر چندین مفسر برای آن در متن وجود داشته باشد مناسب ترین و نزدیک ترین آن ها انتخاب خواهد شد.	۶) برخی از عبارت های اسمی ذاتاً با یکدیگر هم مرجع هستند به عنوان مثال «پرفسور زاده» و «بنیان گذار منطق فازی»، این دو عبارت هر کجا که باشند به یک موجودیت به نام «پرفسور زاده» اشاره خوانند داشت. به عبارت دیگر چنین ارتباطی یک ارتباط هم مرجعی خالص نامیده می شود
۶) به درک بهتر متن و رفع ابهام برخی از واژگان در متن کمک می کند	۷) بیشتر برای استخراج اطلاعات در خصوص یک یا چند موجودیت در متن به کار می رود

را با در نظر گرفتن چند حوزه مشخص نمایش می دهد. ما این حوزه ها را با توجه به مفاهیم هم مرجعی و پیشایندی تفسیر خواهیم نمود. دو عبارت اسمی در صورتی که با یکدیگر مرتبط باشند، می توانند هم دارای ارتباط هم مرجعی و هم دارای ارتباط پیشایندی باشند. مانند آنچه در مثال (۴) آمده است.

مثال ۴: (علی کریمی) Ant.1 در نیمه دوم، ۱۰ گل به ثمر رساند. به طوری که به عنوان (محبوب ترین بازیکن پرسپولیس) Ana.1 در این بازی تشویق شد.

در میان عبارت های اسمی ای که دارای چنین ارتباط دوجانبه ای هستند می توان به برخی از ضمایر و مراجع آن ها نیز اشاره کرد. البته مثال های (۵) و (۶) نشان می دهند که همه ضمایر و مراجعشان در این رابطه صادق نیستند.

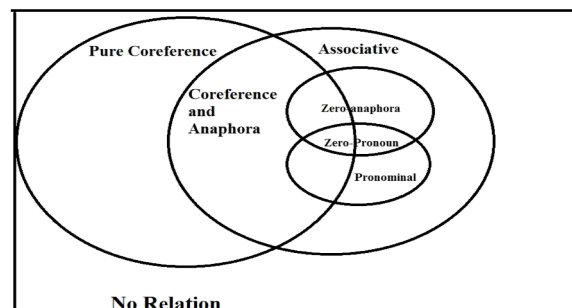
مثال ۵: (هر سگی) Ant(زندگی (خودش) Ana را دارد

مثال ۶: اگر (یک خانه) Ant داشتیم (آن) Ana را به تو اجاره می دادم
در مثال (۵)، ارتباط میان «هر سگی» و «خودش» یک ارتباط پیشایندی از نوع «ارتباط دهنده»^{۲۹} یا «باهم آیی» می باشد. و یا در مثال (۶)، «یک خانه» در دنیای واقعی وجود خارجی ندارد. بنابراین طبق تعریف ارتباط هم مرجعی، «یک خانه» نمی تواند به عنوان هم مرجع برای «آن» در نظر گرفته شود. [11] همان طور که در شکل (۱) مشاهده می شود، عبارت های اسمی که از نوع پیشایندی هستند اما ارتباط هم مرجعی ندارند در حوزه

همان طور که در مثال (۳) مشاهده می شود، دو عبارت اسمی فوق با یکدیگر هم مرجع نمی باشند اما دارای یک مرجع گفتاری مشابه هستند یا به تعبیر [10] با یکدیگر هم-واژگان^{۲۸} هستند. در این مثال ، عبارت اسمی «درب» به عبارت اسمی «اتاق» وابستگی دارد به عبارت بهتر، منظور از «درب»، درب اتاق ذکر شده است نه درب ساختمان یا درب خودرو یا هر درب دیگری.

