شبکه واژگانی زبان عربی
با استفاده از فرآیند نیمه‌خودکار در داده‌های علوم اسلامی

حبیب سریانی
مدرّس ادبیات عرب و پژوهشگر مرکز تحقیقات کامپیوتری علوم اسلامی(نور)
Seryani@noornet.net

چکیده
غنی‌سازی محتواهای علوم انسانی و اسلامی، از اهمیت ویژه‌ای برخوردار است. استفاده از شبکه واژگانی، از مهم‌ترین راهبردهای تحقیقاتی در رشته‌های مرتبط با علوم انسانی است. شبکه واژگانی، مجموعه‌ای از کلمات است که به واسطه ارتباطات معنایی به یکدیگر متصل شده و در سطحی جامع‌تر از یک لغتنامه در یادگیری یا استخراج اطلاعات مورد نیاز محققان کاربرد دارد. روند فعالیت در دستیابی خودکار به یک شبکه واژگانی منسجم، در گرو پردازش لفظی و پردازش معنایی بر اساس متن‌کاوی داده‌های عربی در منابع زبانی دقیق و مناسب است. استفاده از فرآیند ماشینی در هریک از این مراحل پردازش و یافتن منابع عربی دقیق علوم اسلامی، با چالش‌های مختلفی روبهرو بوده که در این مقاله، به بررسی برخی جزئیات این طرح، به‌ویژه بر اساس دادگان موجود در مرکز تحقیقات کامپیوتری علوم اسلامی و روش مقابله با چالش‌ها پرداخته شده است.

کلیدواژگان: شبکه واژگانی، وردنت عربی، داده‌کاوی، متن‌کاوی، پردازش زبان‌های طبیعی.

درآمد
وقتی کارگروه‌های اختصاصی در کشف دانش و داده‌کاوی (۱)، در همایشی در سیاتل آمریکا سال ۲۰۰۴ گرد هم آمده و بحث‌هایی را در مورد چگونگی ارتباط و پیوندهای متن‌کاوی (۲) و پردازش زبان‌های طبیعی (۳) مورد بررسی قرار دادند، نتایج در خور توجهی را در مباحث مختلف و بهره‌گیری از نقاط قوت یکدیگر دربرداشت.[۱] اگرچه طرح مباحث شبکه واژگانی (۴) از سالیان قبل در دانشگاه پرینستون نظر اندیشمندان این عرصه را به خود جلب کرده بود، اما ورود محققانِ متن‌کاوی و پردازش هوشمند زبان به این حوزه و استفاده از توانمندی‌های آنان، تلاش‌های قابل توجهی نظیر ارائه کتاب و مقاله و بهره‌برداری در زبان‌های مختلف را به این عرصه معطوف داشت.

داده‌کاوی، در حقیقت، تجزیه و تحلیل داده‌های مشاهده‌ای در جهت یافتن ارتباط‌های پنهان و ارائه خلاصه‌ای از آن داده‌ها در شیوه‌های جدیدی است که هر دو، قابل فهم و کاربرد برای مالکان آن داده‌ها باشد. این ارتباط و خلاصه حاصل از داده‌کاوی، غالباً با یک الگو استخراج می‌گردد؛ مانند استفاده از معادلات خطی، قوانین، خوشه‌بندی، گراف، ساختار درختی و دیگر الگوها.[۲] نه تنها در عرصه علوم اسلامی، بلکه حلّ مشکلات ناشی از تحلیل مقادیر بسیاری از داده‌های علوم دیگر و حتی شرکت‌ها و سازمان‌ها، در گرو استفاده بهینه از علوم داده‌کاوی است. این در حالی است که تنها درصد کمی از داده‌های جمع‌آوری‌شده در سازمان‌ها و شرکت‌ها مورد تحلیل و تجزیه قرار گرفته‌اند و این درصد در مقابل رشد حجم داده‌ها، تصمیم‌گیری و تحلیل نهایی را در مورد محتوای پایگاه داده‌ها غیرممکن ساخته است.[۳] ازاین‌رو، ضروری است که در علوم اسلامی بیش از تمرکز بر تولید داده‌های جدید، با استفاده از کتاب‌ها و میان‌رشته‌های نوظهور، بر داده‌کاوی دقیق و تحلیل و تجزیه داده‌های علوم اسلامی که غالباً ثابت بوده و مهم‌ترین آنها را قرآن، روایات و کتب تاریخی تشکیل می‌دهند، همت گمارد.

بی‌تردید، یک گام مهم در استفاده از پایگاه داده‌ها و استنتاج متون جهت استخراج ارتباطات و اطلاعات، بهره‌گیری از شبکه واژگانی (وردنت) است.[۴] استفاده از شبکه واژگانی، نه تنها در متن‌کاوی و تحلیل داده‌ها، بلکه در یادگیری و آموزش نیز کاربرد فراوان دارد. در واقع، استفاده از دانش واژگان(۵) در یادگیری و خواندن متون زبان‌های خارجی ضروری است.[۵] به طور کلی، معلمان زبان از دو مهارت جهت درک معنای کلمات ناآشنا برای آموزش شاگردان خود استفاده می‌کنند: ۱. مهارت حدسزدن (به عنوان مثال، توانایی استفاده از قرائن متنی و ساختاری برای تحصیل معنای صحیح کلمه، همچون هم‌خانواده‌ها)؛ ۲. توانایی استفاده مناسب از فرهنگ لغات (به عنوان مثال، ارجاع سریع به صفحه‌ای که شامل کلمه مورد نظر بوده و خواندن توضیحات مربوط).[۶]

