چکیده
یکی از چالش‌های پیش‌روی پردازش زبان طبیعی زبان عربی، رفع ابهام میان تحلیل‌های صرفی ممکن یک کلمه به تناسب جایگاه آن کلمه در جمله است. یک تحلیلگر صرفی، به ازای هر کلمه ورودی، ممکن است بیشتر از یک جواب صرفی داشته باشد. تشخیص اینکه کدام‌یک از تحلیل‌های ممکن، تحلیل صحیح کلمه مورد نظر است، توسط ابزارهای رفع ابهام صورت می‌گیرد. این مقاله، به یکی از قوی‌ترین ابزارهای رفع ابهام اشاره می‌کند که به صورت خاص، برای تحلیلگر صرفی نور (Noor Morphological Analyzer) طراحی شده است. این ابزار که ابهام‌زدای صرفی نور (Noor Morphological Disambiguation) نامیده شده، ترکیبی از الگوریتم‌های یادگیری و قاعده‌محور است. آزمایش‌های این تحقیق نشان می‌دهد که رفع ابهام صرفی نور می‌تواند با دقت ۸۸ درصد خروجی‌های تحلیلگر صرفی نور را رفع ابهام نماید.
کلمات کلیدی: صرف زبان عربی، یادگیری ماشین، اِعراب‌گذاری خودکار، برچسب ادات سخن.

ابهام‌زدایی هوشمند صرفی نور

محمدحسین الهی‌منش*
elahimanesh@noornet.net

مقدمه
پردازش زبان طبیعی، یکی از معروف‌ترین زمینه‌های علمی حال حاضر جهان به شمار می‌آید. این زمینه علمی، یکی از زیرشاخه‌های بااهمیت در حوزه گسترده علوم رایانه، هوش مصنوعی، و نیز دانش زبان‌شناسی محاسباتی است که به تعامل بین رایانه و زبان‌های (طبیعی) انسانی می‌پردازد. یکی از ویژگی‌های این علم، وابستگی آن به زبان مورد نظر است. ازاین‌رو، لازم است برای هر زبان، تحقیقات خاصی صورت گیرد. زبان عربی، یکی از زبان‌های اجتماعی پُرکاربرد با زبان‌شناسی متفاوت از زبان‌های هندی اروپایی محسوب می‌گردد.[۱] این زبان داری قواعد صرفی غنی است که نه تنها صرف فعل، بلکه واژک‌هایی همچون: ضمیر، ترکیب عطفی و حروف جر را نیز داراست.[۲] پردازش زبان طبیعی این زبان، مشکلات فراوانی را به همراه دارد؛ زیرا دارای تحلیل صرفی پیچیده است.[۳] تحلیل صرفی یک کلمه، به معنای تعیین تعداد بسیاری از ویژگی‌هاست که از جمله آن، برچسب ادات سخن (برای نمونه: فعل، اسم و حرف)، معلوم و مجهول، جنسیت، تعداد و اطلاعات در خصوص هریک از قطاع‌های کلمه است. فهرست ویژگی‌هایی را که تحلیلگر صرفی نور به عنوان خروجی می‌دهد، می‌توان کامل‌ترین نمونه از این ویژگی‌ها دانست. فهرست ذیل، ویژگی‌های مختلف خروجی این تحلیلگر را نشان می‌دهد:

* کارشناس متن‌کاوی معاونت فنی نور.

شماره	اسم ویژگی	توضیح	نوع کلماتی که این ویژگی را دارند
۱	Affix	نوع وند	اسم، فعل و حرف
۲	Root	ریشه	اسم و فعل
۳	Lemma	پیراسته	اسم، فعل و حرف
۴	(POS(Part Of Speech	برچسب ادات سخن	اسم، فعل و حرف
۵	(KOL(Kind of Letter	نوع حرف	حرف
۶	Num	مفرد، مثنی، جمع	فعل
۷	Case	اعراب	اسم، فعل و حرف
۸	Categ	باب	اسم و فعل
۹	Time	زمان	فعل
۱۰	(TOV(Type of Verb	نوع فعل	فعل
۱۱	Voice	معلوم و مجهول	فعل
۱۲	Dervt	نوع اسم	اسم