در جدول (۱) برخی از ویژگی های هر کدام از روابط پیشایندی و هم مرجعی را قرار داده ایم. توجه به این ویژگی ها به درک بهتر نقاط تشابه و اختلاف این دو رابطه کمک فراوانی می نماید

به طور کلی، دو عبارت اسمی می توانند به طور هم زمان هم رابطه ی هم مرجعی و هم رابطه ی پیشایندی داشته باشند. البته این بدین معنا نیست که هرگاه دو عبارت اسمی با یکدیگر هم مرجع باشند، با یکدیگر ارتباط پیشایندی نیز داشته باشند و یا بالعکس.



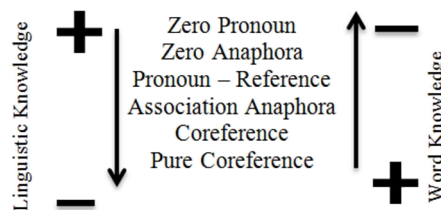
شکل ۱: انواع حوزه های ممکن برای ارتباط میان دو عبارت اسمی

مستندات متقاطع مشاهده می‌شود. با تکیه بر این مفهوم برخی از پژوهشگران مانند [17] به تحلیل مرجع مشترک در متون متقاطع پرداخته‌اند.

نکته‌ی دیگر اینکه، در همه پژوهش‌هایی که برای شناسایی مرجع ضمیر در زبان‌هایی مانند انگلیسی انجام شده است، همیشه مرجع ضمیر پیش از ضمیر قرار می‌گیرد. اما در زبان فارسی این امکان وجود دارد که مرجع یک ضمیر پس از آن قرار گیرد. مثال (۸)، نمونه‌ای از زبان فارسی می‌باشد که این مسئله را نشان می‌دهد.

مثال ۸: سی.ان.ان به آمریکا و جهان چنان شوکی وارد کرد که حتی (خود) Ana, I (ترنر) Ant, I نیز آن‌را باور نمی‌کرد.

بطور کلی، آنچه مبرهن است برای شناسایی و بررسی هر کدام از روابط پیشاندی و یا هم‌مرجعی، نیاز به دانش‌های واژگانی و زبان شناسی در زبان مورد نظر می‌باشد. به طوری که هر چقدر یک رابطه هم‌مرجعی بیشتر پیشاندی باشد به دانش زبان شناسی بیشتری نیاز دارد. در همین راستا، شکل (۲)، میزان ارتباط میان حوزه‌های گفته شده را با هر کدام از انواع دانش‌های زبان شناسی و واژگانی نمایش می‌دهد.



شکل ۲: میزان ارتباط میان حوزه‌ها با انواع دانش‌های زبان شناسی و واژگانی

۳- تحلیل پیشاندی:

یکی از وظایف مهم پردازش زبان طبیعی است که موجب کاهش ابهام عبارت‌های اسمی به کار رفته در متن می‌شود. به طور کلی این فرآیند وابستگی و پیوند میان دو عبارت تالی و مقدم را به گونه‌ای تشخیص می‌دهد که عبارت مقدم، عبارت تالی را تفسیر نماید.

۴- تحلیل مرجع مشترک:

با توجه به آنچه گفته شد، تحلیل مرجع مشترک، به فرآیند مرتبط کردن تمام عبارت‌هایی گفته می‌شود که به یک موجودیت واحد در دنیای واقعی اشاره دارند. این عبارت‌های اسمی هم‌مرجع، با هم تشکیل یک زنجیره واحد را می‌دهند که شامل ضمائر صفر تا اسمی خاص خواهد بود. تمام اعضای تشکیل دهنده‌ی این زنجیره باید از نظر تعداد، جنس و .. با یکدیگر هم تراز باشند.

۵- تقابل تحلیل مرجع مشترک و تحلیل پیشاندی:

با توجه به آنچه تا کنون درباره دو مفهوم تحلیل مرجع مشترک و تحلیل پیشاندی گفته شد، این دو وظیفه با وجود شباهت زیادی که با یکدیگر