با پیشرفت و گسترش فرهنگ لغات الکترونیکی و نرم‌افزارهای مبتنی بر داده، زبان‌آموزان نیز با کاهش زمان جست‌وجو دسترسی کاملی به دانش واژگان پیدا کرده‌اند. فراتر از فرهنگ لغت، با استفاده از فرآیند ماشینی، اطلاعات زبانی بیشتری امکان‌پذیر شده است؛ مانند ترجمه ماشینی که در گذشته مهم‌ترین انگیزه طراحی این شبکه واژگانی بوده است، ابهام‌زدایی واژگان (۶)، طبقه‌بندی اسناد، طراحی هستان‌شناسی در وب و همچنین غنی‌سازی محتوای لغت‌نامه‌ها مانند تعداد تکرار کلمات، تنوع استعمال و چگونگی ساختارهای گرامری[۷] و بسیاری دیگر از فواید مرتبط با متن‌کاوی به‌ویژه در علوم اسلامی.[۸]

در دانش واژگان، هر کلمه به واسطه ویژگی‌هایی که دارد، توصیف می‌گردد؛ ویژگی‌هایی شامل: شیوه خواندن و نوشتار، معنا و مفهوم، دستور زبان، استعمال، پیوستگی، هم‌نشینی، فراوانی و کاربرد.[۹] این ویژگی‌ها در لغات عربی با پیچیدگی بیشتری همراه بوده که مباحث عمیق و گسترده صرفی و نحوی و تداخل آن با مباحث فقه اللغه، گویای آن است. ازاین‌رو،

داده‌کاوی، در حقیقت، تجزیه و تحلیل داده‌های مشاهده‌ای در جهت یافتن ارتباط‌های پنهان و ارائه خلاصه‌ای از آن داده‌ها در شیوه‌های جدیدی است که هر دو، قابل فهم و کاربرد برای مالکان آن داده‌ها باشد. این ارتباط و خلاصه حاصل از داده‌کاوی، غالباً با یک الگو استخراج می‌گردد

استفاده از فرآیند ماشینی جهت یادگیری از داده‌های نامتوازن، به عنوان یک چالش جدید در مجموعه‌های یادگیری ماشینی(۷)، داده‌کاوی و متن‌کاوی در زبان عربی بوده است. البته همان‌گونه‌که زبان مادری همواره تأثیری منفی در روند یادگیری واژگان زبان خارجی دارد، اما این نقیصه از زبان فارسی به عربی به سبب نزدیکی این دو زبان و وجود منابع دینی عربی به یک فرصت تبدیل شده و می‌تواند روند یادگیری زبان عربی و دسترسی به متون قرآنی و روایی را برای کاربران سرعت بخشد.

در دانش واژگان، هر کلمه به واسطه ویژگی‌هایی که دارد، توصیف می‌گردد؛ ویژگی‌هایی شامل: شیوه خواندن و نوشتار، معنا و مفهوم، دستور زبان، استعمال، پیوستگی، هم‌نشینی، فراوانی و کاربرد. این ویژگی‌ها در لغات عربی با پیچیدگی بیشتری همراه بوده که مباحث عمیق و گسترده صرفی و نحوی و تداخل آن با مباحث فقه اللغه، گویای آن است

ایزوتسو، زبان‌شناس معاصر، فهم درست و دقیق مفاد و معانی واژه‌ها، به‌ویژه متون دینی همچون قرآن کریم را منوط به قرارگرفتن این کلمات در یک نظام مرتبط معنایی می‌داند. در حقیقت، قرآن کریم یک مجموعه اتفاقی و تصادفی از کلمات بدون نظم و قاعده نبوده؛ بلکه واژه‌های آن با قرارگرفتن در نظام و شبکه خاص قرآنی در ارتباط با سایر کلمات کلیدی، رنگ معناشناختی خاصی به خود می‌گیرند که اگر خارج از این نظام و بدون لحاظ بافت و سیاق آیات در نظر گرفته شوند، هرگز آن معنا را دربرنمی‌گیرند.[۱۰]

در این نوشتار، بر آنیم تا فارغ از مدل‌های ارائه شبکه واژگانی که مرتبط با علوم رایانه‌ای است، این مبحث را تنها از بُعد تحقیقات زبانشناسی در زبان عربی و بر اساس پایگاه داده علوم اسلامی که بستر آن در مرکز تحقیقات کامپیوتری علوم اسلامی فراهم آمده است، بررسی نماییم و روند اجرایی طرح برای دستیابی به یک شبکه واژگانی جامع را مورد تحقیق و مطالعه قرار دهیم.

شبکه واژگانی(۸)

شبکه واژگانی که بر مبنای یافته‌های روان‌شناسی زبان ساخته شده است[۱۱]، تلاشی در جهت بازنمایی چیزی است که در ذهن انسان‌ها از واژه‌ها و روابط آنها وجود دارد. در حقیقت، یک مدل طراحی‌شده از روابط معنایی بین واژگان است تا حداکثر واژه‌های موجود در یک زبان را به صورت شبکه‌ای از روابط در خود قرار دهند.[۱۲] این مبحث در سال‌های اخیر در مباحث پردازش هوشمند زبان‌های طبیعی (NLP) نیز کاملاً پذیرفته شده است. علم وجود یا هستان‌شناسی شبکه واژگان (۹)، اوّلین‌بار در زبان انگلیسی توسط دانشگاه پرینستون (۱۰) و در چهار مقوله: اسم، فعل، صفت و قید برای هر واژه ساخته شد[۱۳] و سپس، با مقوله پنجم تکمیل شد. معنا و مفهوم لغوی، رابطه بین این واژگان قرار گرفت و مجموعه کلماتی را که دارای معنای مشابه بودند، هم‌معنا (۱۱) نامیده شد. بنابراین، در تعریفی کوتاه «وردنت» یا شبکه واژگانی، مجموعه‌ای از واژگان است که به‌واسطه ارتباطات معنایی به یکدیگر متصل شده‌اند. این شبکه به‌مانند یک لغت‌نامه یا اصطلاح‌نامه وسیع و جامع با ساختاری نموداری و گراف مانند است.[۱۴] از آنجا که هر زبانی دارای ویژگی‌های منحصربه‌فرد خود است، ازاین‌رو، نمی‌توان روابط جامعی را برای تمامی زبان‌ها در نظر گرفت. زبان عربی نیز با توجه به قواعد پیچیده صرف و نحو و مباحث عمیقی که در بحث فقه اللغه مطرح می‌گردد، دارای خصوصیاتی در شبکه واژگانی خود است که الگوبرداری کامل از دیگر شبکه‌های زبانی را ناممکن می‌سازد. نوع ساختار صرفی فعل و اسم، تغییر کامل کلمات در حالت‌های اعلالی، وجود جمع‌های مکسر، تغییر ساختار کلمه در ساختن اسم مصغر و منسوب، انصراف و عدم انصراف در برخی اسما و قواعد دیگر، همگی از مواردی هستند که لغات عربی و قواعد مربوط بدان‌ها را از باقی زبان‌ها متمایز می‌سازد. ازاین‌رو، استفاده از رویکرد ساخت بالا به پایین (Top Down Approach) و ایجاد یک شبکه واژگانی مستقل نیازمند است تا خصوصیات زبان عربی به‌خوبی حفظ شده و سپس، با دیگر زبان‌ها مانند وردنت پرینستون یا فارسنت توانایی تلفیق داشته باشد.[۱۵]

