

معرفی رویکردی ماشینی با استفاده از الگوریتم لسک و برچسب‌دهی نحوی جهت رفع ابهام از معنای کلمات

الهام علایی ابوذر *

دکتری زبانشناسی همگانی

استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

دریافت: ۹۵۰۹/۱۴ | پذیرش: ۹۶/۰۲/۱۱

چکیده:

پژوهش حاضر به معرفی رویکردی ماشینی برای چگونگی رفع ابهام معنایی از کلمات می‌پردازد. در زبان فارسی، که ساخت‌واژه پیچیده دارد، هم‌نگاره‌های بسیاری ساخته می‌شوند که معانی گوناگونی در بافت‌های گوناگون دارند. یکی از راه‌هایی که کمک می‌کند رفع ابهام از معنای کلمات مبهم (هم‌نگاره‌ها) با سهولت و دقت بیشتری انجام شود، تخصیص برچسب درست نحوی به کلمات است. بنابراین اگر برچسب‌دهی نحوی قبل از مرحله رفع ابهام معنایی از کلمات، صورت پذیرد، رفع ابهام معنایی از هم‌نگاره‌ها با دقت بیشتری انجام خواهد گرفت. از آنجائی که فراوانی هم‌نگاره‌های اسمی و صفتی در متون فارسی، در مقایسه با سایر هم‌نگاره‌ها، بالا است، پس از تخصیص برچسب نحوی به کلمات، لازم است رفع ابهام از برچسب نحوی هم‌نگاره‌ها نیز صورت گیرد. در این مقاله ابتدا روش‌های ماشینی موجود در جهت رفع ابهام از معنای کلمات معرفی می‌شود و سپس، الگوریتم لسک (که یکی از روش‌های یادگیری ماشینی بدون‌نظارت/ بدون-سرپرست برای رفع ابهام معنایی از کلمات مبهم موجود در متون گوناگون است) معرفی می‌شود و در نهایت رویکردی ماشینی جهت رفع ابهام از معنای کلمات با استفاده از نتیجه مرحله برچسب‌زنی نحوی به کلمات و رفع ابهام از برچسب نحوی کلمات و الگوریتم لسک معرفی می‌شود. انجام برچسب‌دهی نحوی و رفع ابهام از برچسب نحوی هم‌نگاره‌ها باعث می‌شود الگوریتم لسک تنها، معانی مرتبط با برچسب‌های نحوی را در رفع ابهام معنایی از کلمات در نظر گیرد و در نتیجه عمل رفع ابهام از معنای کلمات با دقت و سهولت بیشتری انجام پذیرد.

فصلنامه علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا(چاپی) ۲۲۵۱-۸۲۲۳
شاپا(الکترونیکی) ۲۲۵۱-۸۲۳۱
نمایه در SCOPUS و ISC
<http://jlist.irandoc.ac.ir>
دوره XX | شماره X | صص XX-XX
۱۳XX X

نوع مقاله: پژوهشی

به این مقاله به شکل زیر استناد کنید:

دورن متن:

(علایی ابوذر ، زودآیند)

در فهرست منابع:

علایی ابوذر، الهام. زودآیند. معرفی رویکردی ماشینی با استفاده از الگوریتم لسک و برچسب‌دهی نحوی جهت رفع ابهام از معنای کلمات پژوهشنامه پردازش و مدیریت اطلاعات. <http://jipm.irandoc.ac.ir> (دسترسی در روز/ماه/سال)

کلیدواژه‌ها: رفع ابهام از معنای کلمات، هم‌نگاره، برجسب‌دهی نحوی، الگو ریتم
لسک، رفع ابهام از برجسب نحوی کلمات

*پدیدآور رابط فونت Alayi@irandoc.ac.ir

۱. مقدمه

هنگام خواندن متون گوناگون فارسی، خوانندگان با چالش‌های مختلفی مانند انواع مترادف‌ها، متجانس‌ها، کلمات خارجی یا بومی و نظایر آن، مواجه می‌شوند؛ به عنوان مثال «ماه‌گرفتگی» و «خسوف» مترادف هستند و یا «تهران» و «طهران»، دو گونه نوشتاری صحیح در نظر گرفته می‌شوند (که نوعاً دومی در متون قدیم یافت می‌شود) و یا کلمات «سرگذشت‌نامه»، «قلب‌نگاری»، «پرتونگاری»، و «دیوان‌سالاری»، معادل‌های فارسی برای کلمات قرضی «بیوگرافی»، «کاردیوگرافی»، «رادیولوژی»، و «بوروکراسی» هستند، علاوه بر موارد ذکر شده، برخی کلمات عربی نیز بیش از یک صورت نوشتاری دارند مانند «زکات» و «زکوة»، یا «اسماعیل» و «اسمعیل» (سلطانی و فانی: ۱۳۶۲). اکثر مردم نسبت به این مسئله که چه اندازه زبان انسان مبهم است، آگاهی ندارند و در مواجهه با این مسئله که رایانه توانایی درک زبان و ارتباطات زبانی را، همانند آنچه انسان انجام می‌دهد، ندارد، ناامید می‌شوند. زبان‌شناسی رایانشی با چالشی جدی تحت عنوان

«ابهام»^۱ در تمام سطوح زبانی مواجه است. گوستاد معتقد است با *vagueness* و *ambiguity* تفاوت است (البته معادل فارسی هر دو واژه انگلیسی، «ابهام» است)؛ وی *vagueness* را حالتی در نظر می‌گیرد که یک کلمه یا عبارت، تنها یک معنی دارد که مرزهای آن به طور دقیق مشخص نشده است، در صورتی که *ambiguity*، حالتی است که یک کلمه می‌تواند چندین معنی در بافت‌های گوناگون داشته باشد. در زبان‌شناسی رایانشی وقتی سخن از ابهام و رفع ابهام از معنای کلمات و جملات است، منظور *ambiguity* است؛ گوستاد انواع گوناگون ابهام را ذکر می‌کند که از میان آن‌ها می‌توان به موارد زیر اشاره کرد (Gaugstad 2004):

الف. ابهام ساخت‌واژی^۲: یک کلمه با توجه به ساخت درونی خود می‌تواند مبهم باشد؛ کلمات مرکب، منبع معمول ابهام ساخت‌واژی هستند. به عنوان مثال، کلمه مرکب *computertaalkunde* در زبان هلندی را به دو طریق می‌توان تجزیه کرد: *computer-taalkunde* به معنی زبان‌شناسی رایانه‌ای و *computertaal-kunde* به معنی برنامه‌ریزی دانش زبانی.