پیشاندی‌های ارتباط‌دهنده قرار می‌گیرند [12, 13] اغلب پژوهشگران این نوع پیشاندی را زمانی به کار می‌برند که در جفت عبارت (مقدم و تالی)، عبارت مقدم به وضوح در متن ظاهر نشده باشد و با استفاده از برخی از موجودیت‌هایی که قبلاً در متن اشاره شده است می‌توان وجود آن‌ها را ثبت کرد. مثال (۳)، یک نمونه مناسب برای این نوع از پیشاندی‌ها می‌باشد. در این مثال ارتباط میان دو عبارت «درب» و «اتاق» از نوع مرونی‌میک^۲ است به این معنا که «درب» جزئی از «اتاق» محسوب می‌شود.^۲ یکی دیگر از انواع ارتباط پیشاندی، پیشاندی صفر^۳ است. مثال (۷) نمونه‌ای از این ارتباط را نشان می‌دهد.

مثال ۷: (دو جاده به سمت اصفهان) Ant می‌رود، (یکی اتوبانی و سریع) Gap1 و (دیگری خاکی و خطرناک) Gap2 است.

در حقیقت پیشاندی‌های صفر، یکی از انواع پیشاندی‌های کاهش یافته می‌باشند به این ترتیب که لزوماً به صورت عبارت‌های اسمی نیستند بلکه به صورت یک شکاف در یک عبارت ظاهر می‌شوند. [14] در مثال (۷) منظور از «اتوبانی و سریع» و «خاکی و خطرناک»، دو جاده‌ای است که به سمت اصفهان می‌روند. هدف از تحلیل پیشاندی‌های صفر، شناسایی چنین روابطی است.

یکی از حالت‌های ارتباط پیشاندی صفر، ضمائر صفر^{۳۲} می‌باشد که تحلیل آن در برخی زبان‌ها مانند ژاپنی [15]، اسپانیایی [16] و حتی فارسی بسیار حیاتی است. در این نوع از پیشاندی‌ها، عبارت اسمی مقدم حذف می‌شود. نمونه‌ای از این نوع پیشاندی، مثال (۶) می‌باشد که منظور از گوینده چنین است: «اگر من یک خانه داشتم آن‌را به تو اجاره می‌دادم»

در نهایت دو عبارت اسمی می‌توانند تنها از نوع هم‌مرجع خالص^{۳۳} باشند. ما این ارتباط را با مثال «پرفسور زاده»، «پدر منطق فازی» و «بنیان‌گذار منطق فازی» تفسیر می‌نماییم. همان طور که می‌دانید هر سه عبارت مذکور به شخص «پروفسور لطفعلی عسگرزاده» اشاره دارند اما اگر عبارت اسمی اول، یعنی «پرفسور زاده» در متن عنوان نشود، باز هم می‌دانیم که عبارت‌های اسمی «پدر منطق فازی» و «بنیان‌گذار منطق فازی» با یکدیگر هم‌مرجع هستند. نمونه‌ی دیگری از یک ارتباط هم‌مرجع خالص، ارجاع انواع عبارت‌های اسمی به «باراک اوباما»، در پاراگراف‌ها، مکالمات با حتی متون متمایز است. اغلب نمونه‌هایی از روابط هم‌مرجع خالص در

^۲ یکی از ویژگی‌های مهم برای تحلیل مرجع مشترک و تحلیل پیشاندی استفاده از شبکه واژگان (WordNet) می‌باشد یک شبکه واژگان در کنار تعریف واژه، اطلاعاتی در خصوص چهار بخش گفتاری (اسم، صفت، فعل و قید) را در اختیار ما قرار می‌دهد. همچنین برخی از ارتباطات لغوی را میان واژگان مشخص می‌کند از جمله این ارتباطات Antonymy، Meronymy و Hypernymy است.

Antonymy: رابطه‌ی میان عبارت‌هایی که از نظر معنایی با یکدیگر متضاد هستند را نشان می‌دهد

Hypernymy: یک واژه مانند انسان به عنوان Hypernym بازیکن در نظر گرفته می‌شود چون انسان به عنوان یک ابر کلاس برای بازیکن می‌باشد.

Meronymy: زمانی اتفاق می‌افتد که یک واژه یک بخش از واژه‌ای دیگر را نشان دهد به عنوان مثال درب Meronym از اتاق است. برعکس این رابطه Holonymy نامیده می‌شود [25].

