رفع ابهام معنایی واژگان مبهم فارسی با مدل موضوعی LDA

بابک مسعودی[1] و سعید راحتی قوچانی[2]

[1]گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور زابل، سیستان و بلوچستان، ایران

[2]گروه برق، دانشگاه آزاد اسلامی واحد مشهد، مشهد، ایران

چکیده

ابهام‌زدایی از واژگان مبهم و دارای معانی متعدد، موضوع مهمی در حوزهٔ پردازش زبان‌های طبیعی است. در این مقاله، مدلی برای رفع ابهام از واژگان مبهم فارسی با استخراج ویژگی‌های جدید پیشنهاد شده است. برای ایجاد این مدل دو دسته ویژگی واژگان و نشانه‌های همراه واژهٔ مبهم و ویژگی‌هایی که با به‌کاربردن روش‌های مدل‌سازی موضوع به‌دست می‌آید، استفاده شده است. یک مدل موضوعی، مدلی آماری برای استخراج چکیدهٔ موضوع‌های موجود در اسناد یک پیکره است. در مقاله حاضر ما از روش بدون سرپرستی تخصیص پنهان دریکله (LDA) برای این منظور استفاده کرده‌ایم. نتایج آزمایش‌ها برای پانزده واژهٔ مبهم پرتکرار در زبان فارسی که از پیکرهٔ پژوهشکده پردازش هوشمند علائم استخراج شد، دقّت حدود ۹۷٪ را نشان می‌دهد که بیان‌گر تأثیر این روش در یافتن معنی مناسب واژگان مبهم است.

واژگان کلیدی: تخصیص پنهان دریکله، چندمعنایی، رفع ابهام معنایی، مدل‌سازی موضوع.

۱- مقدمّه

در اکثر زبان‌های طبیعی، وجود برخی کلمات باعث ایجاد ابهام در متون آن زبان می‌شود. به این کلمات مبهم که دارای ساختار نوشتاری یکسان و معانی متفاوت هستند، هم‌نگاره (هم‌نویسه)[1] گفته می‌شود. ابهام‌زدایی از کلمات مبهم[2] موضوع مهمی در حوزهٔ پردازش زبان‌های طبیعی است و پایهٔ بسیاری از دیگر مباحث مطرح در این حوزه نیز هست. تاکنون پژوهش‌های گسترده‌ای با رویکردهای متعدد در زبان‌های دیگر انجام و نتایج چشم‌گیری در این زبان‌ها حاصل شده است. این موضوع سنگ‌بنای بسیاری از مباحث در حوزهٔ پردازش زبان طبیعی، همچون ترجمهٔ ماشینی، خطایاب، تبدیل متن به گفتار، تشخیص گفتار و درک متن است. به‌عنوان مثال مترجم زبان فارسی به انگلیسی برای ترجمهٔ واژه « شیر» با سه معنی "Lion"، "Valve"و"Milk" مواجه است که باید با توجه به بافت متن، مناسب‌ترین واژه را انتخاب کند. فعالیت بر روی این موضوع پژوهشی می‌تواند راه‌گشای این مباحث در زبان فارسی باشد.

۲- مروری بر پژوهش‌های پیشین

روش‌های بسیاری برای رفع ابهام از واژگان مبهم در زبان‌های مختلف پیشنهاد شده است؛ همچنین به‌منظور ایجاد انگیزه در بین پژوهش‌گران، جهت پژوهش در این زمینه از سال ۱۹۸۸، هر سه سال رقابت‌هایی تحت عنوان SemEval[3] برگزار می‌شود.

در (یاروسکی ،۲۰۰۰) الگوریتمی با سرپرستی بر پایه فهرست‌های تصمیم‌گیری[4] پیشنهاد شده است. در این الگوریتم مجموعه کاملی از ویژگی‌های همسایگی، ساخت واژی و نحوی استفاده و دقّت سامانه ۷۸/۱٪ گزارش شده است. (ونسترا و همکاران ، ۲۰۰۰) روشی مبتنی بر حافظه برای رفع ابهام معنایی لغات پیشنهاد می‌کند که دقّت به‌دست آمده از این سامانه ۷۵/۱٪ گزارش شده است. در

[1] Homograph
[2] Word-sense disambiguation
[3] http://aclweb.org/
[4] Decision lists

(فلورین و همکاران ، ۲۰۰۲) روشی با ترکیب چند دسته‌بندی‌کنندهٔ نامتجانس (بردار شباهت کسینوسی، مدل‌های بیزی[1] و فهرست‌های تصمیم‌گیری) پیشنهاد شده و مجموعه‌ای از ویژگی‌های استخراج‌شده از بافت متن نظیر روابط دستوری و برچسب‌های POS را برای مشخص‌کردن معنای مناسب واژه، استفاده کرده است. سامانهٔ GAMBEL (دکادت و همکاران ، ۲۰۰۴) بر پایهٔ یادگیری با سرپرستی مبتنی بر حافظه پیشنهاد شد که از ویژگی‌های بافت متن و منابع دانشی نظیر WordNet استفاده می‌کند. در SemEval-2007 سامانه NUS-Ml معرفی شد که از تخصیص پنهان دریکله[2](LDA) - یک مدل احتمالاتی بیزی سلسله‌مراتبی سه‌لایه و ویژگی‌های نحوی و موضوعی استفاده می‌کرد. (استیونسون و همکاران، ۲۰۱۲) نشان دادند کارایی یک الگوریتم بدون سرپرستی WSD با افزودن اطلاعات بافت متن بهبود می‌یابد. در (پریسیس و همکاران، ۲۰۱۳) به ویژگی‌های استخراج شده توسط سه سامانه مرسوم رفع ابهام معنایی بدون سرپرستی و یک سامانه با سرپرستی، ویژگی موضوع بافت استخراج‌شده توسط LDA افزوده شد، که نتایج به‌دست آمده نشان داد کارایی سامانه‌ها به‌طور قابل ملاحظه‌ای افزایش می‌یابد.