اگرچه از سال ۲۰۰۶ برخی از زبان‌شناسان عرب‌زبان قدم‌های مؤثری را در ایجاد یک شبکه واژگانی عربی برداشتند (AWN : Arabic WordNet)[۱۶]، ولی ضرورت طرح مبحث در داده‌های علوم اسلامی، این ضرورت را ایجاب نمود تا بر اساس سیستم‌های هوشمندی که در مرکز نور به کار گرفته شده و نسبتاً به موارد مشابه از امتیازات قابل توجهی برخوردار است، به طراحی یک شبکه واژگان منسجم و دقیق

ضروری است که در علوم اسلامی بیش از تمرکز بر تولید داده‌های جدید، با استفاده از کتاب‌ها و میان‌رشته‌های نوظهور، بر داده‌کاوی دقیق و تحلیل و تجزیه داده‌های علوم اسلامی که غالباً ثابت بوده و مهم‌ترین آنها را قرآن، روایات و کتب تاریخی تشکیل می‌دهند، همت گمارد

در این زمینه اقدام نموده و آن را مقدمه‌ای برای ورود به ایجاد یک پیکره دادگان جامع قرار دهیم.

ارتباطات شبکه‌ای

شبکه واژگانی که با استفاده از یک نوع ارتباط معانی بین واژگان فراهم آمده است، بر پایه مختصات زبانی بوده و چه‌بسا در هر زبان دارای ویژگی‌های مخصوص خود باشد. بااین‌حال، دسته‌ای از روابط در مفاهیم عمومی زبان‌شناسی ثابت بوده و در زبان عربی نیز قابل ساخت است. برخی از این ارتباطات بین لغات و شیوه پیشبرد آنها را با توجه به منابع زبانی موجود در پایگاه داده علوم اسلامی مرکز نور بررسی خواهیم نمود.

- **ترادف(۱۲)**: رابطه بین دو مفهوم معادل یا نزدیک که یکی قابلیت جانشینی دیگری را داشته باشد. ازاین‌رو، رابطه ترادف، رابطه‌ای دو سویه و متقارن است؛ مانند سه واژه: فهم، درک و علم.

- **تضاد(۱۳)**: رابطه بین دو مفهوم مخالف است. با توجه به گستره وسیع مفهوم متضاد که شامل هر سه نوع ارتباط تلازم (مانند مفهوم بدهکار و طلبکار) تتمیم و تکمیل (تناقض) و تضاد منطقی است، باید امکان ایجاد شبکه‌ای برای بیان تمامی متضادها مناسب با زبان عربی آماده شود.

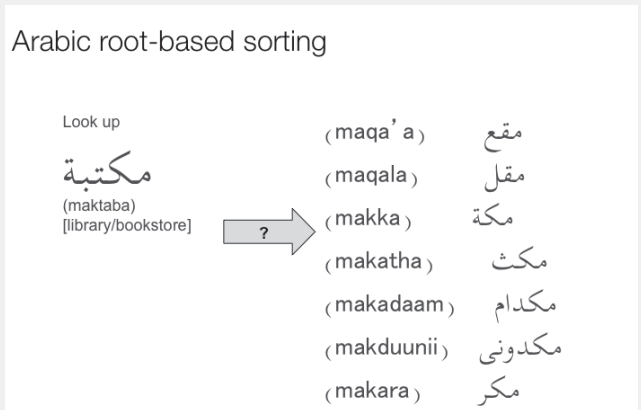
- **شمول/ رابطه کلی(۱٤) و جزئی(۱۵)**: رابطه بین دو مفهوم که یکی اعم از دیگری باشد؛ مانند «الحیوان» و «الفرس» (حیوان و اسب)، «الزهرة» و «الخزامی» (گل و گل لاله). این ارتباط در بازیابی اطلاعات در زمینه‌های دسته‌بندی متون و خوشه‌بندی آنها بسیار مفید خواهد بود؛ به طور مثال، اگر در متنی از کلمات «السیارات» و «الدراجات النارية» (اتومبیل و موتور سیکلت) استفاده شده است، می‌توان آن را در دسته المَرکبات (وسایل نقلیه) قرار داد.

- **جزءواژگی/ رابطه جزء(۱٦) و کل(۱۷)**: رابطه بین دو مفهوم که یکی جزئی از دیگری باشد؛ مانند «الزهرة» و «البتلة» (گل و گلبرگ) یا عضوی از دیگری باشد؛ مانند «الغابات» و «الشجرة» (جنگل و درخت) و یا یک ماده ساخته شده از آن باشد؛ همچون «الزجاج» و «النافذة» (شیشه و پنجره). این رابطه، در تشکیل شبکه واژگانی هسته اصلی علم هستان‌شناسی و مرتبط به آن بوده و می‌تواند در علوم انسانی نقش مهمی ایفا کند.

- **سببیت(۱۸)**: رابطه یک مفهوم با سبب پدیدآورنده آن؛ مانند «الموت» (مرگ) که مسبب «القتل» (کشتن) است.