ایست که در آن دقت^{۳۵} تحلیل پیشابندی از تحلیل مرجع مشترک پیشی می‌گیرد.

طبق آنچه در فرآیند تحلیل پیشابندی رایج است. هنگام شناسایی مقدم برای هر عبارت تالی، حرکت از تالی به سمت عبارت‌های اسمی ماقبل آن انجام خواهد شد تا به این ترتیب، جفت عبارت‌های (مقدم و تالی) شناسایی شوند. همان طور که پیش از این اشاره شد، فرآیند تشخیص مرجع ضمیر نیز از همین رویکرد پیروی می‌نماید، با این تفاوت که در برخی موارد مقدم (در اینجا منظور مرجع ضمیر است) بعد از تالی قرار می‌گیرد (مانند مثال ۸). بنابراین حرکت الگوریتم (راست به چپ) نمی‌تواند به شناسایی مرجع صحیح برای ضمیر مورد نظر منجر شود. در حالی که اجرای وظیفه‌ی تحلیل مرجع مشترک، با شناسایی عبارت‌های هم‌مرجع، تا حد زیادی این مشکل را برطرف خواهد نمود. حرکت الگوریتم در تحلیل مرجع مشترک نیز از راست به چپ خواهد بود با این تفاوت که نقطه شروع آن می‌تواند انتهای متن باشد، هر عبارت اسمی جدید که با عبارت‌های پیش از خود هم‌مرجع باشد به زنجیره‌های موجود اضافه می‌شود، در غیر این صورت خودش تشکیل زنجیره‌ای جدید را خواهد داد. در چنین حالتی فراخوانی و دقت^{۳۶} در تحلیل مرجع مشترک نسبت به تحلیل پیشابندی پیشی می‌گیرد.

به همین ترتیب نیز راهکارهای ارائه شده برای هر کدام از این فرآیندها تا حدودی با یکدیگر متفاوت می‌باشد. به عنوان نمونه یک رویکرد در فرآیند تحلیل پیشابندی افزاز پس از آموزش الگوریتم یادگیری است که تنها به منظور افزاز جفت‌های (مقدم و تالی) صورت می‌گیرد. به این ترتیب که پس از ایجاد نمونه‌های مثبت و منفی که عموماً توسط یک رده‌بند ایجاد می‌شوند، نزدیک‌ترین و مناسب‌ترین مقدم برای تالی مورد نظر مشخص می‌شود. در مقابل برخی پژوهشگران برای تحلیل مرجع مشترک دو اصطلاح تالی و مقدم را نادیده گرفته و به جای آن اصطلاحاتی مانند «قابل برچسب^{۳۷}» و «اشاره^{۳۸}» را جایگزین کرده‌اند. توجه این جایگزینی این است که ما در تحلیل مرجع مشترک به مفاهیم مفسر یا ارجاع دهنده نیاز نداریم و به جای آن اصلاح مستقل اشاره را جایگزین می‌نماییم. هر اشاره به عنوان یک ارجاع به هر گروه از موجودیت‌ها در نظر گرفته می‌شود. هر اشاره می‌تواند یکی از انواع اسم عام، اسم خاص و حتی ضمیر (یا شبه ضمیر) باشد. همچنین سایر عبارت‌های موجود در متن که برای تحلیل مرجع مشترک کاربرد ندارند، به عنوان خارج از اشاره^{۳۹} در نظر گرفته می‌شوند. به طور خلاصه و مفید برخی از روش‌های استفاده شده برای این دو فرآیند را در جدول (۲) آورده‌ایم.

دارند، دو وظیفه مجزا محسوب می‌شوند. بدین ترتیب که با به‌کارگیری فرآیند تحلیل مرجع مشترک، سیستم می‌تواند تمام اطلاعاتی که در مورد یک موجودیت در متن وجود دارد را شناسایی نماید. از طرفی دیگر تفسیر درست پیشابندها در متون به رفع ابهامات موجود در متن می‌انجامد. لذا استفاده از این دو وظیفه با در نظر گرفتن حوزه‌های تشابه و اختلاف آن‌ها به درک و استخراج اطلاعات مفید در متن کمک فراوانی خواهد نمود. در این بخش با استفاده از مثال‌هایی لزوم اجرای این دو فرآیند را در کنار یکدیگر بررسی می‌کنیم.

