پیکره متون زبان طبیعی

(طراحی، ساخت و مدیریت)

حمیده اسدی*

دانشجوی دکتری علم اطلاعات و دانش‌شناسی- بازیابی اطلاعات دانشگاه تهران؛ تهران، ایران.

نادر نقشینه

دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛ دانشگاه تهران، تهران، ایران.

ملوک‌السادات حسینی بهشتی

دکتری زبان‌شناسی – همگانی؛ دانشیار؛ پژوهشگاه علوم و فناوری طلاعات (ایرانداک)؛ تهران، ایران.

دریافت: ۱۴۰۲/۰۴/۰۷ | پذیرش: ۱۴۰۲/۰۹/۰۴ | مقاله برای اصلاح به مدت ۴ روز نزد پدیدآوران بوده است.

چکیده: با توجه به نقش پیکره‌ها در حوزه‌های مطالعاتی گوناگون و لزوم ساخت یک پیکره عمومی برای افزایش کارایی و اثربخشی در پردازش‌هایی که مستلزم بهره‌جویی و استفاده از متن زبان طبیعی است، هدف این مطالعه، تمرکز بر طراحی و ساخت خودکار پیکره متون زبان طبیعی و نرم‌افزاری برای مدیریت آن است.

در این پژوهش، از روش مبتنی بر فناوری برای ساخت پیکره تک زبانه و به زبان فارسی استفاده شده است. این پیکره به صورت خودکار و با گردآوری داده‌های وبی تولیده شده و منابع آن را متون خبری مندرج در خبرگزاری‌های فارسی زبان تشکیل داده است.

در این مطالعه، پیکره‌ای از متون زبان طبیعی به زبان فارسی ساخته شده است. با توجه به خودکار بودن فرایند ساخت پیکره، نرم‌افزاری برای مدیریت آن چه در مرحله ساخت و چه در مرحله استخراج اطلاعات نیاز است که در این مطالعه طراحی، ساخته و پیاده‌سازی شده است.

ساخت پیکره‌ای عمومی از متون زبان طبیعی برای اهداف پژوهشی گوناگون کاربرد دارد و روش پیشنهادی و استفاده از ابزارهای معرفی شده در این مطالعه می‌تواند ساخت پیکره را تسهیل کند. همچنین طراحی نرم‌افزاری برای مدیریت پیکره، صرفه‌جویی در زمان و هزینه ساخت را به همراه خواهد داشت و امکان استخراج اطلاعات از آن را فراهم خواهد

نشریهٔ علمی (رتبه ین المللی)
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۸۲۲۳-۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، SCOPUS و ISC
http://jipm.irandoc.ac.ir
دوره xx | شمارهٔ x | ص ص xx-xx
۱۳xx x

نوع مقاله: پژوهشی

به این مقاله به شکل زیر استناد کنید:

درون متن:

(اسدی ،نقشینه، حسینی بهشتی زودآیند)

در فهرست منابع:

اسدی، حمیده. نقشینه، نادر. حسینی بهشتی، ملوک السادات. زودآیند. پیکره متون زبان طبیعی: طراحی، ساخت و مدیریت. *پژوهشنامه پردازش و مدیریت اطلاعات*.

http://Jipm.irandoc.ac.ir (دسترسی در روز/ماه/سال)

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات
X xx ۱۳ | دورهٔ xx | شمارهٔ x

کرد.

کلیدواژه‌ها: پیکره، دادگان، پردازش زبان طبیعی، زبانشناسی پیکره‌ای، هوش مصنوعی

*حمیده اسدی asadi1366@gmail.com

۱. مقدمه

برای انجام وظایف مختلف پردازش زبان طبیعی، حجم زیادی از داده‌های متنی نیاز است که بتواند با استفاده از هوش مصنوعی تقلید درک و دریافت انسانی را بهبود بخشد و استفاده از آن، فناوری را به توانایی انسانی نزدیک سازد (بحرانی و دیگران ۱۳۸۶؛ ورما[۱] و خندلول[۲] ۲۰۱۹). بدین منظور مهمترین کار، تهیه و ساخت پیکره‌ای است که به صورت خودکار انجام می‌شود.