ب. ابهام نحوی^۳: مثال بارز چنین ابهاماتی هنگامی رخ می‌دهد که یک گروه حرف اضافه‌ای به عناصر جمله اضافه می‌شود، مانند جمله:

The man saw the girl with the telescope

چنین جمله‌ای مبهم است و با توجه به اینکه گروه حرف اضافه‌ای «with the telescope» را به «the man» یا «the girl» نسبت دهیم، دو خوانش متفاوت دارد.

پ. ابهام کاربردی^۴: این نوع ابهام در بسیاری از زبان‌ها عمدتاً در تفسیر و تعیین مرجع ضمیر اتفاق می‌افتد. به عنوان مثال، جمله زیر را در انگلیسی در نظر بگیرید:

Mary`s mother is a gardener. John likes her

¹ ambiguity

² morphological ambiguity

³ syntactic ambiguity

⁴ pragmatic ambiguity

مرجع ضمیر her جمله دوم می‌تواند هم Mary باشد و هم Mary's mother.

ت. ابهام معنایی واژگانی^۱: این نوع ابهام زمانی اتفاق می‌افتد که یک کلمه (در واقع، یک صورت نوشتاری واحد) بیش از یک معنی داشته باشد. در متون قدیمی، اینگونه کلمات (که در واقع کلمات ظاهراً متشابهی هستند که چند مفهوم را می‌رسانند) مشترک لفظی خوانده می‌شوند و تفکیک اینگونه کلمات، فارغ از بافت، مستلزم استفاده از توضیح‌گر است (سلطانی و فانی: ۱۳۶۲). در همه زبان‌های دنیا، کلمات بسیاری وجود دارد که بسته به بافت، معانی گوناگونی دارند. به عنوان مثال، جملات زیر را در نظر بگیرید:

الف. هنگام شکار، شیر ابتدا طعمه را زیر نظر می‌گیرد و سپس هنگامی که طعمه تنها و جدا از گله است، به آن حمله می‌کند.

ب. از سوپر مارکت شیر هم بگیر.

در این دو جمله، صورت نوشتاری «شیر»، دو معنی متفاوت در دو بافت متفاوت دارد.

مقاله حاضر به بررسی چگونگی رفع ابهام معنایی واژگانی و ساخت‌واژی کلمات می‌پردازد. رفع ابهام از معنای کلمه^۲، فرآیندی برای تعیین معنای واقعی کلمه مبهم بر اساس موقعیت مشخص (بافت) است. به عنوان مثال، کلمه «شیر» معانی مختلفی دارد از جمله: «حیوانی درنده و گوشتخوار در جنگل»، «نوعی از لبنیات» و «وسیله‌ای که آب از آن جاری می‌شود»؛ چنین کلماتی ذاتاً مبهم هستند. فرآیندی که طی آن معنی مناسب یک کلمه مبهم در یک بافت خاص، مشخص می‌شود، رفع ابهام از معنای کلمه خوانده می‌شود. منظور از کلمات مبهم، کلماتی است که صورت نوشتاری یکسان، اما منشاء، معنی یا تلفظ متفاوت دارند و اصطلاحاً هم‌نگاره / هم‌نگاشت^۳ نامیده می‌شوند (فرهنگ لغت مریم وبستر^۴). در زبان‌هایی که ساخت‌واژه پیچیده دارند، مانند زبان فارسی، هم‌نگاره‌های بسیاری ساخته می‌شوند. هم‌نگارگی در زبان فارسی ناشی از بازنامایی واجی

¹ lexical semantic ambiguity

² word sense disambiguation (WSD)

³ homograph

⁴ Merriam Webster

و صرفی عناصر زبانی در خط فارسی است، به این صورت که رابطه‌ای چند-به-چند و در برخی موارد، غیرنظام‌مند میان عناصر واجی و صرفی زبان فارسی و تظاهر نوشتاری آن‌ها در خط فارسی دیده می‌شود. بی‌جن خان و مرادزاده (۱۳۸۳، نقل از محسنی: ۱۳۸۷) هم‌نگاره‌های فارسی را به چند دسته تقسیم می‌کنند که از میان آن‌ها می‌توان به موارد «الف» تا «ت» اشاره کرد:

الف. واژگانی: هم‌نگاره‌هایی هستند که در فرهنگ‌های لغات به صورت مدخل‌های مجزا ذکر شده‌اند مانند: «شیر» و «شیر»؛

ب. هم‌نگاره‌هایی که در اثر عدم نمایش علائم زیر و زبری در خط فارسی، ایجاد شده‌اند. مانند: «مرد» /mard/ و «مرد» /mord/.

پ. هم‌نگاره‌هایی که در اثر عدم تناظر یک-به-یک میان واج‌ها و نگاره‌ها در فارسی ایجاد شده‌اند. مانند: «رود» /rud/ و «رود» /ravad/.

ت. هم‌نگاره‌هایی که در اثر یکسانی تظاهر واجی و نوشتاری تکواژها به وجود می‌آیند. مانند: یکسان بودن نمود نوشتاری تکواژ یای نکره، یای اسم‌ساز (اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم معنی یا اشیا، تصغیر و تحییب، اسم مصدر یا حاصل مصدر)، شناسه دوم شخص مفرد و یای صفت‌ساز (صفت فاعلی و مفعولی، صفتی که دال بر نسبت است) که همه نمود نوشتاری «-ی» را دارند. به عنوان مثال، کلمه «آسمانی» در اثر اضافه شدن پسوند تصریفی نکره‌کننده (یای نکره) یا پسوند اشتقاقی صفت‌ساز، به وجود آمده است و فارغ از بافت، می‌تواند اسم نکره یا صفت باشد. از میان هم‌نگاره‌های ذکر شده، هم‌نگاره‌هایی که در اثر اضافه شدن پسوند «-ی» اسم‌ساز یا صفت‌ساز ایجاد می‌شوند، فروانی بالایی در پیکره‌های متنی فارسی دارند (علایی: ۱۳۹۵).