- **نتیجه(۱۹)**: رابطه بین دو مفهوم که یکی حاصل دیگری باشد؛ مانند آنچه در زبان عربی به عنوان مصدر و حاصل مصدر بین دو مفهوم «الغَسل» و «الغُسل» (شستن و شست‌وشو) برقرار است.

روابط متعدد دیگری نیز در وردنت پرینستون و دیگر شبکه‌های واژگانی مطرح شده است؛ مانند رابطه تضمین(۲۰)، اعتباریت(۲۱)، شِبهیت (۲۲) و نوعیت (۲۳).[۱۷] اما آنچه مورد نظر این نگاشته است، تکیه بر مواردی است که در منابع زبانی عربی به‌راحتی قابل دسترس بوده و بتوان به صورت ماشینی یا نیمه‌خودکار آنها را استخراج نمود.



منابع زبانی

در ساخت یک شبکه واژگانی دقیق که مبتنی بر ارتباطات صحیح باشد، به منابع معتبری نیازمند هستیم که در خود توضیحات روشن و دقیقی از معانی و یا برخی ارتباطات در اختیار مخاطبان خود بگذارد. ازاین‌رو، آنچه به عنوان پایه اطلاعات در مباحث وردنت مطرح است، لغت‌نامه‌ها هستند. تمامی اطلاعات موجود در لغت‌نامه‌ها از اولویت اوّل در پردازش لفظی و معنایی برای هر واژه و عبارت برخوردارند. بیش از ۶۰ عنوان لغت‌نامه در مرکز تحقیقات علوم اسلامی نور تایپ و فرمت‌گذاری شده که با استفاده از فرآیندهای ماشینی متن‌کاوی، استخراج اطلاعات مورد نیاز از آنها با سختی کمتری میسر است.

با توجه به تحقیقات صورت‌گرفته در کتاب‌هایی که به بیان کلمات مترادفات پرداخته‌اند، به نظر می‌رسد، کتاب «المَکنز العربی المعاصر»(۱۹۹۳) به عنوان نمونه کار اوّلیه، گزینه مناسبی است. ترتیب الفبایی و چینش روان کلمات و تقسیم معانی مختلف یک کلمه، از نقاط قوت آن است. چنانچه یک فعل یا اسم در استعمال با حرف خاصی همراه شده باشد، آن حرف را به شکل مجزا بیان کرده است. نکته اصلی در این کتاب، تأخر آن از باقی منابع اصیل و قدیمی است که نگارندگان با دقت در منابع معتبر سعی داشته‌اند نمونه‌ای مناسب را در اختیار دانش‌پژوهان بگذارند.

کتاب دیگری که در حوزه مترادفات لغات عربی نوشته شده و بسیار جامع‌تر از «المکنز» است، «المعجم المفصل فی المترادفات»(۲۰۰۹) است که از مجموعه، کتاب‌های ارزشمند «الخزانة اللغوية» است. از همین مجموعه کتاب‌هایی در کلمات متضاد، کلمات اضداد (کلماتی که دارای دو معنای متضاد هستند) و نکات ظریف لغات عربی نگاشته شده است که می‌تواند در استخراج ماشینی ارتباطات راهگشا باشد.

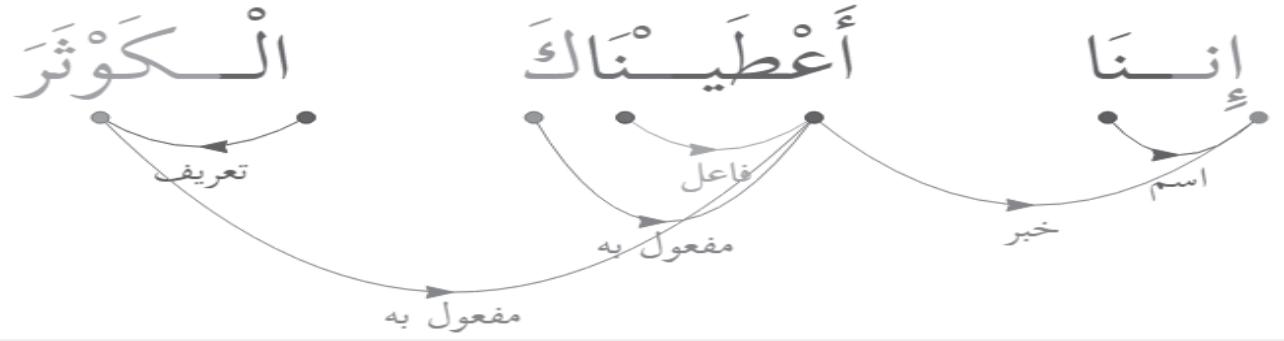
در زمینه ارتباطات شمول یا جزءواژگی و حتی سببیت، کتاب‌هایی همچون «المخصص» نوشته ابن سیده (قرن پنجم) موجود است. المخصص، لغت‌نامه‌ای ۱۷ جلدی است که ترتیب واژه‌های آن برحسب موضوعات و معانی است. کتاب از گسترده‌ترین معاجمی است که بر اساس معنای کلمات و به‌صورت موضوعی تدوین شده است. مؤلف، اوّلین موضوعی که در کتاب خود بعد از مقدمه‌اش ذکر می‌کند، انسان است و سعی کرده در ترتیب موضوعات و به‌تبع آن ترتیب باب‌ها، یک سیر منطقی را رعایت کند. مؤلف، در ذیل هر موضوع، چند باب ذکر می‌کند؛ مثلاً در ذیل موضوع «خلق انسان» بابی به نام «الحمل و الولادة» است که کلمات آن را بر اساس اوّلین روزهای انعقاد نطفه تا ایام بزرگسالی مرتب کرده و شرح داده است. [۱۸] تلخیص دو جلدی از کتاب المخصص نیز با نام «الإفصاح فی فقه اللغة» (۱۳۴۸ه.ق) منتشر گردیده که می‌تواند جهت پرهیز از تطویل کتاب اصلی در استخراج اطلاعات مفید باشد.