پژوهش حاضر، یکی از قوی‌ترین شیوه‌های رفع ابهام تحلیلگر صرفی زبان عربی را ارائه کرده که اسم آن NoorMD (Noor Morphological Disambiguation) است. این شیوه، از تحلیلگر صرفی نور (NoorMA) به عنوان تحلیلگر خود بهره برده، مراحل متعددی را برای رفع ابهام خروجی این تحلیلگر به کار می‌برد. ابهام موجود در خروجی تحلیلگرها، یک امر طبیعی برای آنهاست. ابهام در خروجی یک تحلیلگر، به این معناست که کلمه مورد پرسش، دارای کاربردها و معانی مختلف و به‌تبع تحلیل‌های مختلف است؛ برای نمونه، کلمه «فحکم» را در نظر بگیرید. این کلمه، در کاربردهای مختلف، هم به‌صورت اسم و هم به صورت فعل دیده شده است. جدول ذیل، چند نمونه از ۲۰۱ پاسخ مختلف NoorMA برای این کلمه را نشان می‌دهد.

کلمه با اعراب	توکن‌ها	POS پیشوند	POS میانوند
فَحِکَمِ	ف حکم	حرف	اسم
فَحَکَمٍ	ف حکم	حرف	اسم
فَحْکُمْ	فح کم	فعل	اسم
فَحَکَمَ	ف حکم	حرف	فعل

شیوه ابداعی این تحقیقات، تاکنون بهترین نتایج را در زمینه رفع ابهام صرف زبان عربی به ثبت رسانیده است. علاوه بر این، باتوجه به پوشش بسیار قوی NoorMA، ویژگی‌های متعددی از صرف کلمات عربی به عنوان خروجی ارائه می‌گردد که این سطح از جزئیات، در نوع خود بی‌نظیر است. در ادامه، توضیح بیشتری نسبت ساختار شیوه پیشنهادی این تحقیقات ارائه شده است. ازاین‌رو، طی چهار محور، به بحث و بررسی در این باره می‌پردازیم:

کارهای مرتبط، الگوریتم پیشنهادی، ارزیابی الگوریتم پیشنهادی و نتیجه‌گیری.

کارهای مرتبط

تحقیقات بسیاری در گذشته برای برچسب‌گذاری ادات سخن و تحلیل صرفی زبان عربی صورت گرفته است. این تحقیقات را می‌توان از چند جنبه تقسیم‌بندی نمود؛ یکی از تقسیم‌بندی‌های ممکن برای آن، از لحاظ انواع زبان عربی است. عربی استاندارد مدرن و عربی مصری، دو نوع زبان عربی است که بیشتر در تحقیقات گذشته به آن پرداخته شده است. برخی از تحقیقات گذشته، تمرکز خود را روی برچسب‌گذاری ادات سخن عربی استاندارد مدرن گذاشته‌اند ([۳][۴][۵]) و همچنین برخی روی عربی مصری متمرکز شده‌اند ([4][۶]). متناسب با هریک از این نوع‌ها، پیکره‌هایی نیز انتشار یافته که سعی شده پوشش خوبی روی آن نوع باشد؛ برای نمونه، پیکره Penn Arabic Tree Bank را می‌توان یکی از معروف‌ترین پیکره‌ها در رده عربی استاندارد مدرن دانست. [۷] این پیکره، از چهار بخش تشکیل شده که در دو سطح بر روی کلمات این پیکره غنی‌سازی صورت گرفته است؛ سطح اوّل، به برچسب ادات سخن کلمات مربوط است. در برچسب ادات سخن این پیکره، متن به واحدهای لغوی شکسته شده و به ازای هر واحد، ویژگی‌های همچون: زمان، معلوم و مجهول و جنسیت مشخص شده است. سطح دوم، به بانک درخت عربی مربوط است. پیکره‌ای شبیه به Penn Arabic Treebank برای زبان