ا ز فعالیت‌هایی که در زمینه رفع ابهام واژگان فارسی صورت گرفته، به (مکّی، ۲۰۰۸) می‌توان اشاره کرد. در این مقاله روشی مبتنی بر پیکره و یک واژگان برای امتیازدهی به دسته تعلّق مفهومی هر معنی واژهٔ مبهم، پیشنهاد شده است. دقّت میانگین این روش برای پانزده واژهٔ مبهم ۹۱/۴۶٪ گزارش شده است. همچنین یک روش مبتنی بر قاعده در (سعیدی، ۲۰۰۹) پیشنهاد شده که با درنظرگرفتن این قواعد دقّت آن برای یکی از پیکره‌های فارسی ۹۱٪ است.

در مقاله حاضر روشی مبتنی بر پیکره و با سرپرستی جهت ابهام‌زدایی از واژگان مبهم فارسی پیشنهاد شده است. بر این اساس از یک پیکرهٔ فارسی برای هر واژهٔ مبهم، دو دسته ویژگی استخراج می‌شود. دستهٔ نخست ویژگی‌ها، واژگان و نشانه‌های هم‌نشین واژهٔ هدف و دستهٔ دوم موضوع بافت متن مورد نظر است. برای استخراج موضوع بافت در این مقاله از روش تخصیص پنهان دریکله (LDA)) (بلی، ۲۰۰۳) که یکی از روش‌های بدون سرپرستی مدل‌سازی موضوع[3] است، استفاده می‌شود. روش‌های مدل‌سازی موضوع، مدل‌های آماری از متون هستند و فرض می‌کنند موضوع‌های پنهان درون پیکره نهفته شده‌اند، که می‌توانند برای استخراج، دسته‌بندی متون و بخش‌بندی گفتگوها مورد استفاده قرار گیرند. از آن‌جا که رفع ابهام از واژگان مبهم به‌نوعی یک مسئلهٔ دسته‌بندی مجزا است، می‌توان با ایجاد مدلی برای هریک از واژه‌های مبهم، به‌کمک روش دسته‌بندی بیشینهٔ بی‌نظمی[4] از آن‌ها رفع ابهام کرد.

ادامه مقاله به این صورت بخش‌بندی شده است: بخش سوم اصول روش پیشنهادی را با توصیف مدل‌های موضوعی، مدل‌های زایا، تخصیص پنهان دریکله و دسته‌بندی بیشینهٔ بی‌نظمی بیان می‌کند. مدل پیشنهادی در بخش چهارم ذکر شده و آزمایش‌ها، نتیجه‌گیری و کارهای آینده در بخش‌های پنجم و ششم بیان شده‌اند.

۳- اصول روش پیشنهادی

۳-۱- مدل‌سازی موضوع

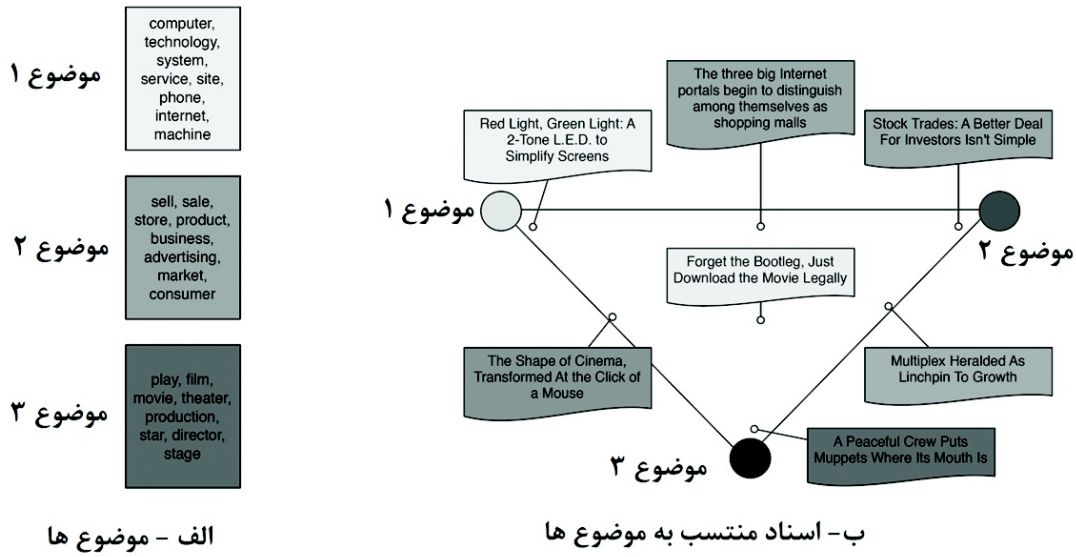
یک مدل موضوعی، مدلی است که موضوع‌های درون پیکره‌ای از اسناد را استخراج می‌کند و به هر سند برخی از این موضوع‌ها را نسبت می‌دهد. درواقع می‌توان گفت یک مدل موضوعی، جعبه‌ای سیاه با دو خروجی است: انتساب واژگان به موضوع‌ها و انتساب چندین موضوع به اسناد. نخستین خروجی، توزیع موضوع بر روی واژگان است. ما به‌طورعمومی موضوع‌ها را به‌صورت فهرست‌هایی از واژه‌ها نشان می‌دهیم. شکل(۱) سه موضوع استخراج شده از پیکره New York Times را نشان می‌دهد. هر واژه با درصدی احتمال، در یک موضوع حضور دارد. در اینجا تعدادی واژه با احتمال بیشتر نشان داده شده است. خروجی دیگر یک مدل موضوعی، امکان انتساب یک سند به چند موضوع مرتبط است.

شکل(۱) نشان می‌دهد چگونه یک سند می‌تواند با یک یا چند موضوع شناخته شود.