سیستم رفع ابهام از معنای کلمات به این صورت عمل می‌کند که ابتدا مجموعه‌ای از کلمات، مثلاً یک جمله، در اختیار سیستم قرار می‌گیرد و سیستم با استفاده از منابع اطلاعاتی که در دسترس آن قرار دارد، مناسب‌ترین معنی برای یک کلمه را در بافت مشخص می‌کند. منبع اطلاعات می‌تواند مجموعه‌ای از متون باشد (پیکره‌های متنی) که کلمات یا برجسب معنایی و نحوی دارند یا ندارند؛ همچنین منبع اطلاعات می‌تواند فرهنگ‌های لغات قابل خواندن برای ماشین باشد

(Sarrafzade et al 2016). بدون مجهز بودن به چنین اطلاعاتی، تشخیص ماشینی معنای هم-نگاره‌ها غیرممکن خواهد بود. در واقع گویشوران به دلیل مجهز بودن به اطلاعات زبانی از قبیل اطلاعات مربوط به ساخت‌واژه زبان، ساخت واجی و ساخت نحوی زبان، قادر به رفع ابهام از هم-نگاره‌ها در بافت نحوی هستند، اما سامانه‌های پردازش متون و رفع ابهام از معنای کلمات، چنانچه به چنین اطلاعات زبانی مجهز نشده باشند، با مشکلاتی در پردازش هم‌نگاره‌ها مواجه می‌شوند. یکی از راه‌هایی که کمک می‌کند رفع ابهام معنایی کلمات (هم‌نگاره‌ها) با سهولت و دقت بیشتری انجام شود، تخصیص برجسب درست نحوی به کلمات است. بنابراین اگر برجسب دهی نحوی^۱ قبل از مرحله رفع ابهام، صورت پذیرد و حاصل آن، پس از رفع ابهام از برجسب نحوی کلمات، در اختیار مرحله رفع ابهام معنایی قرار گیرد، رفع ابهام معنایی از هم‌نگاره‌ها با دقت بیشتری انجام خواهد گرفت.

در این مقاله، پس از مرور پیشینه پژوهش، ابتدا روش‌های ماشینی موجود در جهت رفع ابهام از معنای کلمات معرفی می‌شود و یکی از این روش‌ها معرفی و در پژوهش حاضر استفاده می‌شود و در نهایت رویکردی ماشینی جهت رفع ابهام از معنای کلمات با استفاده از نتیجه مرحله برجسب-زنی نحوی به کلمات و رفع ابهام از برجسب نحوی کلمات، معرفی می‌شود.

۲- پیشینه پژوهش

در این بخش به ذکر نمونه‌هایی از تحقیقاتی که در زمینه رفع ابهام معنایی از کلمات، با تاکید بر هم‌نگاره‌ها انجام شده است، پرداخته می‌شود.

- کاربرد رفع ابهام معنایی در حوزه‌های گوناگون پردازش زبان طبیعی از نظر ایده و ورونیس به صورت زیر است:

الف. ترجمه ماشینی^۲: رفع ابهام معنایی برای ترجمه صحیح کلمات، ضروری است؛ به عنوان مثال، کلمه فرانسوی «grille»، بسته به بافت، می‌تواند در انگلیسی به کلمات «railings»، «gate»، «bar»، «grid»، «scale»، «schedule» و.... ترجمه شود.

¹ part of speech (POS) tagging

² machine translation

ب. بازیابی اطلاعات^۱: زمانی که به دنبال کلیدواژه خاصی می‌گردید، بهتر است کلمه/کلماتی که در متن با معنای نادرست یا نامناسب به کار رفته‌اند، حذف گردند. به عنوان مثال، زمانی که در متون انگلیسی در جستجوی ارجاعات قضائی هستید، بهتر این است که مستندات^۲ی که حاوی کلمه «court» که بیشتر در ارتباط با کلمه «royalty» است تا کلمه «law»، حذف شوند.

پ. تجزیه و تحلیل محتوایی و موضوعی^۳: رویکرد رایج در تجزیه و تحلیل محتوایی و موضوعی، تجزیه و تحلیل توزیع طبقه‌بندی‌های از قبل تعریف شده کلمات (کلماتی که نشان‌دهنده مفهوم و معنای معین هستند) در یک متن است. در پردازش زبان طبیعی، نیاز برای رفع ابهام معنایی در چنین تجزیه و تحلیلی از مدت‌ها قبل شناخته شده است.

ت. تجزیه و تحلیل دستوری^۴: رفع ابهام معنایی برای برچسب‌دهی نحوی مفید است. به عنوان مثال، در زبان فرانسه در جمله «L'étagère plie sous les livres» (قفسه زیر سنگینی بار کتاب‌ها در حال خم شدن است)، رفع ابهام از معنی کلمه «livres» (که می‌تواند به معنی «کتاب‌ها» یا «واحد پولی به نام «پوند» باشد، که در معنای اول، در فرانسه کلمه‌ای است مذکر و در معنای دوم، مونث) ضروری است تا بتوان به درستی به آن برچسب نحوی «اسم، مذکر» را اختصاص داد. رفع ابهام معنایی برای تجزیه و تحلیل‌های خاص نحوی، از قبیل نحوه چسبیدن عبارت اضافی به عبارات دیگر، نیز ضروری است.

ث. پردازش گفتار^۵: آوانگاری درست کلمات در سنتز گفتار و تقطیع کلمه و تمایز گذاشتن میان هم‌آواها، مستلزم رفع ابهام معنایی از کلمات است.

ج. پردازش متن^۵: رفع ابهام معنایی برای تصحیح املاء کلمات (به عنوان مثال، تعیین اینکه چه زمانی باید از علائم زیروبری استفاده شود)، نیز ضروری است.

¹ information retrieval

² content and thematic analysis

³ grammatical analysis

⁴ speech processing

⁵ text processing

ایده و ورونیس، در این تحقیق به بررسی رویکردهای اصلی و مشهور در زمینه رفع ابهام معنایی می‌پردازند و مشکلات ناتمام و رویکردهای تحقیقات آینده را در نظر می‌گیرند (Ide & Veronis 1998).

- در زمینه کاربرد الگوریتم‌ها و تکنیک‌های یادگیری ماشینی برای رفع ابهام معنایی از کلمات، می‌توان به پژوهش انجام شده توسط باکس اشاره کرد؛ وی به کاربرد بین-پیکره‌ای سیستم‌های یادگیری ماشینی با سرپرست برای رفع ابهام معنایی کلمات به منظور آزمودن توانایی تعمیم دادن در پیکره‌های متنی می‌پردازد (Bakx 2006).
- از میان پژوهشگرانی که سعی دارند ضمن بررسی الگوریتم‌های موجود در رفع ابهام معنایی کلمه، روش‌هایی بر اساس بررسی مفهومی و ساختاری کلمات در جهت رفع ابهام کلمات در متون ارائه کنند، می‌توان به راسخ و همکاران (۲۰۱۴) اشاره کرد. آن‌ها راهکاری در جهت رفع مشکلات ترجمه ماشینی و بهبود کیفیت آن ارائه می‌دهند.
- با در نظر گرفتن خصوصیات نوشتاری زبان فارسی، مسعودی و همکاران (۱۳۸۹) روشی مبتنی بر پیکره جهت یافتن معنی صحیح کلمات پیشنهاد کرده‌اند. بر این اساس، ویژگی کلمه و نشانه‌هایی که بلافاصله قبل و بعد از کلمه مبهم به کار رفته‌اند، علاوه بر ویژگی سبدهای همراه هر معنی کلمه، به منظور رفع ابهام، مورد استفاده قرار گرفته است. نتایج دسته‌بندی با استفاده از یک روش بیزی (Naïve Bayes) تغییر یافته برای تعدادی از کلمات مبهم زبان فارسی که از پیکره پژوهشگرانه پردازش هوشمند علائم استخراج شده است، نشان می‌دهد استفاده از ویژگی‌های مذکور در مقایسه با روش‌هایی که تنها از سبدهای بهره می‌جویند، باعث بهبود دقت بازشناسی می‌شود.
- با توجه به مقدار اطلاعات موجود در فرهنگ‌های لغت و مسئله هم‌نگاره‌ها، ویلکس و استیونسون به توصیف دو تحقیق می‌پردازند: در یکی از آن‌ها به بررسی مقدار اطلاعات مربوط به ابهام‌زدایی معنای موجود در اجزا واژگانی کلام در یک فرهنگ لغت

((Machine Readable Dictionary (MRD)) می‌پردازند و در دیگری، که کاربردی‌تر است، سعی در ابهام‌زدایی معنایی همه کلمات قاموسی در یک متن که حاوی هم‌نگاره‌های موجود در فرهنگ لغت مذکور است، دارند. در این تحقیق سامانه برجسب‌دهی معنایی ساده معرفی می‌شود که صحت ۹۴٪ دارد (Wilks & Stevenson 1997).

- یکی از پژوهش‌های انجام شده در زمینه رفع ابهام از معنای کلمات در فارسی، پژوهشی است که توسط عابدینی و همکاران (۱۳۹۰) انجام شده است. آن‌ها روشی جدید ارائه می‌دهند که راهکاری است مبتنی بر آمار و در آن تنها مشخصه‌ای که مورد استفاده قرار گرفته است، کلمات مجاور کلمه مبهم می باشد. این راهکار در ابتدا یک‌سری اطلاعات آماری شامل مشخصه‌های مفید جهت عمل رفع ابهام را از طریق یک احتمال شرطی از پیکره استخراج می‌کند، سپس مشخصه‌های بدست آمده همراه با مقادیری که ارتباط میان این مشخصه‌ها و متون را در پیکره مشخص می‌کند، تشکیل یک پایگاه داده پویا را می‌دهند. در مرحله بعد، از طریق یک طبقه‌بندی، که در اینجا نزدیکترین همسایگی است، رفع ابهام کلمات انجام می‌شود. در نهایت کارایی سیستم مورد بحث قرار می‌گیرد. نتایج آزمایش‌ها تایید می‌کنند که راهکار ارائه شده نسبت به راهکارهای مشابه، دارای دقت مطلوبی می باشد.

یافته‌های عمده پژوهش‌ها را می‌توان به کاربرد رفع ابهام از معنای کلمات در حوزه‌های گوناگون پردازش زبان (مانند: ترجمه ماشینی، بازیابی اطلاعات، تجزیه و تحلیل محتوایی و موضوعی، تجزیه و تحلیل دستوری، پردازش گفتار و پردازش متن) خلاصه کرد و روش‌های پژوهشی مورد استفاده پژوهشگران شامل الگوریتم‌ها و تکنیک‌های یادگیری ماشینی، پژوهش‌های مبتنی بر آمار، بررسی مفهومی و ساختاری کلمات در جهت رفع ابهام کلمات در متون، روشی مبتنی بر پیکره جهت یافتن معنی صحیح کلمات با در نظر گرفتن خصوصیات نوشتاری زبان فارسی، اطلاعات موجود در فرهنگ‌های لغت و مسئله هم‌نگاره‌ها است.

۳- انواع یادگیری ماشینی جهت رفع ابهام از معنای کلمات

پال و همکاران معتقدند از نقطه نظر مقدار نظارت و دانش مورد نیاز جهت رفع ابهام از کلمات، دو روش متداول یادگیری ماشینی استفاده می‌شود که شامل: روش یادگیری بانظارت/ باسرپرست^۱ و روش یادگیری بدون نظارت/ بدون سرپرست^۲ است (Pal et al 2013). البته دو روش دیگر نیز وجود دارد که شامل روش یادگیری نیمه نظارتی^۳ و روش یادگیری مبتنی بر دانش^۴ است. از آنجائی که در تحقیق حاضر از یکی از دو روش اول استفاده می‌شود، به معرفی آن‌ها پرداخته می‌شود.

۳-۱- روش یادگیری ماشینی با نظارت/ با سرپرست

در این روش یادگیری ماشینی، از مجموعه یادگیری^۵ جهت پیش‌بینی معنای واقعی کلمه مبهم با استفاده از تعدادی جملات محدود که کلمات موجود در آن‌ها دارای معنای مشخص هستند و در واقع حاوی برجسب‌های معنایی هستند، استفاده می‌شود. سیستم، معنای واقعی کلمه مبهم در یک بافت خاص را بر اساس مجموعه مشخص و تعریف شده و از قبل وارد شده در سیستم، ارائه می‌دهد. در این روش مجموعه داده‌های یادگیری به صورت دستی تهیه می‌شود؛ در نتیجه، سیستم قادر به تولید قواعد ثابت برای تمام سیستم‌های یادگیری ماشینی باسرپرست جهت رفع ابهام از معنای کلمات مبهم، نخواهد بود. بنابراین، معنای واقعی کلمه‌ای مبهم (مانند هم‌نگاره‌ها)، در یک بافت همیشه قابل شناسایی نیست. یادگیری ماشینی بانظارت / باسرپرست در صورتی به نتیجه درست در زمینه رفع ابهام از کلمات مبهم خواهد رسید که مجموعه یادگیری حاوی اطلاعات

¹ supervised learning

² unsupervised learning

³ semi-supervised learning

⁴ knowledge-based approach

⁵ learning set

مجموعه یادگیری/ آموزشی، مجموعه‌ای است از مثال‌هاست که بر حسب تعدادی از مشخصه-

ها و برجسب معنایی آن‌ها رمزگزاری می‌شوند.

کافی برای همه مفاهیم ممکن یک کلمه مبهم نباشد (Pal et al 2013). نمونه‌هایی از الگوریتم یادگیری ماشینی با نظارت/ باسرپرست که به منظور رفع ابهام معنایی از کلمات به کار می‌روند شامل: یادگیری درخت تصمیم^۱، روش‌های Bayes ساده^۲، شبکه‌های عصبی^۳، روش‌های مبتنی بر مثال^۴ مانند: KNN و SVM است (Sarrafzade et al 2016).

۲-۳ روش یادگیری ماشینی بدون نظارت/ بدون سرپرست

در روش یادگیری ماشینی بدون نظارت/ بدون سرپرست، که معمولاً برای رفع ابهام از معنی کلمات استفاده می‌شود، از تمام تعاریف متناسب به کلمه مبهم که در فرهنگ لغت موجود است، استفاده می‌شود. این تعاریف مربوط به مدخل‌های گوناگون موجود در فرهنگ لغت است که حاوی انواع مختلف مقوله‌های نحوی مانند اسم، فعل، صفت و قید هستند (Pal et al 2013). در نظر گرفتن همه تعاریف و مفاهیم مربوط به کلمه مبهم همراه با برجسب‌های گوناگون متناسب به آن‌ها، مستلزم صرف زمان قابل توجهی است که ممکن است لزومی به صرف چنین زمان زیادی جهت رفع ابهام معنایی از کلمات مبهم نباشد. در صورتی که چنانچه کلمه مبهم حاوی برجسب نحوی خاص در یک بافت مشخص باشد، گویی قبل از رفع ابهام معنایی از کلمه، تخصیص برجسب نحوی به کلمه موجود در بافت صورت گرفته است و بنابراین، به عبارتی دیگر، تنها معانی مربوط به برجسب نحوی کلمه در نظر گرفته خواهد شد.

یکی از روش‌های یادگیری ماشینی بدون نظارت/ بدون سرپرست که جهت رفع ابهام معنایی از کلمات مبهم موجود در متون گوناگون، استفاده می‌شود، الگوریتم لسک^۵ است. الگوریتم لسک در سال ۱۹۸۶ توسط مایکل لسک معرفی شد. این الگوریتم از کلمات در عبارات کوتاه رفع ابهام می‌کند. به این صورت که، برای رفع ابهام معنایی از هر کلمه، تعاریف هر یک از مدخل‌های

¹ decision tree learning

² Naïve Bayes methods

³ neural networks

⁴ example-based methods

⁵ Lesk algorithm

موجود در فرهنگ لغت با تعاریف کلمات دیگر موجود در عبارت مقایسه می‌شود و معنایی به کلمه اختصاص داده می‌شود که بیشترین اشتراک را در تعداد کلمات مشترک موجود در توصیف کلمات موجود در عبارت داشته باشد. این الگوریتم در مواجهه با هر کلمه موجود در عبارت به همین صورت عمل می‌کند و از معنایی که طی عملکرد الگوریتم لسک به کلمه مبهم اختصاص داده شده است، استفاده نمی‌کند؛ به عبارتی دیگر در مواجهه با هر کلمه، فرآیند از نو آغاز می‌شود (Banerjee 2002). به عنوان مثال، عبارت pine cone را در انگلیسی در نظر بگیرید؛ با استفاده از فرهنگ لغت آکسفورد کلمه pine دارای دو معنی و کلمه cone دارای سه معنی است که به ترتیب شامل:

: pine

الف. نوعی درخت که همیشه سبز است و برگ‌های سوزنی شکل دارد (درخت کاج)

ب. از غم و حسرت و بیماری نحیف شدن.

: Cone

الف. مخروط

ب. هر شیء مخروطی شکل (جامد یا توخالی)

ج. میوه درخت کاج

الگوریتم لسک به این صورت عمل می‌کند که هر یک از دو تعریف کلمه pine با هر کدام از سه تعریف کلمه cone مقایسه می‌شود و مشخص می‌شود که عبارت «درخت همیشه سبز» به لحاظ معنایی بیشترین اشتراک را با دو کلمه pine و cone دارد. سپس، معنای به دست آمده، مناسب‌ترین معنا برای این عبارت در نظر گرفته می‌شود، البته زمانی که هر دو کلمه pine و cone با هم استفاده شوند.

نمای کلی عملکرد این الگوریتم در مورد مثال فوق در شکل ۱ آورده شده است.

Example: disambiguate PINE CONE

- PINE
 1. kinds of evergreen tree with needle-shaped leaves
 2. waste away through sorrow or illness
- CONE
 1. solid body which narrows to a point
 2. something of this shape whether solid or hollow
 3. fruit of certain evergreen trees

Pine#1 \cap Cone#1 = 0
 Pine#2 \cap Cone#1 = 0
 Pine#1 \cap Cone#2 = 1
 Pine#2 \cap Cone#2 = 0
 Pine#1 \cap Cone#3 = 2
 Pine#2 \cap Cone#3 = 0

شکل ۱: نمای کلی عملکرد الگوریتم لسک جهت رفع ابهام معنایی از عبارت pine cone

مثال دیگر رفع ابهام از معنی کلمه flies در دو عبارت time flies like an arrow و fruit flies like a banana است. الگوریتم به این صورت عمل می‌کند که در مواجهه با عبارت time flies like an arrow تمام تعاریف موجود در فرهنگ لغت و مربوط به کلمه time را با تمام تعاریف کلمات fly و arrow مقایسه می‌کند و معنایی را به time اختصاص می‌دهد که بیشترین هم‌پوشانی را با معانی کلمات دیگر موجود در عبارت دارد. سپس، همین فرآیند درباره کلمات دیگر عبارت تکرار می‌شود، یعنی پس از رفع ابهام از کلمه time، الگوریتم این بار تمام تعاریف موجود کلمه fly را با معانی کلمات time و arrow مقایسه می‌کند و تخصیص معنا به همان شکل که توضیح داده شد صورت می‌پذیرد (Banerjee 2002).

۴- معرفی رویکردی جهت رفع ابهام معنایی از کلمات در پیکره‌های متنی فارسی با استفاده از برچسب‌زنی نحوی و الگوریتم لسک

ویلکس و استیونسن معتقدند ابهام در بخش برچسب‌دهی نحوی، قسمتی از مشکلات مربوط به رفع ابهام معنایی از کلمات است و چنانچه تخصیص مقوله نحوی به کلمات به درستی انجام شود، گامی مهم در جهت رفع ابهام معنایی از کلمات برداشته خواهد شد. آن‌ها تحقیقات خود را محدود به مسئله هم‌نگاره‌ها کرده‌اند (Stevenson 1998 & Wilks). سیستم‌های گوناگونی جهت برچسب‌دهی نحوی به کلمات در زبان‌ها مختلف استفاده می‌شود که از میان آن‌ها می‌توان

به سه سیستم معروف اشاره کرد: ۱- مدل مارکوفی پنهان^۱: این مدل، یک نوع خاص از زنجیره مارکوف می‌باشد. یک مدل مارکوفی پنهان با پنج تایی (S, K, Π, A, B) تعریف می‌شود که در آن $S = \{S_1, \dots, S_N\}$ مجموعه حالات، $K = \{K_1, \dots, K_N\}$ مجموعه نشانه‌های خروجی، $\Pi = \{\pi_i, i \in S\}$ ماتریس احتمالات حالت اولیه، A ماتریس احتمالات انتقال و B ماتریس احتمالات خروجی است. یک مسیر در یک مدل مارکوفی پنهان، دنباله‌ای از انتقال‌ها پی‌درپی است، به این صورت که حالت نهایی یک انتقال، حالت شروع انتقال بعدی در مسیر می‌باشد. احتمال یک مسیر حاصل ضرب احتمالات انتقال می‌باشد. یک دنباله خروجی می‌تواند به وسیله چندین مسیر تولید شود، اما همیشه مسیری وجود دارد که محتمل‌ترین مسیر برای تولید این خروجی می‌باشد. در برچسب‌گذاری با مدل مارکوفی پنهان، دنباله برچسب‌ها در یک متن، به عنوان یک زنجیره مارکوف در نظر گرفته می‌شود. در این مدل فرض بر این است که برچسب یک کلمه تنها وابسته به برچسب کلمه قبل است و این وابستگی در طول زمان تغییر نمی‌کند. ۲- مدل مبتنی بر حافظه^۲: این نوع یادگیری در واقع شکلی از یادگیری قیاسی بانظارت/باسرپرست است. نمونه‌هایی که یادگیری از روی آن‌ها انجام می‌شود با برداری از ویژگی‌ها نمایش داده می‌شوند و هر نمونه یک برچسب خاص دارد که نشان‌دهنده طبقه آن نمونه می‌باشد. طی عمل آموزش، مجموعه‌ای از نمونه‌ها (مجموعه آموزش) به رده‌بنده داده می‌شود و به حافظه افزوده می‌گردد. برای رده‌بندی یک نمونه، فاصله آن نمونه با نمونه‌های موجود در حافظه محاسبه می‌شود و برچسب نمونه (نمونه‌ها) با حداقل فاصله برای پیش‌بینی برچسب نمونه جدید استفاده می‌شود. بنابراین در برچسب‌گذاری مبتنی بر حافظه، عمل برچسب‌گذاری در اصل به عمل رده‌بندی نمونه‌های جدید براساس نمونه‌های قبلی تبدیل می‌شود، که نمونه‌ها در واقع، کلمات هستند (محسنی: ۱۳۸۷). ۳- مدل مبتنی بر تغییر شکل^۳: این مدل در واقع نمونه‌ای از یادگیری مبتنی بر قواعد است که برای انواع طبقه‌بندی می‌تواند مورد استفاده قرار گیرد و قابل استفاده در حوزه‌های گوناگون پردازش زبان از جمله برچسب‌گذاری اجزا کلام و رفع ابهام از معنای کلمات است. این مدل قادر است عمل برچسب‌دهی واژگانی و برچسب‌دهی مبتنی بر بافت را انجام دهد.

¹ hidden markov model (HMM)

² memory-based model

³ transformation-based model

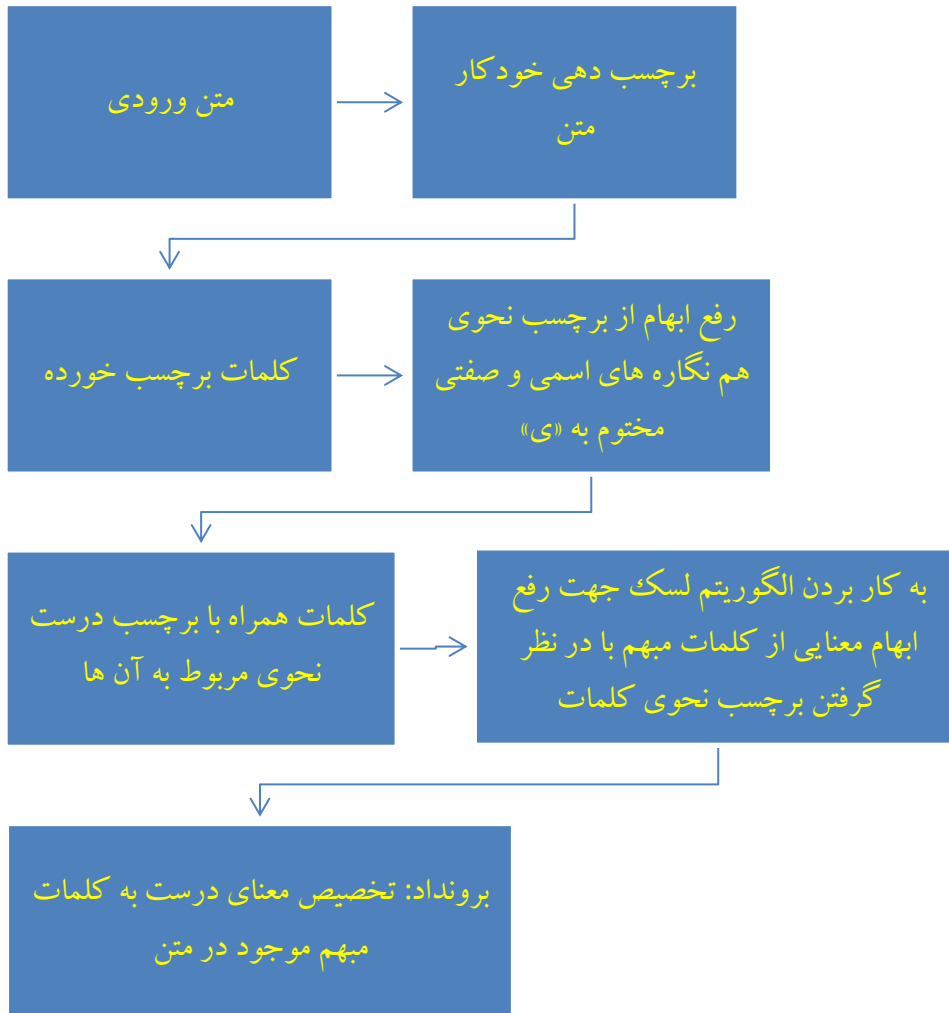
چنانچه کلمه مبهم (یک صورت نوشتاری واحد با بیش از یک معنی) حاوی برچسب نحوی خاص در یک بافت مشخص باشد، در واقع قبل از رفع ابهام معنایی از کلمه، تخصیص برچسب نحوی به کلمه موجود در بافت صورت گرفته است؛ به عبارتی دیگر، تنها معانی مربوط به برچسب نحوی کلمه در نظر گرفته خواهد شد. به عنوان مثال، اگر صورت نوشتاری «کرد» برچسب نحوی «فعل» را داشته باشد، تنها دو تلفظ ممکن /kard/ (در فارسی معیار) و /kerd/ (که در لهجه‌ها یا گویش‌های دیگر فارسی استفاده می‌شود) و نه /kord/ (که می‌تواند «اسم» باشد) و تنها یک معنی که معنی مرتبط با فعل است، برای این صورت نوشتاری اختصاص داده می‌شود. بنابراین، اعمال برچسب‌دهی نحوی، پیش از رفع ابهام از معنی کلمات مبهم، دو مزیت عمده دارد: اولین مزیت این است که عمل رفع ابهام سریع‌تر انجام خواهد شد و دومی مزیت این روش این است که از آنجائی که تنها معانی مرتبط با برچسب نحوی مربوطه لحاظ می‌شود، درجه صحت رفع ابهام بالاتر می‌رود.

مدل پیشنهادی تحقیق حاضر جهت رفع ابهام از معنای کلمات مبهم (هم‌نگاره‌ها) به این صورت عمل می‌کند: ابتدا متن ورودی به عنوان درونداد وارد مرحله برچسب‌گذاری خودکار نحوی می‌شود؛ در این مرحله تخصیص برچسب نحوی به کلمات موجود در متن صورت می‌پذیرد؛ انجام این مرحله خود مستلزم به کار بردن یکی از روش‌های ماشینی رایج جهت تخصیص برچسب نحوی به کلمات موجود در متن به صورت خودکار است. به دلیل پیچیدگی‌های موجود در ساخت‌واژه فارسی از میان هم‌نگاره‌های با فراوانی بالا در پیکره‌های متنی فارسی، فراوانی هم‌نگاره‌هایی که در اثر یکسان بودن نمود نوشتاری یا اسم‌ساز، یا نکره و یا صفت‌ساز ایجاد شده‌اند، بیشتر از هم‌نگاره‌های دیگر است، بنابراین لازم است پس از مرحله برچسب‌دهی خودکار متن، که خود مستلزم استفاده از اطلاعات ساخت‌واژی فارسی است، مرحله دیگری نیز در نظر گرفته شود که طی آن از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختموم به «ی» که فراوانی بالایی در پیکره‌های متنی فارسی دارند، رفع ابهام شود که خود مستلزم تعریف پنجره برای موتور جستجو و استخراج قواعدی است که بر اساس آن‌ها بتوان تعیین کرد برچسب چنین هم‌نگاره‌هایی با توجه به مقوله نحوی کلمات قبل و بعدشان و در واقع با توجه به بافت نحوی متن، اسم است یا صفت. به عنوان مثال، قاعده حساس به بافت نحوی زیر را در نظر بگیرید:

- حرف اضافه (P) + (کمیت نما / QUA) + اسم (N)

طبق این قاعده، چنانچه قبل از هم‌نگارهٔ مختوم به «ی»، حرف اضافه و به طور اختیاری، کمیت-
نما وجود داشته باشد، برچسب نحوی آن هم‌نگاره «اسم» خواهد بود. مانند: به (P) هیچ (QUA)
ارتشی (N) نیز اجازه نمی‌دهد....

پس از تخصیص برچسب نحوی به کلمات موجود در متن و رفع ابهام از هم‌نگاره‌های مختوم
به «ی»، کلمات که حالا حاوی برچسب درست نحوی هستند وارد مرحلهٔ رفع ابهام معنایی می-
شوند که در این مرحله از الگوریتم لسک استفاده می‌شود. از آنجائی که مقولهٔ نحوی کلمات
مبهم، قبل از مرحلهٔ رفع ابهام، مشخص می‌شود، الگوریتم لسک، تنها تعاریفی را در فرهنگ لغت
در نظر خواهد گرفت که با برچسب نحوی مربوط به کلمه هماهنگی داشته باشد. بنابراین، فرآیند
رفع ابهام با سرعت و صحت بالاتری پیش خواهد رفت. مدل پیشنهادی در شکل ۲ آورده شده
است.



شکل ۲: مدل رفع ابهام معنایی از کلمات با استفاده از اطلاعات مربوط به برچسب نحوی کلمات

۵- بحث و نتیجه گیری

در پژوهش حاضر رویکردی ماشینی معرفی شده است که سرعت رفع ابهام از معنای کلمات را بالاتر می‌برد؛ به این صورت که پیش از به کار بردن سیستمی جهت رفت ابهام معنایی از کلمات،

ابتدا عمل تخصیص برچسب نحوی به کلمات به صورت خودکار انجام می‌شود و سپس، از آنجائی که تعداد هم‌نگاره‌های مختوم به «ی» در متون فارسی قابل توجه است، رفع ابهام از برچسب نحوی این هم‌نگاره‌ها نیز انجام می‌شود و در نهایت نتیجه این دو مرحله وارد مرحله رفع ابهام معنایی کلمات می‌شود. بنابراین، به جای آنکه با به کار بردن الگوریتم لسک، به تنهایی، که ناگزیر از در نظر گرفتن همه تعاریف موجود در فرهنگ‌های لغت برای همه کلمات موجود در عبارت و بررسی اشتراکات معنایی هر کدام از کلمات موجود در عبارت با تعاریف مذکور باشد، در این مدل، تنها معانی مرتبط با برچسب نحوی جهت رفع ابهام معنایی برای کلمات در نظر گرفته می‌شود. ضمناً از آنجائی که تخصیص خودکار برچسب نحوی به کلمات موجود در متن، به دلیل وجود هم‌نگاره‌های گوناگون (مخصوصاً هم‌نگاره‌های اسمی و صفتی مختوم به «ی»)، با خطاهایی همراه است، در مدل مذکور این مسئله لحاظ شده است و پیش از رفع ابهام از معنای کلمات، رفع ابهام از برچسب نحوی کلمات صورت می‌پذیرد و حاصل آن در اختیار مرحله آخر که تخصیص معنای درست از میان تعاریف/ معانی گوناگون و موجود در فرهنگ‌های لغت، با توجه به بافت نحوی است، صورت می‌پذیرد.

۶- پیشنهادهایی برای پژوهش‌های بیشتر

تحقیق حاضر در واقع به معرفی رویکردی ماشینی با استفاده از الگوریتم لسک و برچسب‌دهی نحوی جهت رفع ابهام از معنای کلمات می‌پردازد، طی تحقیقات آتی می‌توان این رویکرد را عملیاتی نمود و از نتیجه آن جهت رفع ابهام از معنای کلمات در متون گوناگون فارسی، بهره برد. همچنین می‌توان ابتدا راه کارهایی جهت رفع ابهام از برچسب نحوی دیگر هم‌نگاره‌های موجود در متون فارسی ارائه کرد و نتیجه حاصل از آن را نیز در رویکرد ماشینی معرفی شده در این تحقیق به کار برد.

فهرست منابع

- 1- Banerjee, S. 2002. *Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet*. Department of computer science. University of Minnesota. U.S.A.
- 2- Gaustad, T. 2004. *Linguistic Knowledge and Word Sense Disambiguation*. Netherlands Organization For Scientific Research.
- 3- Ide, N. and Veronis, J. 1998. Introduction to special issue on word sense disambiguation: the state of the art. *Computational linguistics. Special issue on word sense disambiguation*. 24 (1) : 2-40. MIT press Cambridge, MA, USA.
- 4- Pal, A.,R., Munshi, A. , and Saha, D. 2013. An approach to speed-up the Word Sense Disambiguation procedure through sense filtering. *International journal of Instrumentation and Control systems (IJICS)*. Vol. 3. No. 4.
- 5- Sarrafzadeh, B. Yakovets, N. and Cerone, N. 2016. *Two novel approaches for Persian Word Sense Disambiguation. Computational Linguistics Project: Final Report*. University of Waterloo. Ca.
- 6- Wilks, Y and Stevenson, M. 1997. The grammar of sense: using part- of- speech tags as a first step in semantic disambiguation. *Natural language processing*. Cambridge university press.
- 7- Wilks, Y and Stevenson, M. 1998. Word sense disambiguation using optimized combinations of knowledge sources. *Proceedings of the 17th international conference on computational linguistics and the 36th annual meeting of the association for computational linguistics (COLING-ACL` 98)*. Montreal, Canada. Pp 1398-1402.
- 8- <http://www.merriam-webster.com>

۹- راسخ، ا.، صدرالدینی، م. و سیدمصطفی فخر احمد ۲۰۱۴. ابهام‌زدایی کلمات بر اساس ویژگی‌های لغوی و معنایی کلمات، با استفاده از طبقه‌بندی بیزین. *Journal of computing and security*. دانشکده برق و کامپیوتر، دانشگاه شیراز. ۱ (۲).

- ۱۰- سلطانی، پ. و کامران فانی ۱۳۶۲. سرعنوان‌های موضوعی فارسی. چاپ اول. ویرایش اول. کتابخانه ملی ایران.
- ۱۱- عابدینی، م.، جعفری، ش. و رضا جاویدان ۱۳۹۰. یک راهکار جدید مبتنی بر ناظر جهت رفع ابهام معنی کلمه با استفاده از خاصیت نزدیک ترین همسایگی. اولین همایش تخصصی سیستم‌های هوشمند کامپیوتری و کاربردهای آنها. وزارت علوم، تحقیقات و فناوری - دانشگاه پیام نور - دانشگاه پیام نور استان تهران
- ۱۲- علائی، ا. ۱۳۹۵. بررسی ساخت‌واژی هم‌نگاره‌های اسمی و صفتی به منظور برجسب دهی «اسم» به کلیدواژه‌ها در پیکره‌های علمی. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک). پژوهشکده مدیریت دانش.
- ۱۳- محسنی، مهدی ۱۳۸۷. سیستم برجسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی. دانشگاه علم و صنعت. دانشکده مهندسی کامپیوتر.
- ۱۴- مسعودی، ب.، قوچانی، س. و اعظم استاجی ۱۳۸۹. یک روش بیزی برای رفع ابهام معنایی کلمات در زبان فارسی با تأکید بر ویژگی‌های محلی کلمه. اولین کنفرانس ملی محاسبات نرم و فناوری اطلاعات. دانشگاه آزاد اسلامی، واحد ماهشهر.



معرفی رویکردی ماشینی با استفاده از الگوریتم لسک و برجسبدهی نحوی جهت رفع ابهام از معنای کلمات | علایی ابوذر

Introducing a machine-based approach for Word Sense Disambiguation: using Lesk algorithm and Part Of Speech tagging

Elham Alayiaboozar

PhD in General linguistics. Assistant Professor; Iranian Research Institute for Information Science and Technology(Irandoc)

Alayi@irandoc.ac.ir

abstract

Present study introduces a machine-based approach for Word Sense Disambiguation (WSD). In Persian, a morphologically complex language, lots of homographs are made; one way for doing WSD is allocating the right Part Of Speech (POS) tags to words, prior to WSD. Since the frequency of noun and adjective homographs in different Persian text corpuses is high, POS disambiguation of such homographs seems to be necessary for WSD. This paper introduces an approach in which first POS tagging is done, then the output, which is tagged sentences, enters the next step which is POS disambiguation of Persian nouns and adjective homographs; then the output of this step enters the final step which is applying the Lesk algorithm(a kind of unsupervised learning) for WSD. The proposed approach speeds up the WSD procedure by filtering the only relevant glosses (exist in dictionary) and increases the accuracy of the WSD procedure as well.

Key words: homographs, Word Sense Disambiguation, Part Of Speech tagging, disambiguation of Persian nouns and adjective homographs, Lesk algorithm