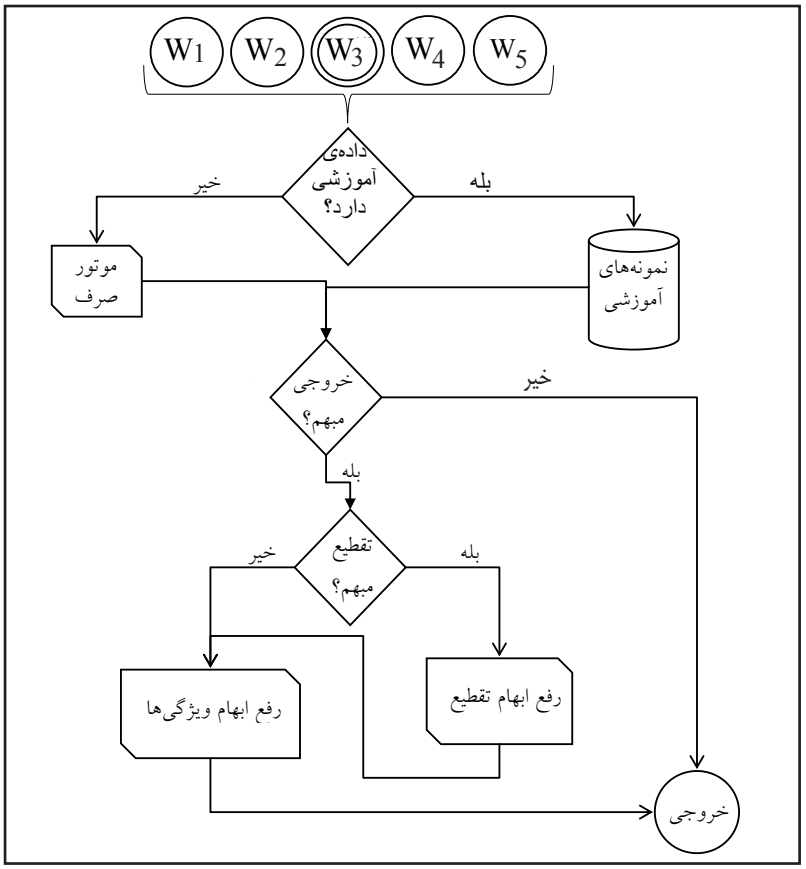
مصری با نام Egyptian Arabic Treebank توسعه یافته که عمده تحقیقات مرتبط به این موضوع در این نوع از زبان عربی روی آن صورت گرفته است. [۸]

یکی دیگر از جنبه‌های رده‌بندی کارهای گذشته در زمینه برچسب‌گذاری ادات سخن عربی، به نوع متون از لحاظ زمانی مرتبط است. برخی کارهای گذشته، تمرکز خود را روی متون قدیمی گذارده‌اند و در این راستا، اقدام به توسعه برچسب‌گذار و تحلیلگر صرفی نموده‌اند. این تحقیقات از لحاظ نوع متون مورد بحث، بیشترین شباهت را با پژوهش‌های ما دارد. Elhadj و همکارانش، از ترکیب Hidden Markov Model و تحلیلگر صرفی برای برچسب‌گذاری ادات سخن متونی از صده سوم هجری بهره برده‌اند. آزمایش‌های این تحقیقات روی پیکره‌ای از کتاب ALJAHEZ صورت گرفته که این پیکره در مجموع ۵۶۳۱۲ کلمه دارد. نتایج ارزیابی آنها در شرایطی که ۱۳ برچسب ادات سخن را به عنوان خروجی می‌دهند، ۹۶٪ دقت دارد. نکته قابل توجه این است که Elhadj و همکارانش با هدف تحلیل و تفسیر قرآن، اقدام به این تحقیقات نموده‌اند. در ادامه، این پژوهش‌ها برای ارزیابی روشی پیشنهادی‌شان روی قرآن، آن را روی هفت شعر از کتاب AL-Fatiha آزمایش نموده‌اند که نتایج اولیه آنها، دقت ۹۴٪ را نشان می‌دهد.

رده‌بندی دیگری را که می‌توان برای تحقیقات گذشته در زمینه برچسب‌گذاری ادات سخن عربی دانست، نوع برچسب‌گذار استفاده شده برای این کار است. در پژوهش‌های گذشته، از ووش‌های HMM، CRF، Memory base tag-ger، Genetic Algorithm و SVM استفاده شده که همین روش‌ها را می‌توان رده‌های مختلف برچسب‌گذارها از لحاظ معماری برچسب‌گذارها دانست. تعداد بسیاری از تحقیقات گذشته، از HMM به عنوان یکی از معروف‌ترین الگوریتم‌های Sequence labeling بهره‌برده‌اند. از جمله این تحقیقات، روش پیشنهادی [۹] است. Ali و همکارانش از الگوریتم GA بهره برده‌اند. [۱۰]