روش‌های متنوعی برای یافتن موضوع‌ها و انتساب آن‌ها به اسناد پیشنهاد شده‌اند. در این مقاله ما بر روی تخصیص پنهان دریکلهٔ (LDA) تمرکز کرده‌ایم. LDA یک مدل احتمالاتی و زایا است. یک مدل احتمالاتی است، زیرا با زبان احتمال بیان شده و زایا است؛ زیرا بیان می‌کند که داده‌ها چگونه به‌وجود آمده‌اند. در این مدل همهٔ جزییات شناخته شده نیستند؛ این قطعات گمشده را متغیرهای

[1] Bayes model
[2] Latent Dirichlet Allocation
[3] Topic Model
[4] Maximum Entropy

پنهان می‌نامیم و بـه روش‌هـای اسـتنتاج آمـاری جسـتجو می‌شوند تا بهترین توصیف را برای داده‌های قابـل مشـاهده فراهم آورند. این مدل از مجموعه فن‌های بـدون سرپرسـتی برای یافتن موضوع‌ها از داده‌های خام محسوب می‌شود.



(شکل- ۱): مثالی از توزیع کلمات بر روی موضوع‌ها و توزیع موضوع‌ها بر روی اسناد در پیکره New York Times (بوید-گرابر ۲۰۱۰)

۳-۲- مدل‌های زایا[1]

یک مدل زایا برای اسناد، براساس قوانین نمونه‌برداری سـادهٔ احتمالاتی توصیف می‌شود. در مدل زایا، هدف یافتن بهترین مجموعه از متغیرهای پنهانی است کـه مـی‌تواننـد داده‌هـای مشاهده‌شده (واژگان موجود در اسـناد) را بـه بهتـرین وجـه توصیف کنند. در شکل (۲) مدل موضوعی با رویکرد مجـزای مدل زایا و استنتاج آماری نشان داده می‌شود.

در سمت چپ، مدل زایا با دو موضوع مختلـف نشـان داده شده است. موضوع‌های ۱ و ۲ وابسته بـه «شـیر معـادل Lion» و «**شیر معادل** Milk» هسـتند و بـه‌عنـوان سـبدهای حاوی توزیع‌های مختلف روی واژگان در نظر گرفته می‌شوند. اسناد مختلف را می‌توان با برداشتن واژگان از یـک موضـوع، بسته به وزن داده‌شده، انتخاب کرد. به‌عنوان مثال اسناد یک و سه به‌ترتیب با نمونه‌برداری از موضوع‌های یک و دو ایجـاد می‌شوند؛ درحالی‌که سـند دو بـا ترکیبـی از هـر دو موضـوع ساخته می‌شود. توجه کنید، اعداد بالانویس نشان می‌دهـد از کدام موضوع برای واژه، نمونه‌برداری شده است.

مدل زایا که در این‌جا توصیف شـد، هـیچ فرضـی در مورد ترتیب واژگانی که در سند ظاهر شده‌انـد، نـدارد. تنهـا اطلاعات وابسته به مدل، دفعات تولید واژه است کـه بـه نـام سبد واژگان[2] شناخته مـی‌شـود و در بسـیاری از مـدل‌هـای آماری زبان، رایج اسـت. البتـه اطلاعـات ترتیـب واژه ممکـن است، حاوی اشاراتی به محتوای یـک سـند باشـند کـه ایـن اطلاعـات توسـط مـدل اسـتفاده نمـی‌شـوند گـریفیتس و همکاران، ۲۰۰۵).

سمت راسـت شـکل (۲) مسـئلهٔ اسـتنتاج آمـاری را نشان می‌دهد؛ به‌طوری که با واژگان مشـاهده شـده در یـک مجموعه از اسناد، تمایل داریم بدانیم کدام مدل موضوعی به توزیع موضوع‌های هر سند، شبیه‌تر است.

۳-۳- تخصیص پنهان دریکله(LDA)

مدل LDA فرایند زایای زیـر را بـرای ایجـاد پیکـره‌ای از M سند با N_d واژه در سند d بـا k موضـوع $\{\beta_1 \dots \beta_k\}$ فـرض می‌کند:

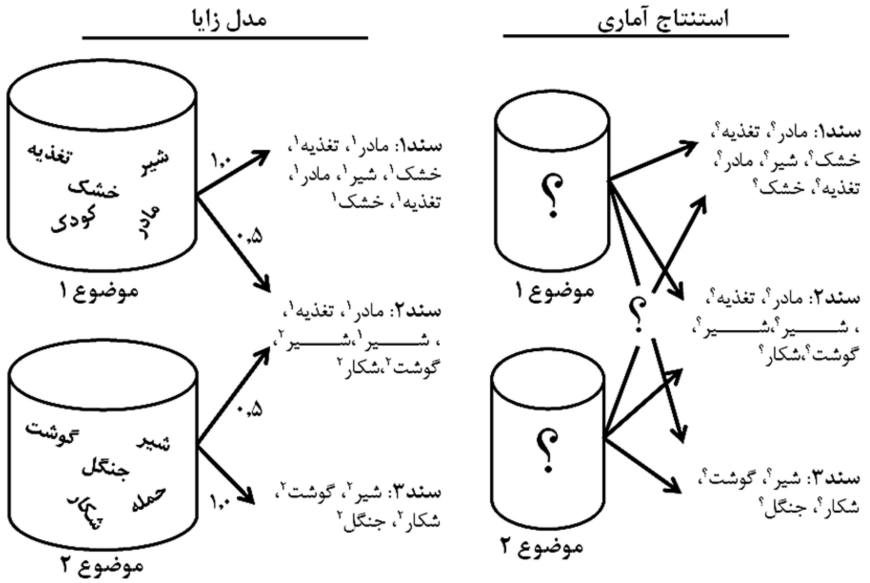
$$\begin{aligned}
&For\ each\ document\ d \in \{1 \dots .\} \\
&\quad a)\ Choose\ the\ document's\ topic\ weights \\
&\qquad \theta_d \sim Dir(\alpha) \\
&\quad b)\ For\ each\ word\ n \in \{1 \dots . n_d\} \\
&\qquad i)\ Choose\ topic\ assignment\ z_{d,n} \\
&\qquad\qquad \sim Mult(\theta_d)
\end{aligned}$$