پس از شناخت فرآیندهای تحلیل مرجع مشترک و تحلیل پیشابند و اینکه یک سیستم چه انتظاراتی را باید برآورده نماید، لازم است تا روال کار را برای اجرای هر فرآیند را مشخص نماییم. استفاده از الگوریتم‌های یادگیری، از پرکاربردترین روش‌های ارائه شده برای شناسایی روابط مثبت و منفی در مرجع مشترک و پیشابند می‌باشد. به این ترتیب که الگوریتم با استفاده از برخی ویژگی‌ها (مانند سون ۲۰۰۱) آموزش می‌بیند تا روابط مثبت و منفی را پیش‌بینی نماید^{۴۰}. البته شایان ذکر است که برخی از ویژگی‌ها برای این دو فرآیند مشترک و برخی دیگر متفاوت می‌باشد. به عنوان مثال ویژگی تطابق رشته‌ای^{۴۱} از جمله مواردی است که برای تحلیل مرجع مشترک بسیار مورد توجه قرار گرفته است اما در فرآیند تحلیل پیشابند کاملاً بی‌معنا است (دو عبارت یکسان چگونه می‌توانند یکدیگر را تفسیر نمایند). به عبارت دیگر فرض بر این است که دو عبارت اسمی که در یک متن تکرار می‌شوند به احتمال خیلی زیاد هم‌مرجع هستند. مثال (۹)، نقض این مسئله را نشان می‌دهد.

مثال ۹: دیروز دو تیم محبوب «الف» و «ب» به مصاف یکدیگر رفتند. تیم «الف»، در نیمه اول بسیار درخشید به طوری که طرفداران این تیم به پیروزی تیمشان امیدوار بودند. اما دیری نپایید که تیم «ب» روال بازی را به نفع خود تغییر داد که منجر به شادمانی طرفداران این تیم شد. همان طور که در مثال (۹) مشاهده می‌شود، عبارت اسمی «طرفداران این تیم» در خطوط ۲ و ۳ عیناً تکرار شده است. بنابراین طبق ویژگی تطابق رشته‌ای در فرآیند تحلیل مرجع مشترک این دو عبارت با یکدیگر هم‌مرجع در نظر گرفته خواهند شد، اما واقعیت این است که این دو عبارت با یکدیگر هم‌مرجع نیستند. به عبارتی قرار گرفتن این دو عبارت در یک زنجیره واحد یکی از خطاهای تحلیل مرجع مشترک محسوب می‌شود. در حالی که استفاده از فرآیند تحلیل پیشابندی موجب رفع این خطا خواهد شد. این کار به این ترتیب انجام خواهد گرفت که عبارت «طرفداران این تیم» در خط ۳ با اولین و نزدیک‌ترین مفسر خود یعنی «تیم پرسپولیس»، تشکیل یک جفت تالی و مقدم را می‌دهند و به همین ترتیب «طرفداران این تیم» در خط ۲ با «تیم استقلال» جفت خواهد شد. این مثال نمونه

^{۳۵} منظور از نمونه‌های مثبت و منفی، جفت کردن تالی با عبارت‌های اسمی ممکن است،

به عنوان نمونه جمله‌ی «سارا به مدرسه‌ی سابق خود رفت که او را یاد کودکی اش می‌انداخت.» را در نظر بگیرید. او و سارا نمونه‌های مثبت هستند. اما او و مدرسه، او و مدرسه سابق، او و سابق از جمله نمونه‌های منفی ممکن می‌باشند.