الگوریتم پیشنهادی

در این بخش، جزئیات الگوریتم پیشنهادی این تحقیقات ارائه گردیده و به طور مفصل، در خصوص هریک از اجزای آن صحبت می‌شود. در حقیقت، الگوریتم پیشنهادی این تحقیقات، از معماری منحصربه‌فردی استفاده می‌کند که هریک از اجزای الگوریتم پیشنهادی در آن نمودی دارند. در این معماری که شکل زیر آن را به تصویر کشیده، ورودی رشته‌ای از کلمات است (w1 w2 w3 w4 w5) که کلمه مورد پرسش در میانه آن قرار دارد. استفاده از پنجره پنج‌کلمه‌ای که کلمه مورد نظر در میان آن قرار گرفته، در بسیاری از تحقیقات پردازش زبان طبیعی مشاهده می‌گردد؛ [۱۱]



روش ارائه‌شده در این تحقیق، یکی از برترین روش‌های رفع ابهام هوشمند زبان عربی محسوب می‌شود که با استفاده از داده‌های آموزشی بسیار غنی و فنون یادگیری ماشین، به دقتی بی‌نظیر در رفع ابهام هوشمند زبان عربی دست یافته است. این رفع ابهام، لازمه استفاده از موتور صرف زبان عربی به شمار می‌رود. کاربردهایی همچون: ریشه‌یابی و جست‌وجوهای مبتنی بر ریشه‌یابی، موتورهای جست‌وجو و مشابه‌یابی متون، برخی از کاربردهای ویژه رفع ابهام هوشمند هستند. چنین کاربردهای، هم‌اکنون در برخی پروژه‌های مرکز تحقیقات کامپیوتری علوم اسلامی مشاهده می‌شوند

[۱۲] علت آن نیز تعادل خوب میان کیفیت و سرعت فعالیت‌های پردازش زبان طبیعی روی این اندازه از پنجره است. در انتهای این معماری، تحلیل‌های صرفی کلمه مورد نظر برحسب کیفیت مرتب شده و به عنوان خروجی ارائه می‌گردد. در ادامه، هر یک از مراحل این معماری توضیح داده شده است.

- داده آموزشی دارد؟

اولین مرحله از الگوریتم پیشنهادی، تشخیص حضور کلمه مورد نظر در داده‌گان آموزشی است. این مرحله، ظاهری ساده دارد؛ اما در درون خود روال نسبتاً پیچیده‌ای را طی می‌کند. در این مرحله، با توجه به اینکه دادگان آموزشی این تحقیقات دارای اعراب است، ابتدا کلمه ورودی توسط موتور اعراب نور، اعراب‌گذاری می‌گردد. این موتور که از پیکره اعراب نور استفاده کرده، کیفیت بسیار بالایی در تشخیص اعراب صحیح کلمات دارد.[۱۳] استفاده از این موتور، یکی از مهم‌ترین مراحل الگوریتم پیشنهادی به شمار می‌آید. علت این استفاده، وابستگی شدید ابهام صرفی به اعراب‌دار بودن یا نبودن کلمات است. در حقیقت، بسیاری از ابهام کلمات در زبان عربی با به‌دست‌آوردن اعرابِ صحیح کلمه از بین می‌رود. آزمایش‌های این تحقیقات نشان می‌دهد که در صورت بدون اعراب بودن کلمات، صرف آنها به‌طور میانگین دارای ۵۱ جواب مختلف است؛ در صورتی که پس از اعراب‌گذاری، این تعداد به سه جواب کاهش می‌یابد. این آزمایش روی ۵۰۰۰ کلمه از پیکره نور صورت گرفته است که عمده آنها، از کتاب الکافی نمونه‌بردای گردیده است.

- نمونه‌های آموزشی

در این مرحله، پاسخ‌های متناظر با کلمه‌ای که نمونه آموزشی برای آن وجود دارد، از داده‌های آموزشی بیرون کشیده می‌شود. در شرایطی که یک کلمه دارای داده آموزشی باشد، می‌توان گفت که تحلیل‌های آمده در داده آموزشی، همان تحلیل‌های ممکن آن کلمه است. در حقیقت، بسیاری از تحلیل‌هایی که یک تحلیلگر صرفی به عنوان خروجی می‌دهد، کاربردی نیستند. ازاین‌رو، در نظر نگرفتن آنها، به کاهش ابهام کمک بسیاری می‌کند. بنابراین، در این تحقیقات در شرایطی که کلمه دارای نمونه آموزشی باشد، دیگر به موتور تحلیلگر مراجعه نمی‌گردد.