[1]Generative Models

[2]Bag-of-Words

$ii)\ Choose\ word\ w_{d,n}\ \sim Mult(\beta_{z_{d,n}})$

در این فرآیند توزیع دریکله و توزیع چندجمله‌ای به‌صورت زیر تعریف می‌شود:



(شکل- ۲): نمایشی از مدل زایا و مسئله استنتاج آماری در تفسیر مدل موضوعی

$$\mathrm{Dir}(\theta|\alpha_1,\ldots,\alpha_k)=\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}\prod_k \theta_k^{\alpha_k-1} \qquad (۱)$$

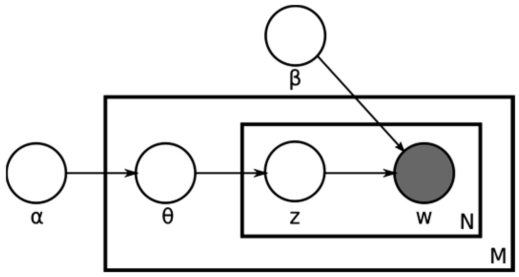
$$\mathrm{Mult}(n|\Theta 1,\ldots,\Theta k)=\frac{(\sum_k n_k)!}{\sum_k n_k}\prod_k \theta_k^{n_k} \qquad (۲)$$

α و β نیز پارامتر هستند.

توزیع (θ_d) برداری به طول k (تعداد موضوعات در مدل) است و برای توصیف موضوع منتسب‌شده z_n به هر یک از n واژه از سند d استفاده می‌شود.

k امین موضوع β_k برداری به طول V است، که هر بخش آن متناظر با احتمال حضور واژه در آن موضوع است. در شکل۱ (ب)، β_{zn} یکی از لغات ذخیره‌شده در فهرست‌ها و θ_d موقعیت قرارگرفتن اسناد را نشان می‌دهد.

ابزار دیگری که برای توصیف این مدل به‌کار برده می‌شود، نمودار شکل(۳) است. این نمودار فرایند LDA را نشان می‌دهد.



(شکل- ۳): نمودار فرآیند زایای LDA

متغیرهای پنهان، داده‌های قابل مشاهده و پارامترها هر کدام با یک گره در این نمودار نشان داده می‌شوند. هر وابستگی آماری توسط خط بین گره‌ها مشخص می‌شود، داده‌های قابل مشاهده با گره سایه‌دار و تکرار داده با زمینهٔ مستطیل نشان داده می‌شود، پایین‌نویس سمت راست پایین نشان‌دهنده تعداد تکرار متغیرهای درون مستطیل است.

از آن‌جا که فقط واژگان جمع‌آوری‌شده از اسناد قابل مشاهده است، یافتن موضوعات $\beta_{1:k}$، انتساب موضوع $z_{1:d}$ و توزیع موضوع $\theta_{1:D}$ برای هر سند، یک مسئله استنتاج آماری است.

۳-۴- دسته‌بندی بیشینه آنتروپی

روش دسته‌بندی آماری استفاده‌شده در این مقاله دسته‌بندی بیشینه آنتروپی است. بیشینهٔ آنتروپی فنی عمومی برای تخمین توزیع احتمال، از یک مجموعه داده است. اگر هیچ دانش نخستینی از داده‌ها در دسترس نباشد، آنتروپی بیشینه، یکنواخت‌ترین توزیع را انتخاب می‌کند؛ به‌طوری‌که همهٔ وقایع دارای احتمال برابر باشند.

درصورتی‌که مجموعهٔ داده در دسترس باشد، داده‌های دارای برچسب مجموعهٔ آموزش، به‌عنوان تعدادی ویژگی که محدودیت‌های مدل از آن‌ها استخراج می‌شود، در نظر گرفته می‌شوند. بنابراین درحالی‌که توزیع باید به‌گونه‌ای باشد تا آنتروپی را بیشینه کند، مدل نیز باید محدودیت‌های اعمال‌شده توسط برچسب‌های دادهٔ آموزشی را برآورده کند. بنابراین یک مدل بیشینهٔ آنتروپی، مدلی برای ارضای تمامی محدودیت‌های اعمال‌شده روی مجموعه داده‌ها است. این مدل براساس رابطهٔ زیر بنا می‌شود:

$$p(c|x)=\frac{1}{Z}\exp\left(\sum_i \lambda_i f_i(x,c)\right) \qquad (۳)$$

که در آن تابع ویژگی $f_i(x,c)$ تعداد دفعاتی است که ویژگی i استفاده شده تا برچسب c را برای رویداد x پیدا کند و وزن‌های λ_i برای بیشینه‌کردن درست‌نمایی در فرایند آموزش انتخاب می‌شوند، تا مقدار آنتروپی p بیشینه شود. مهم‌ترین برتری مدل بیشینهٔ آنتروپی آن است که این مدل چارچوبی برای یک‌پارچه‌کردن داده‌های ناهمگون به دست آمده از منابع مختلف اطلاعاتی، فراهم می‌کند (ویکام ، ۲۰۰۲).

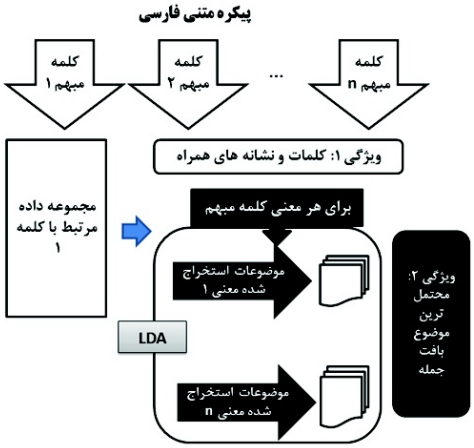
مدل احتمالاتی بیشینهٔ آنتروپی در بسیاری از کاربردهای حوزهٔ پردازش زبان‌های طبیعی نظیر برچسب‌گذاری اجزا کلام[1] (POS) و تعیین مرز جملات با موفقیت استفاده شده است (رتناپارکی، ۱۹۹۸).