name	learning framework	Decoding
Cai et al, ۲۰۱۱	Compute hyperedge weights on ۳۰% of training data	Recursive ۲-way Spectral clustering (Agarwal, ۲۰۰۵)
sapena et al, ۲۰۱۰	Decision Tree + Relaxation Labeling	Iterative
stoyanov et al, ۲۰۱۰	Averaged perceptron	Single-link clustering by computing transitive closure between pairwise positives.
nugue et al	Logistic Regression (LIBLINEAR)	Closest-first clustering for pronouns and Best-first clustering for non-pronouns
santos et al	ETL (Entropy guided Transformational Learning) committee and Random Forest (WEKA)	Limited number of preceding mentions ۶۰ for automatic and ۴۰ given gold boundaries; Aggressive-merge clustering (Mccarthy and Lenhart, ۱۹۹۵)
song et al	MaxEnt (OpenNLP)	Pre-clusters, with singleton pronoun pre-clusters, and use closest-first clustering. Different link models based on the type of linking mentions – NP-PRP, PRP-PRP and NP-NP
kobdani et al ۲۰۱۱	Decision Tree	Best-first clustering. Threshold of ۱۰۰ words used for long documents
yang et al	MaxEnt (MALLET)	Maximum ۲۳ sentences to the left; Constrained clustering
xinxin et al	ILP/Information gain	Best-first clustering followed by ILP optimization
Bunescu et al, ۲۰۱۲	Rule-based (lee et al ۲۰۰۱)	Adaptive clustering
Denic et al, ۲۰۰۹	ILP/Information gain	closest-rst link-clustering
bansal et al ۲۰۱۲	Stoyanov et al. (۲۰۰۹). + Decision tree	Web-scale distributional clustering
Chen et al ۲۰۱۱	Maximum Entropy	best-first clustering
lee et al ۲۰۱۱	Rule-based	Multi-pass Sieves
chang	Learning Based Java	Best link and All links strategy; with and without constraints – Best link without constraints was selected for the official run
uryupina	Decision Tree. Different classifiers for Pronominal and non-Pronominal mentions	mention pair model without ranking as in Soon ۲۰۰۱
sobha	CRF for non-pronominal and salience factor for pronominal resolution	Pronominal: all preceding NPs in the sentence and preceding ۴ sentences
klenner	Rule-based. Salience measure using dependencies generated from training data	Incremental entity creation
zhou	SVM tree kernel using BC portion of the data	
charton	Multi-layer perceptron	MLP with score of ۰.۵ used for linking and ۱۰ mentions
hao	MaxEnt	Beam search (Luo, ۲۰۰۴)
zhang	SVM	Window of ۱۰۰ markables
kummerfield	Unsupervised generative model	Pre- and post- resolution filters
zhokova	TIMBL memory based learner	From last possible mention in document

۶- نتیجه گیری

یک ارتباط پیشابندی میان دو عبارت حاکم است. اینگونه از پژوهش ها به دنبال شناخت موجودیت های بکار رفته در متن می باشند و عموماً هر عبارت اسمی را در قالب یک اشاره در نظر می گیرند به به یک موجودیت واحد در دنیای واقعی ارجاع دارد.

از طرفی دیگر گاهی تفسیر و رفع ابهام عبارت های بکار رفته در متن برای ما بسیار اهمیت دارد مانند آنچه در ترجمه ماشینی یک زبان نیاز داریم لذا در چنین حالتی تحلیل ارتباطات پیشابندی می تواند بسیار موثر واقع شود. ما باور داریم شناخت و توجه انواع ارتباطات میان عبارت های اسمی بکار رفته در متن می تواند به درک متن و رفع برخی از ابهامات موجود کمک فراوانی نماید. لذا استفاده از این دو فرآیند به طور همزمان می تواند در پردازش متن موثر واقع شود. از طرفی دیگر، شناخت دقیق این روابط برای انتخاب راه کار های مناسب برای تحلیل آنها نیز مفید است چراکه این روابط می تواند برای هر زبانی ویژگی های خاص خود را داشته باشد.