- موتور صرف

بخشی از کلمات ورودی را نمی‌توان در داده‌های آموزشی پیدا نمود. این نوع از کلمات، به عنوان کلمات خارج از لغت‌نامه (Out Of Vocabulary) OOV یا کلمات ناشناخته (Unknown Word) شناخته می‌شوند. تحلیل این کلمات نسبت به کلماتی که در دادگان آموزشی هستند، مشکل‌تر است. علت آن هم، نبود یک توزیع احتمال میان تحلیل‌های مختلف ممکن برای این کلمات است. ازاین‌رو، چنین کلماتی در الگوریتم پیشنهادی به تحلیلگر صرفی ارجاع داده شده و تمامی جواب‌های این تحلیلگر، به عنوان یک تحلیل قابل قبول در نظر گرفته می‌شود. همان‌طور که گفته شد، در شرایطی که اعراب برای کلمه وجود نداشته باشد، به‌طور میانگین، هر کلمه ۵۱ آنالایز مختلف خواهد داشت؛ حتی در حالت وجود اعراب، باز هم به‌طور میانگین، هر کلمه سه تحلیل مختلف خواهد داشت. این سطح از ابهام کار، تشخیص پاسخ صحیح را برای کلمات OOV سخت کرده است. برای مقایسه با حالتی که کلمه درون دادگان آموزشی قرار گرفته، میزان ابهام را بیشتر توضیح می‌دهیم.

در شرایطی که کلمه درون پیکره آموزشی قرار گرفته باشد، احتمال اینکه تحلیل‌های مختلفی برای آن یافت گردد، کم است. آمارهای این تحقیقات نشان می‌دهد، برای هر کلمه به‌طور میانگین ۱.۵ پاسخ متفاوت در پیکره آموزشی حضور دارد. این در حالی است که کلمه‌ای که به تحلیلگر فرستاده می‌شود، به‌طور میانگین ۵۱ پاسخ خواهد داشت. گفتنی است که در این آزمایش، کلمات بدون اعراب در نظر گرفته شده‌اند.

- خروجی مبهم

این مرحله با بررسی پاسخ‌های آمده از مراحل دوم و سوم، در صورت عدم ابهام و تک‌جوابی بودن، پاسخ رسیده را به خروجی می‌فرستد. در شرایطی وجود ابهام نیاز به رفع ابهام خروجی‌هاست. ازاین‌رو، مراحلی طراحی شده که سطح به سطح، سعی در رفع ابهام میان پاسخ‌ها و رسیدن به بهترین ترتیب خروجی‌ها دارد. در ابتدای این مراحل، به هر پاسخ، یک مقدار کیفیت متصل شده و در طی مراحل، این کیفیت دستخوش تغییرات می‌گردد. در ابتدا کیفیت تمام پاسخ‌ها برابر با یک است و به‌مرور، پاسخ‌های ضعیف‌تر دچار افت کیفیت می‌شود.

- تقطیع مبهم

ابهام تقطیع، به معنای ابهام در شیوه جداسازی پیشوندها، میانوند و پسوندهاست. مثالی از این ابهام را می‌توان در جدول مرتبط با کلمه «فحکم» مشاهده نمود (فحکم: ف حکم/ فح کم). در حقیقت، در صورت ابهام در پاسخ‌های ممکن برای یک کلمه، ابهام موجود در تقطیع‌های مختلف آن کلمه، اساسی‌ترین ابهام آن به حساب می‌آید. پس، در این مرحله کلماتی که علاوه بر ابهام در پاسخ‌ها، دارای ابهام تقطیع هم هستند، جدا شده و به مرحله ششم که رفع ابهام تقطیع است، ارجاع داده می‌شود. کلماتی نیز که ابهام آنها در سطح تقطیع نیست، به بخش رفع ابهام ویژگی‌ها ارسال می‌گردد. ویژگی‌هایی که مرتبط با رفع ابهام تقطیع هستند، عبارت‌اند از: Entry، Slice و Seq. این ویژگی‌ها، در بخش رفع ابهام ویژگی‌ها وجود ندارند. در حقیقت، ویژگی‌هایی که مرتبط با قطعه‌بندی کلمه هستند، در بخش رفع ابهام تقطیع می‌شوند و سایر ویژگی‌ها در بخش رفع ابهام ویژگی‌ها، مورد بررسی قرار می‌گیرند.