۴- مدل پیشنهادی

برای ایجاد مدلی مؤثر جهت رفع ابهام واژگان هم‌نگاره فارسی از دو دسته ویژگی استفاده می‌شود. دستهٔ نخست، واژگان و نشانه‌های همراه واژه هدف و دستهٔ دوم، موضوع بافت جمله است که به‌کمک روش LDA استخراج شده است.

[1]Part of speech

شکل (۴)، نمودار استخراج ویژگی برای ایجاد مدل را نشان می‌دهد.



(شکل- ۳): نمودار استخراج ویژگی‌ها برای ایجاد مدل پیشنهادی

استخراج دو دسته ویژگی به‌شرح زیر است:

دستهٔ نخست- واژگان و نشانه‌های همراه: این ویژگی‌ها با درنظرگرفتن خصوصیات زبان فارسی انتخاب شده است. ابهام بسیاری از واژگان زبان فارسی، ناشی از نبود اعراب در متون نوشتاری است. از این جهت واژگان و نشانه‌گذاری‌هایی که بلافاصله قبل و بعد از واژهٔ هدف قرار می‌گیرند، نقش مؤثری در رفع ابهام از این واژگان دارند. در مثال زیر با دقت در نشانه‌گذاری قبل و بعد از واژه « شیر» معنی صحیح آن را می‌توان مشخص کرد:

شیر آب

شیر، آب

برای این کار با پنجره‌ای به مرکز واژهٔ مورد نظر و عرض ۲± کلیهٔ واژگان و نشانه‌گذاری‌های قبل و بعد از آن را به‌عنوان دستهٔ ویژگی نخست استخراج می‌کنیم. بدیهی است درصورت قرار گرفتن واژه در ابتدا یا انتهای جملات، مقادیر این ویژگی فضای خالی خواهد بود، که در نظر گرفته نمی‌شوند.

دستهٔ دوم- مدل‌سازی موضوع: جهت یافتن ویژگی مدل‌سازی موضوع در مرحله آموزش مدل، ابتدا سبد واژگان همراه هر واژهٔ مبهم استخراج می‌شود. برای این منظور یک پنجره به مرکز واژهٔ مورد نظر و عرض ۱۰± اطراف آن در نظر می‌گیریم. کلیهٔ واژگان به‌استثنای حروف ربط، اضافه و اعداد را به‌عنوان واژگانی که بافت جملهٔ مورد نظر را نشان می‌دهند، استخراج می‌کنیم؛ سپس به‌کمک روش تخصیص

پنهان دریکله (که در جعبه ابزار Topic Toolbox [1]Matlab پیاده‌سازی شده است) برای هر یک از معانی واژه‌های مبهم تعدادی موضوع در نظر گرفته و احتمال وقوع هر یک از واژگان استخراج شده و موجود در سبد واژگان را در هر موضوع محاسبه می‌کنیم. جداول ایجادشده برای هر کلمه مبهم جهت رفع ابهام در مراحل بعدی نگهداری خواهند شد. در مرحلهٔ آزمایش با جمع جداگانه احتمال حضور واژگان همنشین کلمه مبهم در هر موضوع، هر یک که احتمال وقوع در آن بیشتر باشد، مقدار ویژگی آن موضوع را برابر یک و برای سایر موضوعات این مقدار برابر صفر در نظر گرفته می‌شود. به این ترتیب برای واژه‌ای مانند شیر با سه معنی، تعداد T×۳ ویژگی استخراج می‌شود؛ که T تعداد موضوعات درنظرگرفته‌شده برای هر معنی است. به‌عنوان مثال در یکی از آزمایش‌ها تعداد موضوعات را سه در نظر گرفته‌ایم؛ بنابراین برای واژه « شیر» تعداد نه ویژگی دوتایی درنظر گرفته می‌شود که تنها یکی از آن‌ها یک و بقیه صفر خواهد بود.

پس از استخراج ویژگی‌ها، جداساز بیشینهٔ آنتروپی برای ایجاد سامانه رفع ابهام آموزش می‌بینند. در این‌جا برای هر کلمه مدلی با استفاده از روش دسته‌بندی بیشینهٔ آنتروپی و با استفاده از ابزار MALLET (مک کالم، ۲۰۰۲) که بر پایه جاوا تهیه شده است، ایجاد می‌شود.

۵- آزمایش‌ها

۵-۱- دادگان مورد استفاده

مجموعه دادهٔ مورد استفاده در این مقاله از پیکرهٔ متنی پژوهشکده پردازش هوشمند علائم[2] جمع‌آوری شده است. این پیکرهٔ متنی، مجموعه‌ای از متون نوشتاری و گفتاری زبان فارسی رسمی است که از منابع واقعی همچون روزنامه‌ها، پایگاه‌ها و ... جمع‌آوری شده، تصحیح شد و برچسب خورده است. حجم این دادگان درحدود یکصد میلیون واژه است و تنوع بسیار زیادی دارد.

برای ایجاد مجموعه‌های دادهٔ مورد نیاز از این پیکره متنی جملاتی را که حاوی پانزده کلمهٔ پر تکرار مبهم فارسی هستند، استخراج می‌کنیم.

جدول (۱) جزئیات کلمات مبهم مورد استفاده در آزمایش‌ها را نشان می‌دهد:

(جدول- ۱): جزئیات واژگان مبهم مورد استفاده در آزمایش‌ها

واژه	معانی	تعداد	تعداد جملات
شیر	Lion	۱۵۷	۶۸۰
	Milk	۴۴۸	
	Valve	۱۵۷	
سیر	Garlic	۱۰۷۲	۱۴۹۸
	Journey	۴۲۶	
مهر	Punch	۲۰۹	۷۲۹
	Love	۱۲۰	
	A Solar Month	۴۰۰	
جو	Atmosphere	۳۰۷	۴۱۷
	Barley	۱۱۰	
شکر	Sugar	۲۶۰	۳۷۱
	Thankfulness	۱۱۱	
حلال	Lawful	۸۲	۱۰۹
	Resolvent	۲۷	
خرد	Wisdom	۲۰۹	۴۰۷
	Tiny	۱۹۸	
نفس	Breath	۲۰۴	۷۶۵
	Oneself	۵۶۱	
محرم	ALunar Month	۲۳۱	۲۵۵
	Intimate	۲۴	
تن	Ton	۲۸۹۲	۴۳۴۹
	Body	۱۴۵۷	
اشکال	Figures	۲۶۴	۵۳۹
	Defect	۲۷۵	
اشراف	Dminance	۵۱	۱۰۷
	Knight,Gentleman	۵۶	
قسم	Oath	۷۱	۱۳۰
	Sort	۵۹	
سرم	My head	۹۳	۱۳۱
	Serum	۳۸	
شرف	Honor	۸۹	۱۳۶
	Soon Expected	۴۷	