مراجع

- [1]. ACE (Automatic Content Extraction) English Annotation Guidelines for Entities. Version 6.06 2008.06.13
- [2]. Tian Ye he, "Coreference Resolution on Entities and Events for Hospital Discharge Summaries", Master thesis, MIT University, 2007

امروزه فرآیند تحلیل مرجع مشترک به عنوان یکی از کلیدی ترین وظایف پردازش زبان طبیعی مورد توجه بسیاری از پژوهشگران قرار گرفته است. این فرآیند به دنبال مرتبط کردن عبارت های اسمی می باشد که همه به یک موجودیت واحد ارجاع دارند.

هنگامی که بخواهیم برای زبان های با منابع کم، مانند زبان فارسی، چنین سیستمی را پیاده سازی کنیم، با موانع فراوانی از جمله نبود سیستم های شناسایی موجودیت نامدار، کشف اشاره و به خصوص تجزیه گر روبرو می شویم. یکی از روش های مناسب در چنین شرایطی استفاده از پیکره های برجسب گذاری شده می باشد و از آنجاییکه چنین پیکره ای برای تحلیل مرجع مشترک در زبان فارسی وجود ندارد لازم است تا شناخت بیشتری نسبت به ماهیت اصطلاح هم مرجع داشته باشیم تا با تکیه بر این شناخت بتوانیم عبارت های اسمی را به درستی برجسب گذاری نمائیم. اولین چالشی که با آن مواجه شدیم، تشابه یک رابطه ی هم مرجع با یک رابطه ی پیشابندی بود، از اینرو، در این مقاله با بررسی دقیق اصطلاح هم مرجع و مقایسه ی آن با اصطلاح بسیار نزدیک و مشابه، و البته متمایز پیشابند تصمیم داریم تا این دو مفهوم را به خوبی بشناسیم.

بررسی پژوهش های انجام شده در دو دهه اخیر، نشان می دهد که اغلب پژوهشگران این دو عبارت را به طور مترادف در نظر می گیرند یا به عبارت بهتر نقطه تشابه این عبارت را هدف پژوهش خود قرار می دهند. از طرفی دیگر گاهی ارتباط میان دو عبارت اسمی، از نوع هم مرجع نیست و تنها