استفاده از فرآیند ماشینی جهت یادگیری از داده‌های نامتوازن، به عنوان یک چالش جدید در مجموعه‌های یادگیری ماشینی، داده‌کاوی و متن‌کاوی در زبان عربی بوده است. البته همان‌گونه‌که زبان مادری همواره تأثیری منفی در روند یادگیری واژگان زبان خارجی دارد، اما این نقیصه از زبان فارسی به عربی، به سبب نزدیکی این دو زبان و وجود منابع دینی عربی به یک فرصت تبدیل شده و می‌تواند روند یادگیری زبان عربی و دسترسی به متون قرآنی و روایی را برای کاربران سرعت بخشد

نکته اساسی در مورد زبان عربی و گستره ارتباطات، این است که آیا به طور مثال، در مترادفات، «ترادف در ریشه» ملاک است یا «ترادف در آخرین سلسله از مشتقات» و یا «گزینه‌ای میانی در سلسله مراتب اشتقاق تصریفی» (باب، نوع فعل و...)؟ باید در نظر داشت که در کلمات گوناگون، گستره ترادف متفاوت است. شاید به‌ندرت، دو ریشه را بتوان یافت که با وجود ترادفشان، تمام مشتقات آن دو نیز با هم مترادف باشند؛ مثلاً اگر «جَاهَدَ» را مترادف با «قَاتَلَ» بدانیم، هیچ‌گاه، «اجتهد» مترادف با باب افتعال ریشه «قتل» و یا هر مشتق دیگری از آن نخواهد بود. همچنین است، موارد مشابه آن. گاه یک کلمه، با یک عبارت مترادف است که در آن عبارت، مشتقی از ریشه همان کلمه و یا ریشه‌ای دیگر وجود دارد؛ مثلاً «انتحر» مترادف با «نحر نفسه» و «قتل نفسه» است و به‌هیچ‌وجه با خود «نحر» و «قتل» مترادف نیست. گاه نیز یک کلمه دارای چندین معنا بوده (۲۴) و تنها برخی از آن معانی دارای مترادف با کلمات دسته دیگر هستند. تنها راه تفکیک دقیق ریشه از مشتقات، استفاده از «پردازش لفظی» هوشمند و دقیق است.

اعلام باید از استاندارد مشخصی مانند: سال ولادت، محل و شغل استفاده کرد. سپس، بر اساس این استاندارد، رابط کاربر مناسبی جهت ورود اطلاعات تهیه گردد و در اختیار پژوهشگران قرار گیرد. عدم توجه به این ضرورت، موجب ناهمگونی دادگان و فعالیت‌های پژوهشگران می‌گردد و در نهایت، به عدم ارزش کافی اطلاعات استخراجی می‌انجامد.

۲. استناد به منابع: مهم‌ترین چالش در این بخش، استناددهی به منابعی است که دادگان از آنها جمع‌آوری شده[۱۹] و باید در طول زمان نیز با وجود احتمال تغییرات در منابع، لینک ارتباط بین این دادگان با منابع به شکلی صحیح و کامل برقرار باشد. اینکه فرمت‌ها، اطلاعات اعلام و یا نمایه و موضوعات از



اگر با نگاهی عمیق‌تر شبکه واژگانی دریچه‌ای به سوی دست‌یافتن به یک پیکره دادگان جامع و وسیع باشد، می‌توان شبکه واژگانی را از حوزه ارتباطات بین معانی گسترش داد و در حوزه اعلام و شخصیت‌ها، اماکن و شهرها، کتاب‌ها و نویسندگان و موارد مشابه ترسیم نمود. استفاده از داده‌های نرم‌افزار «درایة النور» و «تراجم ۲» که حاوی کتاب‌های همچون «اعیان الشیعة» نوشته علامه امین و «الذریعة الی تصانیف الشیعة» نوشته علامه طهرانی در حوزه اعلام و شخصیت‌هاست و یا بهره از کتاب «معجم البلدان» در ۷ جلد نوشته حموی (قرن هفتم) در مورد اماکن و شهرها بوده و در نهایت، استفاده از فرمت‌هایی که توسط گروه‌های مختلف پژوهشی در مرکز نور از کتاب‌ها به‌دست‌آمده، مانند نمایه و کلیدواژه‌های موجود در معاجم موضوعی همچون بحار الأنوار، می‌تواند از دیگر منابع زبانی موجود برای ایجاد یک شبکه واژگانی باشد.

چالش‌های پیش رو در این حوزه نیز عبارت‌اند از:

۱. استانداردسازی اطلاعات دادگان: بسیار ضروری است که در جمع‌آوری منابع و دادگان در هر مقطع و موضوعی، به یک استاندارد مطلوب دست یافته و بر اساس نیازها آیین‌نامه‌ای مدون تهیه گردد؛ به‌طورمثال، در جمع‌آوری دادگان مربوط به

کدام کتاب و نرم‌افزار و با کدام نسخه و ویرایش استخراج شده، از جمله این موارد است. علاوه بر دقت پژوهشگران در مستندسازی دادگان، به نظر می‌رسد که حجم اصلی فعالیت، به بخش فنی کار معطوف باشد.

پردازش خودکار

بین الفاظ، دو نوع ارتباط برقرار است: یکی ارتباط لفظی و دیگری ارتباط معنایی.[۲۰] برقراری ارتباط لفظی بین واژگانی که در یک حوزه معنایی قرار دارند، با توجه به پیچیدگی صرفی کلمات در زبان عربی، اهمیتی مضاعف می‌یابد. توجه به پردازش لفظی، فعالیت متن‌کاوی در وردنت عربی را متمایز از زبان‌ها خواهد کرد. در بین تمام اقدام‌های انجام‌شده در حوزه وردنت عربی، فعالیتی را که با دقت در پردازش ماشینی و لفظیِ در پی ایجاد ارتباطات لفظی و نظام

اشتقاق تصریفی باشد، نیافتیم. ازاین‌رو، پیشبرد فعالیت‌های این حوزه در عین نوگرایی، نیازمند تجربه و خطاست که خود چالشی در دستیابی به اهداف مورد نظر است.