- رفع ابهام تقطیع

در این بخش، تقطیع‌های ممکن کلمه ورودی به‌ترتیب کیفیت مرتب می‌شوند. این کار، تأثیر بسیاری در رفع ابهام خروجی تحلیلگرهای صرفی دارد. بسیاری از ابهام‌ها در ویژگی‌های مختلف، به علت ابهام در تقطیع است. در حقیقت، با شناسایی تقطیع صحیح کلمه، بسیاری از ویژگی‌های مبهم رفع ابهام می‌گردند. در این بخش، ابتدا شبیه‌ترین کلمات موجود در پیکره رفع ابهام نور برای کلمه مورد پرسش فهرست شده، سپس با استفاده از رده‌بند نزدیک‌ترین همسایه، بهترین تقطیع انتخاب می‌گردد.

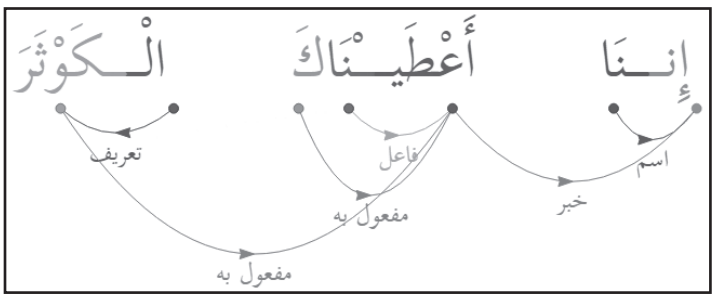
- رفع ابهام ویژگی‌ها

بزرگ‌ترین قسمت الگوریتم پیشنهادی پس از رفع ابهام تقطیع، رفع ابهام ویژگی است. تمامی کلماتی که ابهام دارند، به این مرحله ارسال شده، سپس به خروجی می‌روند. نقش این مرحله، از دو جهت پر رنگ می‌شود؛ ابتدا اینکه با وجود رفع ابهام تقطیع، همچنان ابهام بسیاری در پاسخ‌های منطبق با بهترین تقطیع وجود دارد. از طرفی، اشتباه مرحله رفع ابهام تقطیع، در تشخیص تقطیع برتر می‌تواند در این مرحله و به کمک رفع ابهام ویژگی‌ها خنثا گردد.

همان‌طور که پیش از این اشاره شد، موتور صرف به‌طور میانگین برای هر کلمه، ۵۱ پاسخ متفاوت دارد. آمارهای گرفته‌شده در این تحقیقات نشان می‌دهد که ابهام موجود در پاسخ‌های منطبق بر هر تقطیع نیز بالاست. طبق این آمارها، به ازای هر تقطیع به‌صورت میانگین، ۳۵ پاسخ متفاوت توسط موتور صرف نور تولید می‌گردد. ازاین‌رو، حتی با وجود تشخیص تقطیع صحیح کلمه، همچنان ابهام بالایی در پاسخ‌ها وجود داشته، نیاز به رفع ابهام دقیق‌تر به‌خوبی احساس می‌گردد.

ارزیابی روش پیشنهادی

ارزیابی روش پیشنهادی، روی بخشی از پیکره نور انجام گرفته است. در این روش، ۱۰ درصد این پیکره به‌صورت تصادفی انتخاب گردیده و به عنوان اطلاعات، ارزیابی و سایر داده‌ها به عنوان داده آموزشی مورد استفاده قرار گرفته است. نتایج ارزیابی‌ها نشان می‌دهد که رفع ابهام هوشمند قادر است با دقت ۹۹.۲٪ اسم، فعل یا حرف بودن کلمه را تشخیص دهد. دقت این سیستم در تشخیص ریشه کلمه، ۹۸/۶٪ بوده و پیراسته کلمه (حذف پیشوندها و پسوندها) را با دقت ۹۸٪ به درستی تشخیص می‌دهد.