- [16]. Recasens, M, Marti A, Taule M , “ where anaphora and coreference meet. Annotation in the Spanish CESS-ECE corpus” ,۲۰۰۶
- [17]. Sayeed Asad, Elsayed Tamer, Garera Nikesh ,”Arabic Cross-document Coreference Detection”, 2009
- [18]. Kibble, R.: ۲۰۰۰, Coreference annotation: Whither?, Proceedings of the Second International Conference on Language Resources and Evaluation (LREC- ۲۰۰۰), pp. ۱۲۸۱{۱۲۸۶.
- [19]. Soon, W., Ng, H. and Lim, D.: 2001, A machine learning approach to Coreference resolution of noun phrases, Computational Linguistics 27(4), pages 521-544.
- [20]. Cai J , Mujdricza-Maydt E and Strube M , “Unrestrited Coreference Resolution via Global HyperGraph Partitioning” Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 56–60, Portland, Oregon, 23-24 June 2011
- [21]. Kobdani . H, Schutze H, Schiehlen M, Kamp H ,”Bootsrapping Coreference Resolution Using Word Association”, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 783–792, Portland, Oregon, June 19-24, 2011.
- [22]. Stoyanov V, Cardie C , Gilbert N , Buttlar D , “Coreference Resolution with Reconcile” , Proceedings of the ACL 2010 Conference Short Papers, pages 156–161, Uppsala, Sweden, 11-16 July 2010
- [23]. .Bansal, M and Klein,D , “Coreference Semantics From Web Features”, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 389–398, Jeju, Republic of Korea, 8-14 July 2012.
- [24]. Chen, Weipeng , Zhang, M , Qin, B ,” Coreference Resolution System using Maimum Entropy Classifier” Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 127–130, Portland, Oregon, 23-24 June 2011.
- [25]. Sapena E , Padro L and Turmo J ,”RelaxCor: Aglobal Relaxation Labeling Approach to Coreference Resolution” , Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 88–91, Uppsala, Sweden, 15-16 July 2010.
- [3]. Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L. and Sotirova, V.: ۲۰۰۰, Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies, Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC-۲۰۰۰), pp. ۴۹{۵۸.
- [4]. Hirschman, L. and Chinchor, N.: ۱۹۹۸, Muc-۷ coreference task de_nition. Version ۳, Proceedings of the Seventh Message Understanding Conference (MUC-۷).
- [5]. Davies, S., Poesio, M., Bruneseaux, F. and Romary, L.: ۱۹۹۸, Annotating coreference in dialogues: Proposal for a scheme for mate. <http://www.hcrc.ed.ac.uk/poesio/MATE/annomanual.htm>.
- [6]. Herbert H Clark. 1975. Bridging. In Proc. of the Conference on Theoretical Issues in Natural Language Processing, pages 169–174.
- [7]. Hirst, G.: ۱۹۸۱, Anaphora in natural language understanding: A survey, Lecture Notes in Computer Science, Vol. ۱۱۹, Springer-Verlag Berlin Heidelberg New York.
- [8]. Chinchor, Nancy A. 1998. Overview of MUC-7/MET-2. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. 894.02/related projects/muc/proceedings/muc 7 toc.html.
- [9]. van Deemter, K. and Rodger Kibble. ۲۰۰۰. On coreferring: coreference in MUC and related annotation schemes. *Computational Linguistics*, ۲۶(۴):۶۲۹–۶۳۷.
- [10]. Recasens, M, Marti A, Taule M , “ where anaphora and coreference meet. Annotation in the Spanish CESS-ECE corpus” ,2006
- [11]. Elango, P. Coreference resolution: a survey. University of Wisconsin, Madison.
- [12]. Bunesu Razvan, “Associative Anaphora Resolution:A Web-Based Approach” University of Texas at Austin,
- [13]. Sasano Ryohei and Kurohashi Sadao, “ A Probabilistic Model for Associative Anaphora Resolution, Kayoto University
- [14]. Ngai, G and Wang, C ,” A Knowledge-Based Approach for Unsupervised Chinese Coreference Resolution” , Computational Linguistics and Chinese Language Processing ,Vol. 12, No. 4, December 2007, pp. 459-484
- [15]. Sasano Ryohei and Kurohashi Sadao, “ A Discriminative Approach to Japanese Zero Anaphora Resolution With Large-scale Lexicalized Case Frames, Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 758–766, Chiang Mai, Thailand, November 8 – 13, 2011.

[۲۶] قادریان میثم، "بهیود مدل کاربر در وبسایت بصورت خودکار با استفاده از معنا شناسی با مفاهیم خاص دامنه" پایان نامه کارشناسی ارشد، ۱۳۸۷

-
- ¹ Entity
 - ² Nominal
 - ³ Proper Name
 - ⁴ Co-reference / Coreference
 - ⁵ Information Extraction
 - ⁶ Natural Language Processing
 - ⁷ Data Mining
 - ⁸ Question / Answering
 - ⁹ Text Understanding
 - ¹⁰ Summarizing
 - ¹¹ Information Retrieval
 - ¹² Tokenization
 - ¹³ POS Tagger
 - ¹⁴ Parser
 - ¹⁵ Named Entity Reorganization
 - ¹⁶ Mention Detection
 - ¹⁷ Coreference Resolution
 - ¹⁸ Automatic Content Extraction
 - ¹⁹ Anaphora
 - ²⁰ Coreference Resolution
 - ²¹ Anaphora Resolution
 - ²² Anaphor
 - ²³ Antecedent
 - ²⁴ coreferentially
 - ²⁵ Resolution
 - ²⁶ Message Understanding Initiative
 - ²⁷ Anaphoric
 - ²⁸ Co-Lexical
 - ²⁹ Associative
 - ³⁰ Meronymic
 - ³¹ Zero-Anaphora
 - ³² Zero-Pronoun
 - ³³ Pure Coreference
 - ³⁴ String-Matching
 - ³⁵ Precision
 - ³⁶ Recall
 - ³⁷ Markables
 - ³⁸ Mention
 - ³⁹ Out of Mention