۵-۲- نتایج آزمایش‌های مدل پیشنهادی

به‌منظور ارزیابی مدل پیشنهادی، برای هر کلمه مبهم سبد لغات همنشین آن از مجموعه آموزشی استخراج می‌شود. آزمایش‌ها برای رفع ابهام هر کلمه چندین مرتبه و هر بار با درنظرگرفتن مقداری متفاوت برای تعداد موضوع‌های موجود در هر معنی آن کلمه صورت گرفته و احتمال حضور کلمات همنشین موجود در سبد لغات توسط LDA ، محاسبه می‌شود. جداول۲- ۴ ده واژه (انتخاب شده از کل سبد

[1]http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm
[2]www.rcisp.com

واژگان) با بالاترین احتمال وقوع را در تعداد سه موضوع برای هر معنی واژه« شیر » نشان می‌دهد.

(جدول- ۲): ده واژه با بالاترین احتمال استخراج شده توسط LDA برای معنی ۱ واژه « شیر»

موضوع ۱	احتمال	موضوع۲	احتمال	موضوع۳	احتمال
سر	۰/۰۴۱۶۰	روباه	۰/۰۲۷۰۲	یک	۰/۰۴۷۰۰
آن	۰/۰۳۳۳۰	حق	۰/۰۲۱۶۳	شکار	۰/۰۴۱۱۳
نر	۰/۰۳۳۳۰	نه	۰/۰۲۱۶۳	گرگ	۰/۰۳۵۲۶
هوا	۰/۰۲۴۹۹	فیل	۰/۰۲۱۶۳	خود	۰/۰۲۹۴۰
مجسمه	۰/۰۲۴۹۹	مثل	۰/۰۲۱۶۳	روی	۰/۰۲۳۵۳
البته	۰/۰۱۶۶۹	حیوانات	۰/۰۲۱۶۳	فیلم	۰/۰۱۷۶۶
زبانی	۰/۰۱۶۶۹	حمله	۰/۰۱۶۲۳	مأمور	۰/۰۱۷۶۶
فرار	۰/۰۱۶۶۹	سنگی	۰/۰۱۶۲۳	فعل	۰/۰۱۷۶۶
پارک	۰/۰۱۶۶۹	خورشید	۰/۰۱۶۲۳	جایزه	۰/۰۱۷۶۶
پیدا	۰/۰۱۶۶۹	دین	۰/۰۱۶۲۳	تخته-سنگ	۰/۰۱۱۷۹

(جدول- ۳): ده واژه با بالاترین احتمال استخراج شده توسط LDA برای معنی ۲ واژه « شیر»

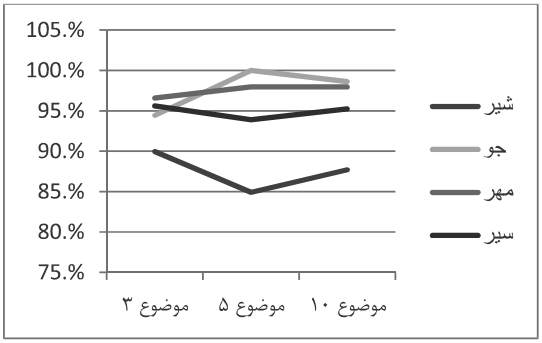
موضوع۴	احتمال	موضوع۵	احتمال	موضوع۶	احتمال
خشک	۰/۰۱۴۳۲	مادر	۰/۰۴۲۶۸	تولید	۰/۰۲۹۵۸
گرم	۰/۰۱۳۱۳	یک	۰/۰۳۰۸۸	هزار	۰/۰۲۰۷۰
گوشت	۰/۰۱۳۱۳	آن	۰/۰۱۸۱۷	تن	۰/۰۱۹۷۲
دیگر	۰/۰۱۱۹۴	قیمت	۰/۰۱۷۲۶	افزایش	۰/۰۱۷۷۵
تولیدی	۰/۰۱۱۹۴	تغذیه	۰/۰۱۷۲۶	آب	۰/۰۱۷۷۵
آنها	۰/۰۱۰۷۵	استفاده	۰/۰۱۶۳۵	خود	۰/۰۱۶۷۷
قرار	۰/۰۱۰۷۵	کارخانه	۰/۰۱۵۴۴	درصد	۰/۰۱۶۷۷
مثل	۰/۰۰۹۵۵	دادن	۰/۰۱۴۵۴	گاو	۰/۰۱۵۷۸
سیر	۰/۰۰۸۳۶	صنایع	۰/۰۱۴۵۴	شرکت	۰/۰۱۵۷۸
کیلو	۰/۰۰۸۳۶	دامداران	۰/۰۱۲۷۲	پنیر	۰/۰۱۴۸۰

جدول (۵) دقت دسته‌بندی روش پیشنهادی با معیارهای دقت و فراخوانی را نشان می‌دهد. لازم به ذکر است در این آزمایش‌ها تعداد پنج موضوع برای هر معنی واژه در نظر گرفته شده است. علت وجود اختلاف در میزان دقّت هر واژه را می‌توان میزان تکرار واژه در پیکره و فراوانی هر کدام از معانی آن و همچنین میزان هم‌پوشانی موضوع‌های مختلف در برگیرندهٔ معانی واژه عنوان کرد.