در انتها، دقت کلی این سیستم، به ازای ترکیب تمام ویژگی‌ها کنار هم، یعنی تشخیص تحلیل کامل کلمات، برابر با ۸۸٪ است.

نتیجه‌گیری

روش ارائه‌شده در این تحقیقات، یکی از برترین روش‌های رفع ابهام هوشمند زبان عربی محسوب می‌شود که با استفاده از داده‌های آموزشی بسیار غنی و فنون یادگیری ماشین، به دقتی بی‌نظیر در رفع‌ابهام هوشمند زبان عربی دست یافته است. این رفع ابهام، لازمه استفاده از موتور صرف زبان عربی به شمار می‌رود. کاربردهایی همچون: ریشه‌یابی و جست‌وجوهای مبتنی بر ریشه‌یابی، موتورهای جست‌وجو و مشابه‌یابی متون، برخی از کاربردهای ویژه رفع ابهام هوشمند هستند. چنین کاربردهای، هم‌اکنون در برخی پروژه‌های مرکز تحقیقات کامپیوتری علوم اسلامی نور مشاهده می‌شوند؛ برای نمونه، نرم‌افزار جامع الأحادیث علاوه بر جست‌وجوی سابق خود، در نسخه جدید از جست‌وجوی پیراسته کلمات نیز بهره می‌برد. جست‌وجوی پیراسته کلمات، به کاربر اجازه می‌دهد با جست‌وجوی یک کلمه، کلماتی را که با حذف و یا اضافه شدن پیشوند یا پسوندی به آن کلمه تولید شده‌اند نیز بیابد.

منابع:

1. M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," *HLT-NAACL 2004 Short Pap.*, pp. 149–152, 2004.

2. E. Mohamed and K. Sandra, "Arabic Part of Speech Tagging," *Evaluation*, pp. 2537–2543.

3. N. Habash and O. Rambow, "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop," *ACL '05 Proc. 43rd Annu. Meet. Assoc. Comput. Linguist.*, pp. 573–580, 2005.

4. A. Pasha, M. Al-badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. M. Roth, "MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," *Proc. 9th Lang. Resour. Eval. Conf.*, pp. 1094–1101, 2014.

5. E. Mohamed, B. Mohit, and K. Oflazer, "Annotating and Learning Morphological Segmentation of Egyptian Colloquial Arabic," *Proc. Eighth Int. Conf. Lang. Resour. Eval.*, pp. 873–877, 2012.

6. N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh, "Morphological analysis and disambiguation for dialectal Arabic," *Hlt-Naacl*, no. June, pp. 426–432, 2013.

7. M. Maamouri and A. Bies, "Developing an Arabic treebank: methods, guidelines, procedures, and tools," *Proc. Work. Comput. Approaches to Arab. Script-based Lang.*, pp. 2–9, 2004.

8. M. Maamouri, A. Biesa, T. Buckwalter, M. Diab, N. Habash, O. Rambow, and D. Tabessi, "Developing and Using a Pilot Dialectal Arabic Treebank," *Proc. Fifth Int. Conf. Lang. Resour. Eval.*, pp. 443–448, 2006.

9. Y. O. Mohamed Elhadj and Y. O. M. Elhadj, "Statistical Part-of-Speech Tagger for Traditional Arabic Texts," *J. Comput. Sci.*, vol. 5, no. 11, pp. 794–800, 2009.

10. B. Benali and F. Jarray, "Genetic Approach for Arabic Part of Speech Tagging," *Int. J. Nat. Lang. Comput.*, vol. 2, no. 3, pp. 1–12, 2013.

11. K. Toutanova, D. Klein, and C. D. Manning, "Feature-rich part-of-speech tagging with a cyclic dependency network," *Proc. 2003 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Vol. 1 (NAACL '03),* pp. 252–259, 2003.

12. C. D. Manning, "Part-of-speech tagging from 97% to 100%: Is it time for some linguistics?," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6608 LNCS, no. PART 1, pp. 171–189, 2011.

13. A. Dastani, B. Minaei-bidgoli, M. R. Vafaei, and H. Juzi, "An Introduction to Noor Diacritized Corpus Diacritization," *Lr. Work.*, no. table 3, pp. 13–17, 2012.■