در مرکز نور جهت پردازش لفظی و تجزیه و تحلیل ماشینی لغات عربی، از دو سیستم هوشمند «موتور صرف» و «موتور استم‌ساز» استفاده می‌شود. موتور صرف، وظیفه تجزیه صرفی کلمات بر اساس قواعد دستوری زبان عربی را بر عهده دارد.[۲۱] این سیستم هوشمند برچسب‌گذار، بر اساس ریشه کلمات عربی، قابلیت برچسب‌گذاری بسیاری از ویژگی‌های کلمات مانند: نوع کلمه(۲۵)، نوع اشتقاق، تذکیر و تأنیث، افراد و جمع را داراست. در کنار این سیستم، از یک برنامه کمکی دیگر، به‌ویژه در کلماتی که فاقد ریشه هستند، با عنوان موتور «استم‌ساز» استفاده می‌شود. استم در یک کلمه، به‌منزله بن‌مایه و ساقه آن کلمه است که فارغ از تغییرات لفظی و اعلالی آن کلمه لحاظ می‌گردد؛[۲۲] به طور مثال، کلمات «فالتحوّلات» و «بالتحویل»، هر دو از ریشه «حول» هستند؛ ولی دارای استم مشترک نیستند؛ زیرا استم یکی «تحول»، و دیگری «تحویل» است.

در مرکز نور جهت پردازش لفظی و تجزیه و تحلیل ماشینی لغات عربی، از دو سیستم هوشمند «موتور صرف» و «موتور استم‌ساز» استفاده می‌شود. موتور صرف، وظیفه تجزیه صرفی کلمات بر اساس قواعد دستوری زبان عربی را بر عهده دارد

بی‌تردید، نقطه قوت بخش متن‌کاوی مرکز تحقیقات نور، بهره‌گیری از این دو سیستم هوشمند شناسایی لغات عربی است که تا حدودی بسیاری نیل به یک شبکه واژگانی منسجم را فراهم آورده است.

پس از گذر از پالایش‌های عمومی، مانند تصحیح فونت و اغلاط متنی، نوبت به پردازش لفظی با استفاده از موتور صرف و استم‌ساز در این حوزه می‌رسد. فعالیت‌های این بخش عبارت‌اند از:

۱. تعیین ریشه استعمالی برای تمامی کلمات، به‌عنوان مقدمه اتصال به لغت‌نامه‌ها و واژگان دیگر: موتور صرف با تحلیل کلمه بر اساس قواعد زبان عربی، ممکن است چندین ریشه منطقی را اعلام کند؛ به طور مثال، کلمه «یستعد» بر اساس منطق صرفی می‌تواند هم از ریشه «ع د د» و هم ریشه «ع د و» باشد؛ درحالی‌که این کلمه بر استعمال در داده‌های علوم اسلامی، تنها در موارد معدودی از ریشه «عدو» به کار رفته است. همچنین، کلمه «المدح» که منطقاً از دو ریشه «م د ح» و «د ح ح» امکانپذیر است.

۲. استفاده از موتور صرف جهت شناسایی کلمات دارای ریشه و استفاده از موتور استم‌ساز و پیراسته‌ساز جهت شناسایی کلمات فاقد ریشه و مطابقت آن با مداخل استاندارد.

۳. استفاده از نظام اشتقاق و تصریف در ایجاد حلقه‌های ارتباط لفظی بین واژگان: در حقیقت، می‌توان به‌واسطه این نظام، افعال، اسما و دیگر مشتقات از یک ریشه را با یک نظام ترتیبی به یکدیگر متصل نمود.

نظر به اقدام‌های انجام‌شده در مطالعات ریشه و مشتق و توانایی حال حاضرِ گروه متن‌کاوی، به نظر می‌رسد تعیین مداخل استانَدارد[۲۳]، از بیشترین اولویت برخوردار است. ضرورت ایجاد این مداخل به همراه تمامی مراحل پردازش و استخراج آنها، نیازمند نگاشته‌ای مجزاست. مداخل استاندارد در زبان عربی را این‌گونه می‌توان تعریف نمود: کلماتی هستند که ساختار صرفی آن، دارای موضوعیت معنایی در لغت‌نامه‌های عربی است. این مداخل، آن دسته از کلماتی هستند که در معاجم لغت برای آنها در ذیل یک ریشه یا یک موضوع، توصیف و یا توضیحی بیان شده است؛ به طور مثال، کلمه «ضارب» معنایی متفاوت از فعل خود «ضرب» نداشته و در شبکه واژگانی، یک مفهومی مستقل از ضرب نیست؛ اما کلمه «مجتهد» که در مورد افراد خاصی در یک سطح علمی به یک اصطلاح تبدیل شده است، کاملاً مفهومی متفاوت از فعل ماضی «اجتهد» که تنها به معنای تلاش و کوشش است، داراست.

اما پردازش معنایی کلمات، در حقیقت، فرایندی جهت اتصال کلمات و ایجاد یک پیکره از دادگان مختلف است. استفاده از برچسب‌های مختلف در شناسایی معنای یک کلمه، می‌تواند آن را در حلقه‌های ارتباطی مختلف از کلمات قرار دهد. در مباحث وردنت، اوّلین و اصلی‌ترین ارتباط معنایی کلمات، بین مترادفات است[۲۴]. کلمه‌ای را که محور قرار گرفته و باقی مترادفات و کلمات با آن حلقه ارتباطی تشکیل می‌دهند، به‌عنوان Lemma یاد می‌شود و حلقه‌های ارتباطی پیرامون یک کلمه را هم‌معنا یا synset می‌نامند.

اوّلین پردازش معنایی نیز توسط موتور صرف صورت می‌پذیرد که با اِعلام نوع کلمات مانند انواع اسماء یا افعال و همچنین ارائه برخی تگ‌ها، اطلاعاتی همچون تأنیث و تذکیر، افراد و جمع، تعریف و تنکیر و نوع باب را ارائه می‌دهد. تگ‌های اِعلامی در ابزار ریشه و مشتق (ابزاری که جهت تفکیک ریشه منطقی و استعمالی تهیه شده است)، مانند تگ اَعلام و تگ معرب و دخیل نیز از دیگر مواردی است که قابل استفاده است. حال با توانایی‌های موجود و اولویت‌های متن‌کاوی، می‌توان فعالیت‌های ذیل را برشمرد:

۱. اضافه‌کردن مترادفات و متضادات به مداخل استاندارد: با توجه به وجود دو کتاب «المعجم المفصل فی المترادفات» و «المعجم المفصل فی المتضادات» و تایپ و نشانه‌گذاری آن دو، بهترین شروع در حوزه پردازش معنایی، مرتبط‌کردن اطلاعات این دو کتاب به مداخل استاندارد است. این دو کتاب، به‌مانند «فرهنگ ابجدی» علاوه بر صیغه ۱ ماضی، مواردی از اسما را که دارای وجه تمایزی در معنا نسبت به افعال خود بوده‌اند، به شکل مجزا آورده‌اند. ازاین‌رو، ابتدا با استفاده از موتور صرف باید در ذیل یک ریشه قرار بگیرند و سپس، با استفاده از نظام اشتقاق، حلقه‌های ارتباطی تشکیل دهند و در نهایت، به مداخل استاندارد اضافه شوند. پیش‌فرض این است که مواردی از مداخل این دو کتاب که در مداخل استاندارد موجود نیستند، باید به مداخل استاندارد اضافه شوند؛ مگر آنکه توسط محقق به‌عنوان مدخل شناسایی نشوند و به افعال خود ملحق شوند.

۲. روش دیگر استفاده از دادگان، معاجم موضوعی مانند نمایه و موضوعات است. در معاجم موضوعی، هریک از نمایه و موضوعات به یک یا چند کلیدواژه متصل شده‌اند. چالش اصلی کار در این پروژه، تفاوت بین کلیدواژه‌ها و مداخل استاندارد است. ازاین‌رو، همان فرایند پردازش لفظی، یعنی استفاده از موتور صرف و موتور استم‌ساز (در کلماتی که موتور صرف پاسخی ندارد) و استفاده از نظام اشتقاق در کلیدواژه‌ها ضروری است. ساماندهی و یکسان‌سازی کلیدواژه‌ها، از اوّلین اولویت‌های این پروژه در این حوزه است.

با نظر به امکانات نرم‌افزاری و ورود به عرصه الکترونیک، بررسی کلمات عربی بر اساس تحلیل ماشینی حروف، چندان دور از دسترس نخواهد بود. اگر شبکه واژگانی به‌گونه‌ای ترسیم شود که کلماتی که دارای اشتقاق کبیر و اکبر هستند، رابطه شبکه‌ای و معنایی با دیگر کلمات داشته باشند، چه‌بسا بتوان چنین نتیجه گرفت که در زبان عربی، نه‌تنها کلمات، بلکه حروف نیز دارای ماهیتی استقلالی بوده و هر حرف در بارِ معنایی کلمات و هم‌معناها دخیل است؛ مطلبی که در هیچ‌یک از دیگر زبان‌ها به آن اشاره‌ای نشده است

شبکه واژگانی بر اساس تحلیل ماشینی حروف

در مباحث دستوری زبان عربی، مبحثی به عنوان «اشتقاق» مطرح است که برخی از اندیشمندان این عرصه آن را به سه نوع: صغیر، کبیر و اکبر تقسیم نموده‌اند. [۲۵] نوع اوّل، مانند اشتقاق ضارِب و مضروب از مصدر «الضرب» که در آن، ترتیب حروف اصلی و ریشه کاملاً رعایت شده است. نوع دوم، مانند کلماتی که دارای همان حروف هستند؛ ولی ترتیب رعایت نشده است؛ مانند «کلم» (به معنای جرح) و «لکم» (به معنای ضربت با دست) که دارای معنایی نزدیک به یکدیگرند و همچنین است کلماتی مانند «کمل» و «ملک». نوع سوم، اشتقاقی است که برخی حروف اصلی تغییر نموده و با حروف هجایی مشابه جایگزین شده است؛ مانند دو کلمه «فصم» و «قصم» که جامع معنای آن دو، شکستن و قطع‌کردن است و همچنین، کلمات «ثلب و ثلم» و «فطر و فطم».

با نظر به امکانات نرم‌افزاری و ورود به عرصه الکترونیک، بررسی کلمات عربی بر اساس تحلیل ماشینی حروف، چندان دور از دسترس نخواهد بود. اگر شبکه واژگانی به‌گونه‌ای ترسیم شود که کلماتی که دارای اشتقاق کبیر و اکبر هستند، رابطه شبکه‌ای و معنایی با دیگر کلمات داشته باشند، چه‌بسا بتوان چنین نتیجه گرفت که در زبان عربی، نه‌تنها کلمات، بلکه حروف نیز دارای ماهیتی استقلالی بوده و هر حرف در بارِ معنایی کلمات و هم‌معناها دخیل است؛ مطلبی که در هیچ‌یک از دیگر زبان‌ها به آن اشاره‌ای نشده است. به‌هرحال، این امر به‌عنوان یک موضوع پیشنهادی جهت ارائه مقالات و پایان‌نامه به دانش‌پژوهان عرصه زبان‌شناسی و پردازش طبیعی زبان پیشنهاد می‌گردد.

پی‌نوشت‌ها:

1. Knowledge Discovery in Database (KDD) and Data Mining (DM).

کشف دانش و داده‌کاوی، دانشی میان‌رشته‌ای با تمرکز بر روش‌های استخراج سودمند اطلاعات از میان داده‌هاست. در حال حاضر، رشد سریع داده‌های آنلاین و استفاده گسترده از پایگاه داده‌ها، نیاز شدیدی را متوجه روش‌های داده‌کاوی نموده است.

2. Text Mining (TM).

3. NLP.

4. Wordnet.

5. Vocabulary Knowledge.

6. Word Sense Disambiguation (WSD).

7. Machine Learning (ML).

۸. شبکه واژگانی که با نام تجاری «وردنت» برای دانشگاه پرینستون به ثبت رسیده است، برای زبان‌های دیگر گاهی با نام «فارس‌نت» و «یورونت» یاد شده و گاهی نیز مقید به زبان می‌شود؛ مانند «وردنت عربی» (AWN) یا «وردنت اروپایی» (EWD).

9. Wordnet Ontology.

10. Princeton University.

11. Synset: Synonym Sets.

12. Synonymy.

13. Antonymy.

14. Hyperonymy.

15. Hyponymy.

16. Metonymy.

17. Meronymy.

18. Causality.

19. Derived from.

20. Implication.

21. Value.

22. Similar to.

23. Troponym.

24. Polysemy.

25. POS Tagging.

منابع:

1. Anne Kao, Stephen R.potee, "Natural Language Processing and Text Mining", Springer-Verlag London Limited 2007, Library of Congress Control Number: 2006927721, P IV.

2. David Hand, Heikki Mannila and Padhraic Smyth, 2001, "Principles of Data Mining", A Bradford Book The MIT Press, Cambridge, Massachusetts London England, P6.

3. M. Rajman, R. Besancon, 1998. "Text Mining: Natural Language techniques and Text Mining applications", Artificial Intelligence Laboratory, Computer Science Department, Swiss Federal Institute of Technology, Switzerland, Published by Chapman & Hall, P 3.

4. Anne Kao, P148.

5. Chen, C.M. & Hsu, S.H. (2008). Personalized intelligent mobile learning system for supporting effective English learning. Educa-

tional Technology & Society, 11(3), 153-180.

6. Sun, K.T. Huang, Y.M. & Liu, M.C. (2011). A WordNet-Based Near-Synonyms and Similar-Looking Word Learning System. Educational Technology & Society, 14 (1), 121–134.

۷. ر.ک: روحی‌زاده و همکاران. «طراحی شبکه واژگانی افعال زبان فارسی». مجموعه مقالات دانشگاه علامه طباطبایی، هفتمین همایش زبان‌شناسی ایران. ۲۲۰. (۱۳۸۶): ۵۱۸ ـ ۵۳۰.

۸. ر.ک: عابدینی، حسین و بهروز مینایی. «کاربردهای داده‌کاوی در علوم اسلامی». ره‌آورد نور، سال دهم. ۳۴. (بهار ۱۳۹۰): ۷ ـ ۱۲.

9. Richards, J. C. (1976). The role of vocabulary teaching. TESOL Quarterly, 10(1), 77-89.

۱۰. ایزوتسو، ۱۳۶۰، به نقل از: عچرش، خیریه و کهندل جهرمی مرضیه. «رسم شبکه معنایی واژه فتنه و مترادفات آن در قرآن». فصلنامه مطالعات قرآنی. ۲۴. (زمستان ۱۳۹۴): ۳۱.

11. Miller G. "Nouns in WordNet: A Lexical Inheritance System", International Journal of Lexicography, vol. 3, no. 4, 1990.

۱۲. حسابی، اکبر. «مقایسه روابط معنایی درون‌زبانی اسلامی در فارسنت، یورونت و وردنت پرینستون». جستارهای زبانی. ۳۲. (مهر و آبان ۱۳۹۵): ۱۴۹ ـ ۱۷۴.

13. WordNet. A lexical database for English, Available online at: http://wordnet.princeton.edu/wordnet/relatedprojects

14. WordNet is a database of English words that are linked together by their semantic relationships. It is like a supercharged dictionary/thesaurus with a graph structure: http://stevenloria.com/tutorial-wordnet-textblob

۱۵. ر.ک: شمس فرد، مهرنوش و سمیه باقربیگی. «روشی نوین در ساخت نیمه‌خودکار شبکه واژگانی افعال زبان فارسی»، مجله دستور ویژه‌نامه فرهنگستان. ۷. (اسفند ۱۳۹۰): ۱۰۸ ـ ۱۶۱.

16. Lahsen Abouenour, Karim Bouzoubaa, Paolo Rosso, On the evaluation and improvement of Arabic WordNet coverage and usability, Lang Resources & Evaluation (2013) 47:891-917, DOI 10.1007/s10579-013-9237-0.

17. See: Zakaria Elberrichi and others, 2008, "Using WordNet for Text Categorization", The International Arab Journal of Information Technology, Vol. 5, No. 1, January 2008, PP 16-24.

۱۸. برگرفته از: بخش کتاب‌شناسی نرم‌افزار «قاموس النور ۲» از محصولات مرکز تحقیقات کامپیوتری علوم اسلامی (نور).

19. When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly cite the source. Citation figures are critical to WordNet funding: https://wordnet.princeton.edu

20. Two kinds of relations are represented by pointers: lexical and semantic. Lexical relations hold between semantically related word forms; semantic relations hold between word meanings: https://wordnet.princeton.edu/man/wninput.5WN.html

۲۱. ر.ک: سریانی، حبیب و بهروز مینایی. «سیستم هوشمند برچسب‌گذار ادات سخن زبان عربی؛ لایه صرف». ره‌آورد نور. سال دهم. ۳۴. (بهار ۱۳۹۰): ۱۸ ـ ۲۸.

22. Sun, K.T. Huang, Y.M. & Liu, M.C. (2011). A WordNet-Based Near-Synonyms and Similar-Looking Word Learning System. Educational Technology & Society, 14 (1), PP 121–134.

23. An IndexWord is a single word and part of speech. An IndexWord can be used to lookup a Synset object. Once you have an IndexWord, you can lookup all the Synset objects associated with that word.

24. The main relation among words in WordNet is synonymy, as between the words shut and close or car and automobile. Synonyms-words that denote the same concept and are interchangeable in many contexts-are grouped into unordered sets (synsets): https://wordnet.princeton.edu

۲۵. السید علی خان. ۱۳۸۹. **الحدائق الندیة فی شرح الفوائد الصمدیة**. ج ۱، ص ۱۷۴. قم: انتشارات هجرت. ■