(جدول- ۴): ده واژه با بالاترین احتمال استخراج شده توسط LDA برای معنی ۳ واژه « شیر»

موضوع۷	احتمال	موضوع۸	احتمال	موضوع۹	احتمال
تنظیم	۰/۰۴۱۹۱	آب	۰/۰۸۳۸۳	فشار	۰/۰۵۱۴۶
فشار	۰/۰۲۳۳۰	شیلنگ	۰/۰۲۷۹۶	باز	۰/۰۴۸۶۰
مثلاً	۰/۰۲۳۳۰	برقی	۰/۰۲۷۹۶	یک	۰/۰۴۲۸۹
تخلیه	۰/۰۱۸۶۵	گاز	۰/۰۲۷۹۶	استفاده	۰/۰۲۵۷۴
اصلی	۰/۰۱۸۶۵	کار	۰/۰۲۲۳۸	کنترل	۰/۰۲۲۸۹
صفحه	۰/۰۱۴۰۰	هنگام	۰/۰۱۹۵۸	آب	۰/۰۲۰۰۳
رانش	۰/۰۱۴۰۰	سیلندر	۰/۰۱۹۵۸	طرف	۰/۰۱۷۱۷
دیگ	۰/۰۱۴۰۰	دو	۰/۰۱۹۵۸	دهانه	۰/۰۱۶۲۶
عمل	۰/۰۱۴۰۰	کردن	۰/۰۱۶۷۹	قرار	۰/۰۱۴۳۲
دما	۰/۰۱۴۰۰	سه	۰/۰۱۴۰۰	هوا	۰/۰۱۴۳۲
دکمه	۰/۰۱۴۰۰	تحریک	۰/۰۱۴۰۰	قطع	۰/۰۱۴۳۲
تنها	۰/۰۱۴۰۰	سمت	۰/۰۱۴۰۰	عنوان	۰/۰۱۴۳۲

در آزمایشی دیگر اثر تغییر در انتخاب تعداد موضوع برای هر معنی واژه بررسی شد. شکل زیر دقت سامانه برای ۳، ۵ و ۱۰ موضوع را نشان می‌دهد.



(شکل- ۴): مقایسه اثر تغییر تعداد موضوع در دقت دسته‌بندی

با توجه به نمودار بالا می‌توان دریافت تغییر در انتخاب موضوع برای معانی مختلف یک واژه، می‌تواند اثرات مختلفی داشته باشد؛ اما به‌طورکلی می‌توان گفت با زیادشدن تعداد موضوع، واژه در موضوع‌هایی با جزییات بیشتر قرار می‌گیرد و با کم‌شدن تعداد آن ها معانی واژه در دسته‌های کلی‌تر قرار داده می‌شوند. افزایش تعداد موضوع‌ها در بعضی واژگان ممکن است باعث افزایش دقّت و در بعضی باعث کاهش دقّت شود. علت این مسئله را می‌توان در نوع مجموعه دادهٔ مورد استفاده و یا به دستهٔ مفهومی که واژه به آن تعلق دارد، یافت؛ که نیاز به پژوهش بیشتر دارد.

(جدول- ۵): کارایی روش پیشنهادی به درصد

واژه	معانی	دقت Precision	فراخوانی Recall	دقت کل
شیر	Lion	۸۹/۴۲	۸۷/۳۹	۹۲/۷۳
	Milk	۹۵/۳	۹۶/۶۴	
	Valve	۹۳/۳۴	۹۰/۵۹	
سیر	Garlic	۹۹/۹۴	۹۸/۲۳	۹۸/۶۴
	Journey	۹۷/۳۵	۹۷/۴۲	
مهر	Punch	۹۹/۳۶	۹۹/۸۵	۹۹/۳۱
	Love	۹۱/۷۱	۹۳/۳۳	
	A Solar Month	۹۹/۶۴	۹۷/۴۱	
جو	Atmosphere	۱۰۰	۱۰۰	۱۰۰
	Barley	۱۰۰	۱۰۰	
شکر	Sugar	۹۸/۶۷	۹۵/۴	۹۷/۳۵
	Thankfulness	۹۸/۷۱	۹۶/۵۶	
حلال	Lawful	۹۹/۸۹	۹۶/۳۲	۹۷/۴۲
	Resolvent	۹۸/۹۷	۹۷/۶۶	
خرد	Wisdom	۹۸/۹۴	۹۴/۶۲	۹۷/۲۸
	Tiny	۹۹/۳۱	۹۶/۸۶	
نفس	Breath	۹۹/۳۸	۹۹/۹۲	۹۹/۲۵
	Oneself	۹۷/۵۶	۹۸/۵۶	
محرم	Lunar Month	۸۳/۷۹	۸۹/۹۶	۹۶/۸۴
	Intimate	۹۷/۵۹	۹۶/۵۶	
تن	Ton	۸۹/۵۷	۹۴/۳	۹۳/۵۱
	Body	۹۰/۳۶	۹۳/۷۸	
اشکال	Figures	۹۸/۴۵	۹۸/۹۴	۹۹/۰۱
	Defect	۱۰۰	۹۷/۷۲	
اشراف	Dminance	۹۵/۷۴	۹۷/۳۷	۹۷/۱۳
	Knight ,Gentleman	۹۴/۴۵	۹۶/۵۴	
قسم	Oath	۹۹/۶۱	۹۹/۱۶	۹۹/۵۸
	Sort	۹۹/۴۸	۱۰۰	
سرم	My head	۱۰۰	۱۰۰	۱۰۰
	Serum	۱۰۰	۱۰۰	
شرف	Honor	۹۹/۷۲	۹۸/۸۲	۹۹/۶۴
	Soon Expected	۹۸/۲۴	۹۹/۳۱	