پیکره‌ها را مجموعه‌ای از داده‌های متنی سازمان یافته می‌دانند که می‌تواند حاوی متن، نقل‌قول، فهرست و حتی لغات باشد که برای اهداف گوناگون تهیه می‌شوند (دشتبانی و دیگران ۱۳۹۱).

در ابتدا پیکره‌های متنی سنتی برای پژوهش‌های زبانشناسی با استفاده از منابع چاپی مانند مقالات روزنامه‌ای و کتاب‌ها ساخته می‌شد. با رشد شبکه جهانی وب به عنوان یک منبع اطلاعاتی، استفاده از داده‌های آن برای وظایف گوناگون در پردازش زبان طبیعی افزایش یافت و مزایایی برای ساخت پیکره از داده‌های وبی نسبت به متن چاپی وجود دارد:

۱. داده‌های وبی به شکل الکترونیک و برای رایانه‌های قابل خواندن است در حالیکه همه داده‌های چاپی به شکل الکترونیک در دسترس نیست.
۲. حجم داده‌های وبی زیاد است و برای آموزش داده تخمین بهتری را رقم می‌زند.
۳. گردآوری داده‌های وبی با استفاده از موتورهای جستجو انجام می‌شود و نیازی به دانلود گرانقیمت/پر هزینه محاسباتی ندارد (لیو[۳] و سوران[۱] ۲۰۰۶).

---

۱. Verma, Parul

۲. Khandelwal, Brijesh

۳. Liu, Vinci

در واقع، ظهور رایانه/فناوری‌های رایانه‌ای به خلق آنچه امروز پیکره نامیده می‌شود، کمک فراوانی نمود و زبانشناسی پیکره‌ای به عنوان رویکردی عملی در پژوهش‌های زبانشناسی مورد توجه بسیاری از پژوهشگران قرار گرفته و سبب شده تا پیکره‌های گوناگونی با اهداف متنوع تهیه شود اما هدف مشترک آنها، مطالعات علمی بر روی زبان طبیعی است که گاه در یک حوزه خاص انجام می‌شود (دشتبانی و دیگران ۱۳۹۱؛ صفری ۱۳۹۴؛ بِنت[۲] ۲۰۱۰).

هر چند قابلیت انجام و کارایی مطالعات پیکره‌ای به روش و حجم پیکره وابسته بوده و بیشتر این مطالعات هم در سطح پیکره‌های کوچک طراحی شده و انجام گرفته است اما پیکره‌هایی که در مقیاس بزرگ تهیه می‌شوند علاوه بر اینکه در پیشبرد پژوهش‌های زبانشناسی موثر هستند برای توسعه نظام/سامانه‌های پردازش زبان طبیعی هم کارایی دارند و اساساً نیز برای آزمون روش‌های گوناگون، پیکره‌های متنی با حجم زیاد نیاز است (دشتبانی و دیگران ۱۳۹۱؛ ذوالفقار و دیگران ۱۳۹۹؛ ثابتی[۳] و دیگران ۲۰۱۸).

امروزه به کارگیری و توجه به پیکره‌ها و زبانشناسی پیکره‌ای برای تحلیل منابع در حوزه‌های مختلف و به ویژه مطالعات زبانشناختی، پژوهش‌های زبانی را اعتبار دیگری بخشیده است و مزایای استفاده از پیکره‌ها در تحلیل‌ها و استدلال‌های زبانی همچون: بهره‌گیری از حجم زیاد داده‌ها، گردآوری نظام‌مند و صرفه‌جویی در زمان سبب شده که با کمک آنها بتوان اطلاعات نهفته در متون را شناسایی و استخراج کرد و اهداف پژوهشی گوناگونی را دنبال نمود (عاصی و قندی ۱۳۹۴؛ میرزایی و صفری ۱۳۹۴؛ صفری ۱۳۹۵؛ قدردوست نخچی و دیگران ۱۳۹۵).

مبنای تجربی پیکره‌ها، سبب پدید آمدن بنیاد و بستری برای شناسایی الگوهای زبانی و چگونگی استفاده از آنها فراهم می‌آورد و بررسی توزیعی پیکره‌ها نشان می‌دهد که ویژگی‌های آوایی، واژگانی، دستوری، گفتمانی یا کاربردشناختی و معنایی زبان چگونه است (رضایی پناه و شوکتی مقرب ۱۳۹۵).

در مجموع می‌توان گفت که پیکره‌های زبانی با هدف استفاده در یادگیری ماشینی برای ترکیب توانمندی‌های انسانی و سرعت پردازش اطلاعات ماشینی در راستای دستیابی به لایه‌های

---

۱. Curran, James R.

۲. Bennett, Gena R

۳. Sabeti, Behnam

مختلف زبانی تهیه می‌شوند و با فراهم آوردن امکان استخراج، پردازش و دسته‌بندی اطلاعات، کاربردهای گوناگونی دارند. با این حال می‌توان گفت که هدف نهایی استفاده از آن برای دستیابی به یک مفهوم است. یعنی از آشکارترین لایه زبانی شروع کرده و ادامه می‌دهد تا در انتها به معنا و مفهوم پنهان در متن دست یابد. پیکره‌ها به راستی نقطه آزمون نظام‌ها و روش‌های ابهام‌زدایی و از ابزارهای ضروری این حوزه بوده و ظهور اینترنت و پیشرفت‌های رایانه‌ای بر توانایی‌های این حوزه افزوده و نتیجه آن را نیز بهبود بخشیده است (دفتری‌نژاد ۱۳۸۵؛ میرزایی و مولودی ۱۳۹۳؛ عاصی و قندی ۱۳۹۴).

با این حال باید یادآوری کرد که گردآوری داده و ساخت پیکره به تنهایی چندان ارزشمند نیست و موفقیت استفاده از آن، در به کارگیری و پیشرفت انواع ابزارها و روش‌های پردازش زبان طبیعی، یادگیری ماشینی، یادگیری آماری و یادگیری عمیق است (کامیابی گل و دیگران ۱۳۹۷؛ لی[۱] و دیگران، ۲۰۱۹).

۲. پیشینه پژوهش

در این قسمت، مطالعات مرتبط با پژوهش حاضر ارائه می‌شوند. برای ارائه پیشینه پژوهش از الگوی پیشنهادی نظری (۱۳۹۲) استفاده شده است. بر اساس این الگو پیشینه پژوهش یا «نقشه پژوهش» با استفاده از دو لنز «موضوعی» و «روش‌شناختی» -که محصول دریافت‌های پژوهشگر از مبانی نظری موضوع هستند- تحلیل و دسته‌بندی می‌شوند. محصول مطالعه پیشینه پژوهش با این رویکرد ترسیم گسست دانشی است که پژوهشگر بناست در پژوهش خود آن را پر نماید. بدین منظور مطالعات پیشین از دو منظر موضوعی و روش‌شناختی تحلیل و ارائه می‌شوند.

۱. تحلیل پیشینه از منظر موضوعی

پژوهش‌های بسیاری درباره پیکره‌ها انجام شده است که به بررسی ساخت انواع پیکره‌های زبانی پرداخته یا با ساخت انواع پیکره‌های زبانی اهداف دیگری را مورد تحقیق و ارزیابی قرار داده است. وجه اشتراک هر دو دسته از پژوهش‌ها، توجه به معیارها و شاخص‌هایی برای ساخت پیکره بوده که آنها را از دیدگاه‌های مختلف می‌توان بررسی کرد. از میان آنها آنچه با موضوع این پژوهش مرتبط است را می‌توان در چهار گروه کلی دسته‌بندی کرد (نمودار ۱):

---

۱. Li, Qin

گروه اول، پیکره‌هایی هستند که داده‌های خود را در اشکال متنی از جمله نوشتاری (ذوالفقار و دیگران، ۱۳۹۹؛ افراشی و دیگران، ۱۳۹۴؛ میرزایی و مولودی، ۱۳۹۳؛ دشتبانی و دیگران، ۱۳۹۱؛ پوستیوسکی و دیگران، ۱۹۹۳) و الکترونیک (علایی ابوذر و دیگران، ۱۴۰۰؛ سلامی و دیگران، ۱۳۹۴؛ نظارات و موسوی میانگاه، ۱۳۹۰؛ میهالسی و دیگران، ۲۰۰۶؛ ثابتی و دیگران، ۲۰۱۸؛ سوکولوا[۱] و بابیسو[۲]، ۲۰۱۸) انتخاب کرده است.

گروه دوم، پیکره‌هایی هستند که بر اساس نوع متن، می تواند به کتاب (ذوالفقار و دیگران، ۱۳۹۹؛ میرزایی و مولودی، ۱۳۹۳؛ دشتبانی و دیگران، ۱۳۹۱؛ پوستیوسکی و دیگران، ۱۹۹۳)، مجله (دشتبانی و دیگران، ۱۳۹۱)، مقاله (علایی ابوذر و دیگران، ۱۴۰۰؛ کامیابی گل و دیگران، ۱۳۹۷؛ سلامی و دیگران، ۱۳۹۴؛ دشتبانی و دیگران، ۱۳۹۱؛ ثابتی و دیگران، ۲۰۱۸؛ پوستیوسکی و دیگران، ۱۹۹۳)، متون وبی (افراشی و دیگران، ۱۳۹۴؛ دشتبانی و دیگران، ۱۳۹۱؛ نظارات و موسوی میانگاه، ۱۳۹۰؛ لیو و سوران، ۲۰۰۶) و داده‌های احساسی (سوکولوا و بابیسو، ۲۰۱۸) تقسیم‌بندی شود.

گروه سوم، پیکره‌هایی هستند که حوزه متون آنها علمی (علایی ابوذر و دیگران، ۱۴۰۰؛ ذوالفقار و دیگران، ۱۳۹۹؛ کامیابی گل و دیگران، ۱۳۹۷؛ افراشی و دیگران، ۱۳۹۴؛ سلامی و دیگران، ۱۳۹۴؛ دشتبانی و دیگران، ۱۳۹۱؛ ثابتی و دیگران، ۲۰۱۸) یا عمومی (میرزایی و مولودی، ۱۳۹۳؛ دشتبانی و دیگران، ۱۳۹۱؛ نظارات و موسوی میانگاه، ۱۳۹۰؛ سوکولوا و بابیسو، ۲۰۱۸؛ لیو و سوران، ۲۰۰۶) است.

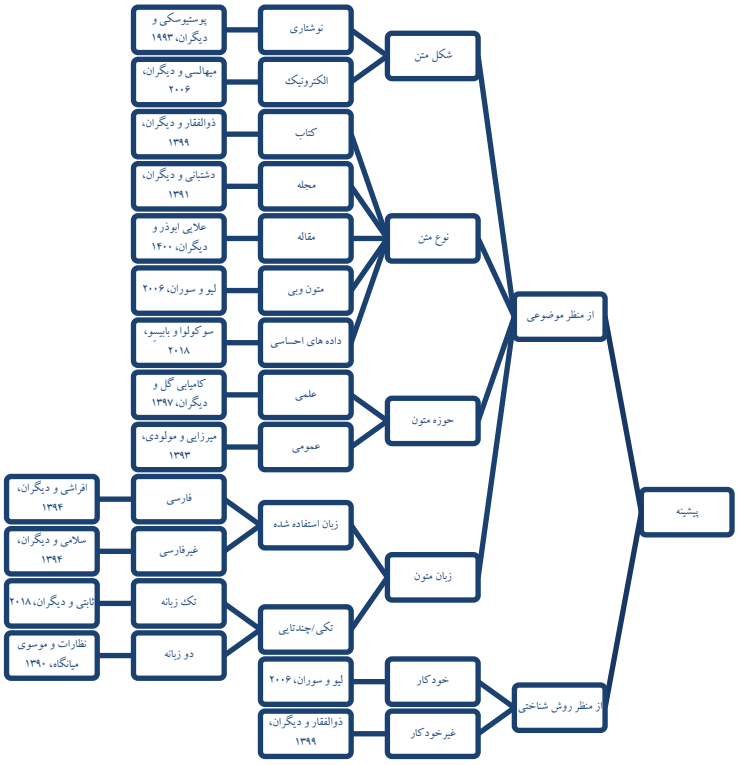
گروه چهارم، پیکره‌هایی هستند که زبان متون آن، فارسی (علایی ابوذر و دیگران، ۱۴۰۰؛ ذوالفقار و دیگران، ۱۳۹۹؛ کامیابی گل و دیگران، ۱۳۹۷؛ افراشی و دیگران، ۱۳۹۴؛ میرزایی و مولودی، ۱۳۹۳؛ دشتبانی و دیگران، ۱۳۹۱؛ ثابتی و دیگران، ۲۰۱۸؛ سوکولوا و بابیسو، ۲۰۱۸) و غیرفارسی (سلامی و دیگران، ۱۳۹۴) بوده و همچنین از لحاظ تک زبانه (کامیابی گل و دیگران، ۱۳۹۷؛ افراشی و دیگران، ۱۳۹۴؛ میرزایی و مولودی، ۱۳۹۳؛ ثابتی و دیگران، ۲۰۱۸؛ سوکولوا و بابیسو، ۲۰۱۸) و دو زبانه (نظارات و موسوی میانگاه، ۱۳۹۰؛ پوستیوسکی و دیگران، ۱۹۹۳) بودن قابل تفکیک می‌باشد.

---

۱. Sokolova, Marina

۲. Bobicev, Victoria

در ادامه این پژوهش‌ها معرفی و دستاورهای آنها ارائه می‌شوند.



نمودار ۱. نقشه پژوهش حاضر

۲. تحلیل پیشینه از منظر روش‌شناختی

در پژوهش‌های بررسی شده، شیوه ساخت پیکره‌ها به صراحت بیان نشده و از شواهد موجود در پژوهش چنین برمی‌آید که پیکره‌ها به صورت خودکار (لیو و سوران، ۲۰۰۶) و یا غیرخودکار (ذوالفقار و دیگران، ۱۳۹۹) تهیه شده است.

به طور کلی از تحلیل مطالعات پیشین اینطور برمی‌آید که در پژوهش‌های گوناگون بسته به هدف آن، یک یا چند معیار درنظر گرفته و بدان اشاره شده است؛ اما در این پژوهش و برای ساخت پیکره، از متون خبری خبرگزاری‌های برخط فارسی زبان به زبان طبیعی، بهره گرفته شده است. شایان توجه است، بر اساس این معیارها، قابلیت به روزرسانی پیکره وجود دارد و از نکات قابل تامل آن به شمار می‌رود.

۳. روش‌شناسی

هدف این پژوهش، پیشنهادی روشی خودکار برای تهیه یک پیکره زبانی از متون زبان طبیعی است. بر این اساس از روش پژوهش مبتنی بر فناوری استفاده می‌شود. در این پژوهش‌ها، استفاده از تجهیزات فناورانه و به خدمت گرفتن کارکنانی که زمینه و سابقه فنی دارند، در اولویت است و بهره‌گیری از آن برای سهولت در انجام آن است. با این حال، مساله‌ای که در این روش وجود دارد، منطبق ساختن روش علمی با فعالیت‌هایی است که شباهت بیشتری به تولید محصول دارند تا پژوهش بنیادی (پاول، ۱۳۸۹).

۱-۳. جامعه پژوهش/داده‌ها

این پیکره متنی، مجموعه‌ای از متون زبان طبیعی زبان فارسی است که از خبرگزاری‌ها جمع‌آوری می‌شود و مورد استفاده قرار می‌گیرد. این پیکره متنی، از نوع تک زبانه و پویا می‌باشد که قابلیت روزآمدسازی نیز دارد.

امروزه متون خبری در جامعه اهمیت بسیاری یافته و جایگاه برجسته آن نزد آحاد افراد جامعه و نیز مسئولان، مدیران، سیاستگذاران و... قابل انکار نیست چرا که:

- با توجه به سرعت و حجم بالای انتشار و میزان تاثیرگذاری آن، بررسی‌های جامع و دقیق از این منابع می‌تواند اطلاعات کاربردی از جامعه در اختیار قرار دهد (مظاهری و دل‌آرا ۱۳۹۸).
- در حوزه متن، خبرگزاری‌هایی که در تولید خبر نقش دارند، معمولا به سبک روزنامه‌ای خاص این کار را انجام داده و بازنشر می‌دهند و علاوه بر آنکه نوع خاصی از اطلاعات به شمار می‌روند، از نظر موضوع، مضامین مرتبط، مالکیت، سرقت علمی و سایر فعالیت‌ها نیز قابل بررسی هستند (کیلگاریف[۱] و گرفنستت[۲] ۲۰۰۳).

---

[۱]. Kilgarriff, Adam

[۲]. Grefenstette, Gregory

- از آنجا که تنوع بیشتری در متن داده‌ها وجود دارد و متن داده از بازه زمانی مختلف انتخاب می‌شوند؛ بدین ترتیب رفتار این اسناد به اطلاعات دنیای واقعی نزدیک‌تر می‌شود (شهشهانی و دیگران ۱۳۹۸).
- از سوی دیگر، امروزه، خبرگزاری‌ها تنها مسئولیت نشر و اطلاع‌رسانی خبر را ندارند بلکه یک خبرگزاری برخط بایستی انواع نیازهای کاربران خود را پاسخ دهد. از جمله آنها می‌توان به دسته‌بندی و سازماندهی درست متون خبری، بازیابی متون خبری با استفاده از کلیدواژه‌های صحیح، سازماندهی و امکان دستیابی به سوابق خبری و... اشاره کرد. این فعالیت‌ها علاوه بر اینکه رضایت کاربران خبرگزاری‌ها را افزایش می‌دهد، باقی ماندن آن را در عرصه رقابت تضمین می‌کند (رباطی ۱۳۹۳).

لازمه بهره‌گیری از روش‌های یادگیری عمیق، حجم زیاد داده‌های متنی، زمان و منبع کافی، برای آموزش و استخراج مدل است و هرقدر داده‌ها حجیم‌تر باشد، تخمین بدست آمده از مدل نیز بهتر خواهد بود. همچنین در بررسی‌های پیکره‌ای هر حوزه دانشی، توجه به انواع پایگاه‌های اطلاع‌رسانی و تولید دانش ضروری است و نیز تنوع مواد اطلاعاتی که محتوای پیکره را تشکیل می‌دهد، نیاز به ارزش‌گذاری و دفاع از صحت داده‌ها دارد (بحرانی و دیگران ۱۳۸۶، شهشهانی و دیگران ۱۳۹۸، روحانیان و دیگران ۱۳۹۹).

خبرگزاری‌ها با توجه به ویژگی‌هایی که پیشتر اشاره شد- به عنوان یکی از بسترهای اصلی تولید دانش زبان فارسی و به شکل زبان طبیعی به عنوان داده این پیکره انتخاب می‌شود و می‌تواند در تحقق پژوهش‌های گوناگون بر اساس اهداف، روش و... نقش بسیار مهمی ایفا کند.

۲-۳. ساخت پیکره و نرم‌افزار مدیریت آن

پیکره‌ای که به صورت خودکار ساخته می‌شود، نیازمند نرم‌افزاری است که بتوان به کمک آن، پیکره را مدیریت کرد: عموماً نرم‌افزاری که برای مدیریت پیکره طراحی می‌شود هم در مرحله ساخت پیکره و هم در مرحله استخراج اطلاعات از پیکره کاربرد دارد و استفاده از پیکره را تسهیل می‌نماید.

نرم‌افزار مدیریت پیکره موسوم به پیکره نما[۱]، برای مدیریت پیکره هم در دو مرحله ساخت و استخراج اطلاعات طراحی، ساخته و پیاده‌سازی شده است. هدف از طراحی این نرم‌افزار مدیریتی، ذخیره‌سازی، پردازش و جستجوی اطلاعات با سرعتی قابل تحمل است که به سبب حجم زیاد اطلاعات گردآوری شده در رایانه‌های خانگی، بسیار وقت‌گیر خواهد بود.

۱-۲-۳. مشخصات فنی نرم‌افزار

این نرم‌افزار در بستر نرم‌افزاری دات نت[۲] محصول شرکت مایکروسافت[۳]، نسخه ۴/۵/۲ و با استفاده از زبان برنامه‌نویسی سی[۴] و در محیط استودیو ویژوال ۲۰۱۷[۵] پیاده سازی شده و در سیستم عامل ویندوز[۶] قابل استفاده و اجرا است. این نسخه از نرم‌افزار به صورت ۶۴ بیتی[۷] همگردانی[۸] شده و با توجه به وجود متن برنامه، این قابلیت را داراست که برای رایانه‌های ۳۲ بیتی هم بازطراحی شود. بدین ترتیب نرم‌افزار تولیدشده قابلیت اجرا در بسیاری از رایانه‌های متداول امروزی را دارد. شایان ذکر است که با توجه به حجم داده‌ها، استفاده از نرم‌افزار روی رایانه‌هایی با حافظه کم و پردازشگر ضعیف با کندی اجتناب‌ناپذیری مواجه خواهد شد.

بانک اطلاعاتی پیکره، در بستر اِس کیو اِل لایت[۹] طراحی و پیاده‌سازی شده است. اِس کیو اِل لایت، یک سامانه مدیریت پایگاه داده کم حجم و قابل جابجایی است که می‌تواند بانک‌های اطلاعاتی رابطه‌ای را تولید و مدیریت کند. ویژگی ممتاز آن اینست که می‌تواند داده‌های حجیم با ساختار ساده را با کارایی و سرعت زیاد پردازش کند. قابلیت تلفیق با برنامه‌های اجرایی را دارد و برای استفاده از آن نیاز به نصب نرم‌افزار مستقلی ندارد.

---

۱. Corpus Viewer
۲. .Net
۳. Microsoft
۴. C
۵. Visual Studio 2017
۶. Windows
۷. bit
۸. Compile
۹. SQL Lite۰

با تخمین حجم زیاد و ساختار ساده داده‌های پیکره متون زبان طبیعی، اِس کیو اِل لایت، می‌تواند بستر مناسبی برای ذخیره‌سازی و پردازش داده‌های پیکره باشد و با توجه به متن باز بودن بستر اِس کیو اِل لایت، امکان استفاده یا انتقال داده‌ها به هر نرم‌افزار دیگر وجود خواهد داشت.

۲-۲-۳. جستجو و ذخیره داده‌ها

اولین مرحله ساخت پیکره، گردآوری داده است که از طریق خزش اینترنتی و موتورهای جستجو انجام می‌شود. در این روش، گردآوری داده‌ها بر اساس موضوع یا پرسش خاص کاربر صورت می‌پذیرد و مزیت آن اینست که قابل گسترش بوده و می‌تواند اطلاعات جاری/بروز را نیز بازیابی کند (بِنت ۲۰۱۰).

برای ساخت پیکره متون زبان طبیعی، خبرگزاری‌های فارسی‌زبان به سبب دسترس‌پذیری و خوانش‌پذیری انتخاب شد. این خبرگزاری‌ها دامنه گسترده‌ای از مطالب و مقالات خبری را شامل می‌شود که برگرفته از میلیون‌ها صفحه وب خبرگزاری‌هاست.

برای جستجوی این صفحات خزشگر/ربوت اینترنتی کوفکس کاپو[۱] انتخاب شد. کوفکس کاپو، یک ربوت نرم‌افزاری هوشمند و پلتفرم/پایگاه یکپارچه است که با طراحی یک استودیوی بصری، داده‌ها را به شکل هوشمند گردآوری و یکپارچه می‌کند (کوفکس ۲۰۱۶؛ کوفکس ۲۰۱۷).

پلتفرم کوفکس کاپو، با روشی سریع‌تر و کارآمدتر، دسترسی به داده‌های ساختاریافته و ساختارنیافته یک برنامه کاربردی یا منبع داده مجازی مانند پایگاه داده‌ها، نظام داده‌ای و ایمیل، وبگاه‌ها، درگاه‌ها سامانه‌های نرم‌افزاری، سامانه‌های کسب و کار، برنامه‌های رومیزی و دیگر منابع داده‌ای را فراهم می‌آورد و با پشتیبانی از انواع برنامه‌های مبتنی بر ویندوز، جاوا اسکریپ[۲] و آژاکس[۳]، داده‌ها را در قالب اکسل[۴]، ایکس اِم اِل[۵]، ایکس اِل اِس[۱]، پی دی اف[۲]، آر اِس اِس[۳] و

---

۱. Kofax Kapow

۲. JavaScript

۳. AJAX

۴. Excel

۵. XML

ای پی آی[۴] و... استخراج می‌کند و داده‌های استخراج شده را بدون کد ترکیب و یکپارچه می‌سازد (کوفکس ۲۰۱۶؛ کوفکس ۲۰۱۷).

البته این امکان وجود داشت که داده‌ها را با کمک سرویس‌های آر اِس اِس که بیشتر خبرگزاری‌ها ارائه می‌کند نیز استخراج و ذخیره کرد اما این ربوت خودکارسازی فرایندها، یک ظرفیت هوشمند دیجیتال است که در کنار نیروی انسانی، با غلبه بر مهمترین چالش‌های اطلاعاتی یعنی پراکندگی اطلاعات در یک نظام‌ها و سامانه‌های اطلاعاتی، کارایی بیشتر و بهتری ایجاد می‌کند و با انجام تمام وظایف پردازش اطلاعات و در نتیجه آن گردآوری و یکپارچه‌سازی داده‌ها، در چند ثانیه و به صورت خودکار در هزینه، زمان و تلاش صرفه‌جویی بسیاری کرده و عملکرد و دسترس‌پذیری را بهینه می‌سازد (کوفکس ۲۰۱۶؛ کوفکس ۲۰۱۷).

با توجه یو آر اِل‌های[۵] معتبر، کوفکس کاپو، محتوای صفحات را بررسی و محتوای آن را با استفاده از عملگرهایی پردازش می‌کند. مهمترین مراحل فرایند خزش بدین شرح است:

۱. صفحات اِچ تی اِم اِل[۶] واکشی شده و لینک ورودی به این صفحات برای خزش استخراج می‌شود
۲. محتوای صفحات بازیابی شده برای استخراج اطلاعات مورد نیاز تجزیه می‌شود
۳. در آخرین مرحله حوزه‌های/اطلاعات استخراج شده مرتبط با هر صفحه در یک پایگاه داده نمایه (ذخیره) شده و مجموعه جدیدی از یو آر اِل برای ادامه خزش انتخاب می‌شود.

در طی فرایند خزش این امکان وجود دارد که صفحات خاص با محتوای یکسان هم بازیابی شود. این بدلیل ابهام در دسته‌بندی‌ها و کلیدواژه‌های جستجو است. برای حذف محتوای تکراری

---

۱. XLS
۲. PDF
۳. RSS
۴. APIs
۵. URL
۶. HTML

در طول فرایند، فیلترهایی اعمال می‌شود. در نهایت تمام اطلاعات استخراج شده در قالب یک پایگاه داده ذخیره شده و برای اعمال پیش‌پردازش‌های گوناگون آماده می‌شود.

بازه زمانی برای گردآوری داده‌ها سال ