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

Iranian Journal of Information Processing and Management

Received: 04. Dec. 2016 **Accepted: 01, May 2017**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 33 | No. 3 | pp. 1165-1182
Spring 2018

Abstract: The present study introduces a machine-based approach for word sense disambiguation (WSD). In Persian, a morphologically complex language, POS tag which 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 POS tag text corpuses is high, POS tag 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 (existing in dictionary) and increases the accuracy of the WSD procedure as well.

Keywords: Homographs, Word Sense Disambiguation, Part of Speech Tagging, Disambiguation of Persian Nouns and Adjective Homographs, Lesk Algorithm

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)؛
alayi@irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۵/۰۹/۱۴ | **پذیرش:** ۱۳۹۶/۰۲/۱۱ | مقاله برای اصلاح به مدت ۲ روز نزد پدیدآوران بوده است.

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۳ | شمارهٔ ۳ | ص‌ص ۱۱۶۵-۱۱۸۲
بهار ۱۳۹۷

چکیده: پژوهش حاضر به معرفی رویکردی ماشینی برای چگونگی رفع ابهام معنایی از کلمات می‌پردازد. در زبان فارسی، که ساخت‌واژهٔ پیچیده‌ای 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 مرحلهٔ رفع ابهام معنایی از کلمات صورت پذیرد، رفع ابهام معنایی از هم‌نگاره‌ها با دقت بیشتری انجام خواهد گرفت. از آنجا که فراوانی هم‌نگاره‌های اسمی و صفتی در متون فارسی، در مقایسه با سایر هم‌نگاره‌ها بالاست، 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 بدون نظارت/ 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 نتیجهٔ مرحلهٔ 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 که 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.

کلیدواژه‌ها: رفع ابهام از معنای کلمات، هم‌نگاره، برچسب‌دهی نحوی، الگوریتم لسک، رفع ابهام از برچسب نحوی کلمات

۱. مقدمه

هنگام خواندن متون گوناگون فارسی، خوانندگان با چالش‌های مختلفی مانند انواع مترادف‌ها، متجانس‌ها، کلمات خارجی یا بومی و نظایر آن، مواجه می‌شوند. به‌عنوان مثال، «ماه‌گرفتگی» و «خسوف» مترادف هستند و یا «تهران» و «طهران»، دو گونهٔ 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 ذکرشده، 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»[1] در تمام سطوح زبانی مواجه است. «گوستاد» معتقد است ambiguity با vagueness متفاوت است (البته، معادل فارسی هر دو واژهٔ انگلیسی، «ابهام» است). وی vagueness را حالتی در نظر می‌گیرد که یک کلمه یا عبارت، تنها یک معنا دارد که مرزهای آن به‌طور دقیق مشخص نشده است، در صورتی که ambiguity، حالتی است که یک کلمه می‌تواند چندین معنا در بافت‌های گوناگون داشته باشد (Gaustad 2004). 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 (همان):

الف. ابهام ساخت‌واژی[2]: یک کلمه با توجه به ساخت درونی خود می‌تواند مبهم باشد. کلمات مرکب، منبع معمولِ ابهام ساخت‌واژی هستند. به‌عنوان مثال، کلمهٔ مرکب computertaalkunde در زبان هلندی را به دو طریق می‌توان تجزیه کرد: computer-

1. ambiguity
2. morphological ambiguity

taalkunde به معنای زبان‌شناسی رایانه‌ای و computertaal-kunde به معنای برنامه‌ریزی دانش زبانی.

ب. ابهام نحوی[1]: مثال بارز چنین ابهاماتی هنگامی رخ می‌دهد که یک گروه حرف اضافه‌ای به عناصر جمله اضافه می‌شود، مانند جملهٔ:

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[2]: این نوع ابهام در بسیاری از زبان‌ها عمدتاً در تفسیر و تعیین مرجع ضمایر اتفاق می‌افتد. به‌عنوان مثال، جملهٔ زیر را در انگلیسی در نظر بگیرید:

Mary's mother is a gardener. John likes her.

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

ت. ابهام معنایی واژگانی[3]: این نوع ابهام زمانی اتفاق می‌افتد که یک کلمه (در واقع، یک صورت نوشتاری واحد) بیش از یک معنا 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 همهٔ زبان‌های دنیا، کلمات بسیاری وجود دارد که بسته به بافت، معانی گوناگونی دارند. به‌عنوان 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 معنای متفاوت در دو بافت متفاوت دارد.

در مقالهٔ حاضر به بررسی چگونگی رفع ابهام معنایی واژگانی و ساخت‌واژی کلمات پرداخته می‌شود. رفع ابهام از معنای کلمه[4]، فرایندی برای تعیین معنای واقعی کلمهٔ مبهم بر اساس موقعیت مشخص (بافت) است. به‌عنوان مثال، کلمهٔ «شیر» معانی مختلفی دارد،

1. syntactic ambiguity 2. pragmatic ambiguity 3. lexical semantic ambiguity
4. word sense disambiguation (WSD)

از جمله: «حیوانی درنده و گوشتخوار در جنگل»، «نوعی از لبنیات» و «وسیله‌ای که آب از آن جاری می‌شود». چنین کلماتی ذاتاً مبهم هستند. فرایندی که طی آن معنای مناسب یک کلمهٔ مبهم در یک بافت خاص، مشخص می‌شود، رفع ابهام از معنای کلمه خوانده می‌شود. منظور از کلمات مبهم، کلماتی هستند که صورت نوشتاری یکسان، اما منشأ، معنا یا تلفظ متفاوت دارند و اصطلاحاً هم‌نگاره/ هم‌نگاشت[1] نامیده می‌شوند (MerriamWebster Dictionary). در زبان‌هایی که مانند زبان فارسی ساخت‌واژهٔ پیچیده‌ای دارند، هم‌نگاره‌های بسیاری ساخته می‌شوند. هم‌نگارگی در زبان فارسی ناشی از بازنمایی واجی 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. «محسنی» از «بی‌جن‌خان و مرادزاده» نقل می‌کند که آن‌ها هم‌نگاره‌های فارسی را به چند دسته تقسیم می‌کنند که از میان آن‌ها می‌توان به موارد «الف» تا «ت» اشاره کرد (محسنی ۱۳۸۷ نقل از بی‌جن‌خان و مرادزاده ۱۳۸۳):

الف. واژگانی: هم‌نگاره‌هایی هستند که در فرهنگ‌های لغت به‌صورت 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 به‌وجود می‌آیند، مانند: یکسان بودن نمود نوشتاری تک‌واژ یای نکره، یای اسم‌ساز (اسم مکان، اسمی که دال بر شغل یا محافظت و دارندگی است، اسم معنا یا اشیا، تصغیر و تحبیب، اسم مصدر یا حاصل مصدر)، شناسهٔ دوم شخص مفرد و یای صفت‌ساز (صفت فاعلی و مفعولی، صفتی که دال بر نسبت است) که همه نمود نوشتاری «-ی» را دارند. به‌عنوان مثال، کلمهٔ «آسمانی» در اثر اضافه‌شدن پسوند تصریفی نکره‌کننده (یای نکره) یا پسوند اشتقاقی صفت‌ساز به‌وجود آمده است و فارغ از بافت،

1. homograph

می‌توانـد اسـم نکـره یـا صفـت باشـد. از میـان هم‌نگاره‌هـای ذکرشـده، هم‌نگاره‌هایـی کـه در اثـر اضافه‌شـدن پسـوند «-ی» اسم‌سـاز یـا صفت‌سـاز ایجـاد می‌شـوند، فراوانـی بالایـی در پیکره‌هـای متنـی فارسـی دارنـد (علایـی ۱۳۹۵).

سیسـتم رفع ابهـام از معنـای کلمـات بـه ایـن صـورت عمـل می‌کنـد کـه ابتـدا مجموعـه‌ای از کلمـات، مثـلاً یـک جملـه، در اختیـار سیسـتم قـرار می‌گیـرد و سیسـتم بـا اسـتفاده از منابـع اطلاعاتـی کـه در دسـترس آن قـرار دارد، مناسـب‌ترین معنـا بـرای یـک کلمـه را در بافت مشـخص می‌کنـد. منبـع اطلاعـات می‌توانـد مجموعـه‌ای از متـون باشـد (پیکره‌هـای متنـی) کـه کلمـات یـا برچسـب معنایـی و نحـوی دارنـد یـا ندارنـد. همچنیـن، منبـع اطلاعـات می‌توانـد فرهنگ‌هـای لغـات قابـل خوانـدن بـرای ماشـین باشـد (Sarrafzade, Yakovets, and Cerone 2016). بـدون مجهزبـودن بـه چنیـن اطلاعاتـی، تشـخیص ماشـینی معنـای هم‌نگاره‌هـا غیرممکـن خواهـد بـود. در واقـع، گویشـوران به‌دلیـل مجهزبـودن بـه اطلاعـات زبانـی از قبیـل اطلاعـات مربـوط بـه سـاخت‌واژهٔ زبـان، سـاخت واجـی و سـاخت نحـوی زبـان، قـادر بـه رفـع ابهـام از هم‌نگاره‌هـا در بافـت نحـوی هسـتند، امـا سـامانه‌های پـردازش متـون و رفـع ابهـام از معنـای کلمـات، چنانچـه بـه چنیـن اطلاعـات زبانـی مجهـز نشـده باشـند، بـا مشـکلاتی در پـردازش هم‌نگاره‌هـا مواجـه می‌شـوند. یکـی از راه‌هایـی کـه کمـک می‌کنـد رفـع ابهـام معنایـی کلمـات (هم‌نگاره‌هـا) بـا سـهولت و دقـت بیشـتری انجـام شـود، تخصیـص برچسـب درسـت نحـوی بـه کلمـات اسـت. بنابرایـن، اگـر برچسـب‌دهی نحـوی[1] قبـل از مرحلـهٔ رفـع ابهـام صـورت پذیـرد و حاصـل آن، پـس از رفـع ابهـام از برچسـب نحـوی کلمـات، در اختیـار مرحلـهٔ رفـع ابهـام معنایـی قـرار گیـرد، رفـع ابهـام معنایـی از هم‌نگاره‌هـا بـا دقـت بیشـتری انجـام خواهـد گرفـت.

در ایـن مقالـه، پـس از مـرور پیشـینهٔ پژوهـش، ابتـدا روش‌هـای ماشـینی موجـود در جهـت رفـع ابهـام از معنـای کلمـات معرفـی می‌شـود و یکـی از ایـن روش‌هـا معرفـی و در پژوهـش حاضـر اسـتفاده می‌شـود و در نهایـت، رویکـردی ماشـینی جهـت رفـع ابهـام از معانـی کلمـات بـا اسـتفاده از نتیجـهٔ مرحلـهٔ برچسـب‌زنی نحـوی بـه کلمـات و رفـع ابهـام از برچسـب نحـوی کلمـات معرفـی می‌شـود.

1. part of speech (POS) tagging

۲. پیشینهٔ پژوهش

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

◇ کاربرد رفع ابهام معنایی در حوزه‌های گوناگون پردازش زبان طبیعی از نظر Ide & Veronis (1998) به‌صورت زیر است:

الف. ترجمهٔ ماشینی[1]: رفع ابهام معنایی برای ترجمهٔ صحیح کلمات ضروری است. به‌عنوان مثال، کلمهٔ فرانسوی "grille"، بسته به بافت، می‌تواند در انگلیسی به کلمات "railings"، "gate"، "bar"، "grid"، "scale"، "schedule" و ... ترجمه شود.

ب. بازیابی اطلاعات[2]: زمانی که به‌دنبال کلیدواژهٔ خاصی می‌گردید، بهتر است کلمه یا کلماتی که در متن با معنای نادرست یا نامناسب به‌کار رفته‌اند، حذف گردند. به‌عنوان مثال، زمانی که در متون انگلیسی در جست‌وجوی ارجاعات قضائی هستید، بهتر این است مستنداتی که حاوی کلمهٔ "court" که بیشتر در ارتباط با کلمهٔ "royalty" است تا کلمهٔ "law"، حذف شوند.

پ. تجزیه و تحلیل محتوایی و موضوعی[3]: رویکرد رایج در تجزیه و تحلیل محتوایی و موضوعی، تجزیه و تحلیل توزیع طبقه‌بندی‌های از قبل تعریف‌شدهٔ کلمات (کلماتی که نشان‌دهندهٔ 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[4]: رفع ابهام معنایی برای برچسب‌دهی نحوی مفید است. به‌عنوان مثال، در زبان فرانسه در جملهٔ L` étagère plie sous les livres (قفسهٔ زیر سنگینی بار کتاب‌ها در حال خم‌شدن است)، رفع ابهام از معنای کلمهٔ "livres" (که می‌تواند به معنای «کتاب‌ها» یا «واحد پولی» به نام «پوند» باشد، که در معنای اول، در فرانسه کلمه‌ای است مذکر و در معنای دوم، مؤنث) ضروری است تا بتوان به‌درستی به آن برچسب نحوی «اسم، مذکر» را اختصاص داد. رفع ابهام

1. machine translation
2. information retrieval
3. content and thematic analysis
4. grammatical analysis

معنایی برای تجزیه و تحلیل‌های خاص نحوی، از قبیل نحوهٔ چسبیدن عبارت اضافی به‌عبارات دیگر نیز ضروری است.

ث. پردازش گفتار[1]: آوانگاری درست کلمات در سنتز گفتار و تقطیع کلمه و تمایزگذاشتن میان هم‌آواها، مستلزم رفع ابهام معنایی از کلمات است.

ج. پردازش متن[2]: رفع ابهام معنایی برای تصحیح املاء کلمات (به‌عنوان مثال، تعیین این‌که چه زمانی باید از علائم زیروزبری استفاده شود)، نیز ضروری است.

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

- ◇ در زمینهٔ کاربرد الگوریتم‌ها و تکنیک‌های یادگیری ماشینی برای رفع ابهام معنایی از کلمات می‌توان به پژوهش انجام‌شده توسط «باکس» اشاره کرد. وی به کاربرد بین پیکره‌ای سیستم‌های یادگیری ماشینی باسرپرست برای رفع ابهام معنایی کلمات به‌منظور آزمودن توانایی تعمیم‌دادن 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 برای رفع ابهام کلمات در متون ارائه کنند، می‌توان به Rasekh Sadreddini and Fakhrahmad (2014) اشاره کرد. آن‌ها راهکاری در جهت رفع مشکلات ترجمهٔ ماشینی و بهبود کیفیت آن ارائه می‌دهند.
- ◇ با در نظر گرفتن خصوصیات نوشتاری زبان فارسی، «مسعودی، قوچانی و استاجی» روشی مبتنی بر پیکره جهت یافتن معنای صحیح کلمات پیشنهاد کرده‌اند. بر این اساس، ویژگی کلمه و نشانه‌هایی که بلافاصله قبل و بعد از کلمهٔ مبهم به‌کار رفته‌اند، علاوه بر ویژگی سبد لغات همراه هر معنای کلمه، برای رفع ابهام مورد استفاده قرار گرفته است (۱۳۸۹). نتایج دسته‌بندی با استفاده از یک روش «بیزی»[3] تغییریافته برای تعدادی از کلمات مبهم زبان فارسی که از متون موجود در پیکرهٔ پژوهشکدهٔ پردازش هوشمند علائم استخراج شده، نشان می‌دهد که استفاده از ویژگی‌های

1. speech processing 2. text processing 3. Naïve Bayes

مذکور در مقایسه با روش‌هایی که تنها از سبد لغات بهره می‌جویند، باعث بهبود دقت بازشناسی می‌شود.

- با توجه به مقدار اطلاعات موجود در فرهنگ‌های لغت و مسئلهٔ هم‌نگاره‌ها، «ویلکس و استیونسون» به توصیف دو تحقیق می‌پردازند: در یکی از آن‌ها به بررسی مقدار اطلاعات مربوط به ابهام‌زدایی معنای موجود در اجزای واژگانی کلام در فرهنگ لغت Machine Readable Dictionary (MRD) می‌پردازند و در دیگری، که کاربردی‌تر است، سعی در ابهام‌زدایی معنایی همهٔ کلمات قاموسی در یک متن که حاوی هم‌نگاره‌های موجود در فرهنگ لغت مذکور است، دارند. در این تحقیق سامانهٔ برچسب‌دهی معنایی ساده معرفی می‌شود که صحت ۹۴ درصد دارد (Wilks & Stevenson 1997).
- یکی از پژوهش‌های انجام‌شده در زمینهٔ رفع ابهام از معنای کلمات در فارسی، پژوهشی است که توسط «عابدینی، جعفری، و جاویدان» انجام شده است. آن‌ها روشی جدید ارائه می‌دهند که راهکاری است مبتنی بر آمار و درآن تنها مشخصه‌ای که مورد استفاده قرارگرفته، کلمات مجاور کلمهٔ مبهم است. 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 به‌دست‌آمده همراه با مقادیری که ارتباط میان این مشخصه‌ها و متون را در پیکره مشخص می‌کند، یک پایگاه دادهٔ پویا تشکیل می‌دهند. در مرحلهٔ بعد، از طریق یک طبقه‌بندی، که دراینجا نزدیک‌ترین همسایگی است، رفع ابهام کلمات انجام می‌شود. در نهایت، کارایی سیستم مورد بحث قرار می‌گیرد. نتایج آزمایش‌ها تأیید می‌کنند که راهکار ارئه‌شده نسبت به راهکارهای مشابه دارای دقت مطلوبی است (۱۳۹۰).

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

۳. انواع یادگیری ماشینی جهت رفع ابهام از معنای کلمات

«پال، مونشی و ساها» معتقدند که از نقطه‌نظر مقدار نظارت و دانش مورد نیاز جهت رفع ابهام از کلمات، دو روش متداول یادگیری ماشینی به‌کار می‌رود که شامل: روش یادگیری بانظارت/ باسرپرست[1] و روش یادگیری بدون نظارت/ بدون سرپرست[2] است (Pal, Munshi, and Saha 2013). البته، دو روش دیگر نیز وجود دارد که شامل روش یادگیری نیمه‌نظارتی[3] و روش یادگیری مبتنی بر دانش[4] است. از آنجا که در تحقیق حاضر از یکی از دو روش اول استفاده می‌شود، به معرفی آن‌ها پرداخته می‌شود.

۳-۱. روش یادگیری ماشینی بانظارت/ باسرپرست

در این روش یادگیری ماشینی، از مجموعهٔ یادگیری[5] جهت پیش‌بینی معنای واقعی کلمهٔ مبهم با استفاده از تعدادی جملات محدود که کلمات موجود در آن‌ها دارای معنای مشخص بوده و در واقع، حاوی برچسب‌های معنایی هستند، استفاده می‌شود. سیستم، معنای واقعی کلمهٔ مبهم در یک بافت خاص را بر اساس مجموعهٔ مشخص و تعریف‌شده که از قبل وارد سیستم شده، ارائه می‌دهد. در این روش مجموعهٔ داده‌های یادگیری به‌صورت 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 نتیجهٔ درست در زمینهٔ رفع ابهام از کلمات مبهم خواهد رسید که مجموعهٔ یادگیری حاوی اطلاعات کافی برای همهٔ مفاهیم ممکن یک کلمهٔ مبهم نباشد (Pal, Munshi, and Saha 2013). نمونه‌هایی از الگوریتم یادگیری ماشینی با نظارت/ باسرپرست که به‌منظور رفع ابهام معنایی از کلمات به‌کار می‌روند، شامل یادگیری درخت تصمیم[6]، روش‌های «بیزی» ساده، شبکه‌های عصبی[7]، روش‌های مبتنی بر مثال[8]، مانند KNN و SVM است (Sarrafzadeh, Yakovets, and Cerone 2016).

---

1. supervised learning
2. unsupervised learning
3. semi-supervised learning
4. knowledge-based approach
5. learning set

مجموعهٔ یادگیری/ آموزشی، مجموعه‌ای است از مثال‌ها که بر حسب تعدادی از مشخصه‌ها و برچسب معنایی آن‌ها رمزگزاری می‌شوند.

6. decision tree learning
7. neural networks
8. example-based methods

۳-۲. روش یادگیری ماشینی بدون نظارت/ بدون سرپرست

در روش یادگیری ماشینی بدون نظارت/ بدون سرپرست، که معمولاً برای رفع ابهام از معنای کلمات استفاده می‌شود، از تمام تعاریف منتسب به کلمهٔ 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, Munshi, and Saha 2013). در نظر گرفتن همهٔ تعاریف و مفاهیم مربوط به کلمهٔ مبهم همراه با برچسب‌های گوناگون منتسب 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 اگر کلمهٔ مبهم حاوی برچسب نحوی خاص در یک بافت مشخص باشد، گویی قبل از رفع ابهام معنایی از کلمه، تخصیص برچسب نحوی به کلمهٔ موجود در بافت صورت گرفته است و بنابراین، به‌عبارتی دیگر، تنها معانی مربوط به برچسب 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» است. الگوریتم «لسک» در سال ۱۹۸۶ توسط «مایکل لسک»[1] 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 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 کلمهٔ موجود در عبارت به همین صورت عمل می‌کند و از معنایی که طی عملکرد«الگوریتم «لسک» به کلمهٔ مبهم اختصاص داده شده است، استفاده نمی‌کند. به‌عبارتی دیگر، در مواجهه با هر کلمه، فرایند از نو آغاز می‌شود (Banerjee 2002). به‌عنوان مثال، عبارت pine cone را در انگلیسی در نظر بگیرید. با استفاده از فرهنگ لغت «آکسفورد» کلمهٔ pine دارای دو معنا و کلمهٔ cone دارای سه معناست؛ به‌ترتیب زیر:

1. Michael Lesk

:Pine

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

ب. از غم و حسرت و بیماری نحیف‌شدن.

:Cone

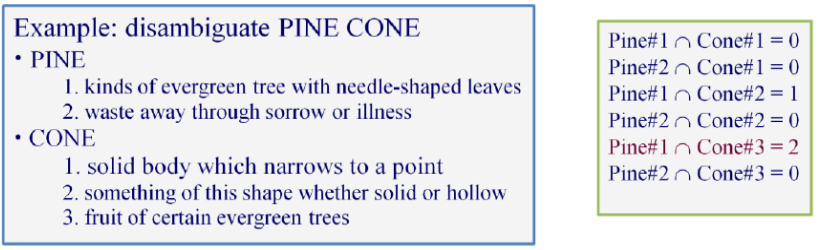
الف. مخروط

ب. هر شیء مخروطی شکل (جامد یا توخالی)

ج. میوهٔ درخت کاج

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

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



شکل ۱. نمای کلی عملکرد الگوریتم «لسک» جهت رفع ابهام معنایی از عبارت pine cone

مثال دیگر، رفع ابهام از معنای کلمهٔ 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 کلمهٔ 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 فرایند دربارهٔ کلمات دیگر عبارت تکرار می‌شود؛ یعنی پس از رفع ابهام از کلمهٔ time، الگوریتم

این بار تمام تعاریف موجود کلمهٔ fly را با معانی کلمات time و arrow مقایسه می‌کند و تخصیص معنا به همان شکل که توضیح داده شد، صورت می‌پذیرد (Banerjee 2002).

٤. 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 مقولهٔ نحوی به کلمات به‌درستی انجام گیرد، گامی مهم در جهت رفع ابهام معنایی از کلمات برداشته خواهد شد. آن‌ها تحقیقات خود را محدود به مسئلهٔ هم‌نگاره‌ها کرده‌اند (Wilks & Stevenson 1998). سیستم‌های گوناگونی جهت برچسب‌دهی نحوی به کلمات در زبان‌های مختلف استفاده می‌شود که از میان آن‌ها می‌توان به سه سیستم معروف اشاره کرد:

۱. «مدل مارکوفی پنهان»[1]: این مدل، یک نوع خاص از زنجیرهٔ «مارکوف» است. یک مدل مارکوفی پنهان با پنج‌تایی (S, K, Π, A, B) تعریف می‌شود که در آن S= {S1,……., SN} مجموعهٔ حالات، K= {K1,……., KN} مجموعهٔ 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 انتقالات پی در پی است؛ به این صورت که حالت نهایی یک انتقال، حالت شروع انتقال بعدی در مسیر است. احتمال یک مسیر حاصل ضرب احتمالات انتقالات است. یک دنبالهٔ خروجی می‌تواند به‌وسیلهٔ چندین مسیر تولید شود، اما همیشه مسیری وجود دارد که محتمل‌ترین مسیر برای تولید این خروجی است. در برچسب‌گذاری با «مدل مارکوفی پنهان»، دنبالهٔ برچسب‌ها در یک متن، به‌عنوان یک زنجیرهٔ «مارکوف» در نظر گرفته می‌شود. در این مدل فرض بر این است که برچسب یک کلمه تنها وابسته به برچسب کلمهٔ قبل است و این وابستگی در طول زمان تغییر نمی‌کند.

۲. مدل مبتنی بر حافظه[2]: 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 نشان‌دهندهٔ

1. Hdden Markov Model (HMM)
2. memory-based model

طبقهٔ آن نمونه است. طی عمل آموزش، مجموعه‌ای از نمونه‌ها (مجموعهٔ آموزش) به رده‌بنده داده می‌شود و به حافظه افزوده می‌گردد. برای رده‌بندی یک نمونه، فاصلهٔ آن نمونه با نمونه‌های موجود در حافظه محاسبه می‌شود و برچسب نمونه (نمونه‌ها) با حداقل فاصله برای پیش‌بینی برچسب نمونهٔ جدید استفاده می‌شود. بنابراین، در برچسب‌گذاری مبتنی بر حافظه، عمل برچسب‌گذاری در اصل به عمل رده‌بندی نمونه‌های جدید بر اساس نمونه‌های قبلی تبدیل می‌شود که نمونه‌ها در واقع، کلمات هستند (محسنی ۱۳۸۷).

۳. مدل مبتنی بر تغییر شکل[1]: 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 اجزای کلام و رفع ابهام از معنای کلمات است. این مدل قادر است عمل برچسب‌دهی واژگانی و برچسب‌دهی مبتنی بر بافت را انجام دهد.

چنانچه کلمهٔ مبهم (یک صورت نوشتاری واحد با بیش از یک معنا) حاوی برچسب نحوی خاص در یک بافت مشخص باشد، در واقع، قبل از رفع ابهام معنایی از کلمه، تخصیص برچسب نحوی به کلمهٔ موجود در بافت صورت گرفته است. به‌عبارتی دیگر، تنها معانی مربوط به برچسب نحوی کلمه در نظر گرفته خواهد شد. به‌عنوان 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 معنا که معنای مرتبط با فعل است، برای این صورت نوشتاری اختصاص داده می‌شود. بنابراین، اعمال برچسب‌دهی نحوی، پیش از رفع ابهام از معنای کلمات مبهم دو مزیت عمده دارد: اولین مزیت این است که عمل رفع ابهام سریع‌تر انجام خواهد شد و مزیت دوم این روش آن است که چون تنها معانی مرتبط با برچسب نحوی مربوطه لحاظ می‌شود، درجهٔ 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 به‌عنوان درون‌داد وارد مرحلهٔ برچسب‌گذاری خودکار نحوی می‌شود. در این مرحله تخصیص برچسب نحوی به کلمات موجود در متن

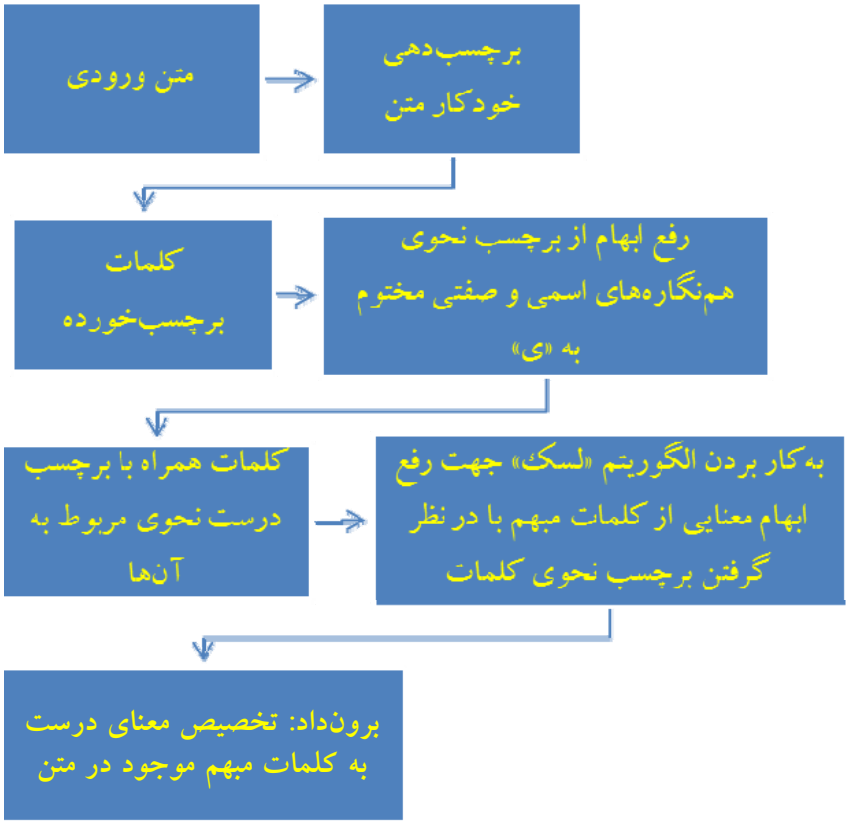
1. transformation-based model

صورت می‌پذیرد. انجام این مرحله خود مستلزم به‌کار بردن یکی از روش‌های ماشینی رایج جهت تخصیص برچسب نحوی به کلمات موجود در متن به‌صورت خودکار است. به‌دلیل پیچیدگی‌های موجود در ساخت‌واژهٔ فارسی از میان هم‌نگاره‌های با فراوانی بالا در پیکره‌های متنی فارسی، فراوانی هم‌نگاره‌هایی که در اثر یکسان‌بودن نمودِ نوشتاری یای اسم‌ساز، یای نکره و یای صفت‌ساز ایجاد شده‌اند، بیشتر از هم‌نگاره‌های دیگر است. بنابراین، لازم است پس از مرحلهٔ برچسب‌دهی خودکار متن، که خود مستلزم استفاده از اطلاعات ساخت‌واژی فارسی است، مرحلهٔ دیگری نیز در نظر گرفته شود که طی آن از برچسب نحوی هم‌نگاره‌های اسمی و صفتی مختوم به «ی» که فراوانی بالایی در پیکره‌های متنی فارسی دارند، رفع ابهام شود که خود مستلزم تعریف پنجره برای موتور جست‌وجو و استخراج قواعدی است که بر اساس آن‌ها بتوان تعیین کرد برچسب چنین هم‌نگاره‌هایی با توجه به مقولهٔ نحوی کلمات قبل و بعدشان و در واقع، با توجه به بافت نحوی متن، اسم است یا صفت. به‌عنوان مثال، قاعدهٔ حساس به بافت نحوی زیر را در نظر بگیرید:

◇ حرف اضافه (P) + (کمیت‌نما/ QUA) + /اسم (N)

طبق این قاعده، چنانچه قبل از هم‌نگارهٔ مختوم به «ی»، حرف اضافه و به‌طور اختیاری، کمیت‌نما 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 می‌شوند که در این مرحله از الگوریتم «لسک» استفاده می‌شود. از آنجا 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، فرایند 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 نتیجه‌گیری

در پژوهش حاضر رویکردی ماشینی معرفی شده است که سرعت رفع ابهام از معنای کلمات را بالاتر می‌برد. به این صورت که پیش از به‌کار بردن سیستمی جهت رفت ابهام معنایی از کلمات، ابتدا عمل تخصیص برچسب نحوی به کلمات به‌صورت خودکار انجام می‌شود و سپس، از آنجا 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، نتیجهٔ این دو مرحله وارد مرحلهٔ رفع ابهام معنایی کلمات می‌شود. بنابراین، به‌جای آن‌که با به‌کار بردن الگوریتم «لسک» به‌تنهایی، که ناگزیر از در نظر گرفتن همهٔ تعاریف موجود در فرهنگ‌های لغت برای همهٔ کلمات موجود در عبارت و بررسی اشتراکات معنایی

هر کدام از کلمات موجود در عبارت با تعاریف مذکور است، در این مدل، تنها معانی مرتبط با برچسب نحوی جهت رفع ابهام معنایی برای کلمات در نظر گرفته می‌شود. در ضمن، از آنجا که تخصیص خودکار برچسب نحوی به کلمات موجود در متن، به‌دلیل 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 مرحلهٔ آخر که تخصیص معنای درست از میان تعاریف و یا معانی گوناگون و موجود در فرهنگ‌های لغت، با توجه به بافت نحوی است، صورت می‌پذیرد.

٦. 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 نتیجهٔ آن جهت رفع ابهام از معنای کلمات در متون گوناگون فارسی بهره برد. همچنین، می‌توان ابتدا راهکارهایی جهت رفع ابهام از برچسب نحوی دیگر هم‌نگاره‌های موجود در متون فارسی ارائه کرد و نتیجهٔ حاصل از آن را نیز در رویکرد ماشینی معرفی‌شده در این تحقیق به‌کار برد.

فهرست منابع

سلطانی، پوری، و کامران فانی. ۱۳۶۲. *سرعنوان‌های موضوعی فارسی*. تهران: کتابخانهٔ ملی ایران.

عابدینی، مجتبی، شهرام جعفری و رضا جاویدان. ۱۳۹۰. یک راهکار جدید مبتنی بر ناظر جهت رفع ابهام معنای کلمه با استفاده از خاصیت نزدیک‌ترین همسایگی. *اولین همایش تخصصی سیستم‌های هوشمند کامپیوتری و کاربردهای آن‌ها*. وزارت علوم، تحقیقات و فناوری؛ دانشگاه پیام نور؛ دانشگاه پیام نور استان تهران.

علایی، الهام. ۱۳۹۵. *بررسی ساخت‌واژی هم‌نگاره‌های اسمی و صفتی به‌منظور برچسب‌دهی «اسم» به کلیدواژه‌ها در پیکره‌های علمی*. تهران: پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک). پژوهشکدهٔ مدیریت دانش.

محسنی، مهدی. ۱۳۸۷. *سیستم برچسب‌گذاری و ابهام‌زدایی خودکار اجزای کلام برای پیکره متنی زبان فارسی*. تهران: دانشگاه علم و صنعت. دانشکده مهندسی کامپیوتر.

مسعودی، بابک، سعید راحتی قوچانی 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. ماهشهر.

Banerjee, S. 2002. *Adapting the Lesk Algorithm for Word Sense Disambiguation to WordNet*. Department of computer science. Minnesota: University of Minnesota.

Bakx, G. E., 2006. *Machine learning techniques for Word Sense Disambiguation*. Universitat Politecnica de Catalunya.

Gaustad, T. 2004. *Linguistic Knowledge and Word Sense Disambiguation*. University of Groningen. Netherlands.http://www.Merriam-Webster.com

Ide, N. and J. Veronis. 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. Cambridge, MA: MIT press.

Pal, A., R., A. Munshi, and D. Saha. 2013. An approach to speed-up the word sense disambiguation procedure through sense filtering. *International journal of Instrumentation and Control systems* (IJICS). 3 (4): 29-41.

Rasekh A., M. Sadreddini and S. Fakhrahmad. 2014. Word sense disambiguation based on lexical and semantic features using Naive Bayes classifier. *Journal of Computing and Security* 2:123-132.

Sarrafzadeh, B., N. Yakovets, and N. Cerone. 2016. *Two novel approaches for Persian Word Sense Disambiguation. Computational Linguistics Project: Final Report*. University of Waterloo. Canada.

Wilks, Y., and M. Stevenson. 1997. The grammar of sense: using part- of- speech tags as a first step in semantic disambiguation. *Natural language processing*. Cambridge university press.

_____. 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. http://www.merriam-webster.com

الهام علایی ابوذر

دانش‌آموخته دکتری تخصصی در رشته زبان‌شناسی همگانی از دانشگاه تهران است. ایشان هم‌اکنون استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

حوزه زبان‌شناسی رایانشی (سیستم‌های تبدیل متن به گفتار، نظام‌های نوشتاری، یادگیری ماشینی و پردازش زبانی)، زبان‌شناسی نظری و پیکره‌ای از جمله علایق پژوهشی وی است.