۶- نتیجه‌گیری و کارهای آینده

این مقاله به بحث واژگان هم‌نگارهٔ فارسی و ارائهٔ روشی برای ابهام‌زدایی از آن‌ها می‌پردازد. اگرچه در بسیاری از زبان‌های طبیعی به‌طور گسترده از مدل‌های موضوعی در کاربردهای مختلف استفاده می‌شود، در زبان فارسی کمتر به استفاده از آن‌ها پرداخته شده است. در این مقاله از مدل موضوعی LDA برای استخراج موضوع‌های مختلف هر معنی واژه استفاده شده است و به این ترتیب موضوع بافت جمله نیز به‌عنوان یک عامل مؤثر در یافتن معنی واقعی کلمه مورد استفاده قرار می‌گیرد. نتایج آزمایش‌ها نشان می‌دهد استخراج ویژگی‌های ذکرشده و استفاده از روش دسته‌بندی بیشینه آنتروپی با دقّت بالایی قادر به مشخص‌کردن معنی مناسب واژه است؛ علاوه‌بر این استفاده از مدل موضوعی LDA برای مجموعه آموزشی عملی وقت‌گیر نبوده و پس از آماده‌سازی مجموعه‌های مورد نیاز کار رفع ابهام، به‌سادگی انجام خواهد گرفت. در مقایسه می‌توان گفت روش پیشنهادی با کاهش فضای ویژگی‌ها و استفاده از ویژگی‌های مؤثر، قابلیت ابهام‌زدایی از واژگان مبهم فارسی را با دقّت بالایی دارد.

در کارهای آینده از وردنت فارسی (فارس نت)[1] در کنار مدل‌های موضوعی (همان‌طور که در برخی زبان‌ها صورت گرفته است) جهت رفع ابهام معنایی از واژگان فارسی استفاده خواهد شد. علاوه‌بر این مدل‌های موضوعی دیگر نظیر..plsa, hdp, را نیز باید آزمود تا مناسب‌ترین مدل برای زبان فارسی به‌دست آید.

۷- مراجع

Blei D. M. and et al., "Latent Dirichlet allocation", Journal of Machine Learning Research., (2003) , vol. 3, pp. 993-1022.

Boyd-Graber, Jordan. L., "Linguistic Extensions of Topic Models", Princeton University, (2010), pp 2-4.

Decadt B. and et al., "GAMBL, Genetic Algorithm Optimization of Memory-Based WSD", Proceedings of the ThirdInternational Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Sens-eval-3), (2004).

Florian R and, et al., "Combining Classifiers for word sense disambiguation", Natural Language Engineering,(2002), vol. 8, pp. 327-341.

Griffiths, T. L., D. M. Blei, et al. "Integrating topics and syntax.", Advances in Neural Information Processing, (2005), 17: 537-544.

Makki R. and Homayoun pour M. M., "Word Sense Disambiguation of Farsi Homographs Using Thesaurus and Corpus", presented at the Proceedings of the ۶th international conference on Advances in Natural Language Processing, (2008), Gothenburg, Sweden.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit

[1] http://nlp.sbu.ac.ir/site/farsnet/

Preiss J., Stevenson M., "Unsupervised Domain Tuning to Improve Word Sense Disambiguation", in Proceedings of NAACL-HLT, (2013), pp. 680-684.

Ratnaparkhi, A., "MAXIMUM ENTROPY MODELS FOR NATURAL LANGUAGE AMBIGUITY RESOLUTION", University of Pennsylvania. Doctor of Philosophy, (1998).

Saedi C. and Shamsfard M., "Translating Persian documents into English using knowledge based WSD", in Digital Information Management, ICDIM 2009 .Fourth International Conference, (2009), on, pp. 1-6

Steyvers M., Griffiths T., "Probabilistic Topic Models", In Landauer, T., Mcnamara, D., Dennis, S., Kintsch, W., Latent Semantic Analysis: A Road to Meaning. ,(2006).

Stevenson M, Agirre, E and Soroa A, "Exploiting domain information for Word Sense Disambiguation of medical documents", Journal of the American Medical Informatics Association, (2012), vol. 19, pp. 235-240.

Veenstra J. and et al., "Memory-Based Word Sense Disambiguation", Computers and the Humanities, (2000), vol. 34, pp. 171-177.

Weikum G., "Foundations of statistical natural language processing", SIGMOD Rec, (2002), vol. 31, pp. 37-38.

Walter D., TiMBL: Tilburg Memory-Based Learner, Tilburg University and CNTS Research Group,University of Antwerp (2006).

YarowskyD., "Hierarchical Decision Lists for Word Sense Disambiguation", Computers and the Humanities, (2000), vol. 34, pp. 179-186.
(http://rahati.mshdiau.ac.ir)

بابک مسعودی مدرک کارشناسی خود را در رشته مهندسی برق – الکترونیک در سال ۱۳۷۹ از دانشگاه سیستان و بلوچستان و مدرک کارشناسی ارشد خود را از دانشگاه آزاد اسلامی مشهد در سال ۱۳۸۹ اخذ کرد. وی هم‌اکنون عضو هیئت علمی دانشگاه پیام نور زابل است. زمینه‌های پژوهشی ایشان پردازش زبان طبیعی و پردازش تصاویر دیجیتال است.

نشانی رایانامهٔ ایشان عبارتست از:

B.masoudi@pnu.ac.ir

سعید راحتی متولد سال ۱۳۴۶ در شهرستان قوچان است. وی در سال ۱۳۶۴ در رشته الکترونیک دانشکدهٔ فنی دانشگاه تهران تحصیلات دوره کارشناسی خود را آغاز کرد. پس از فراغت از تحصیل، دورهٔ کارشناسی ارشد خود را در رشتهٔ مخابرات دانشگاه آزاد اسلامی واحد تهران جنوب در سال ۱۳۷۲ به اتمام رساند؛ سپس مدرک دکترای خود را در سال ۱۳۷۸ در رشته مخابرات دانشگاه آزاد اسلامی واحد علوم و تحقیقات اخذ کرد. با پایان دورهٔ دکترا، به‌عنوان استادیار و از سال ۱۳۹۰ در سمت دانشیاری دانشگاه آزاد اسلامی واحد مشهد مشغول به‌خدمت شد. تاکنون بیش از یکصد مقاله در زمینهٔ سامانه‌های هوشمند از ایشان به چاپ رسیده است.

نشانی رایانامهٔ ایشان عبارتست از:

Rahati@mshdiau.ac.ir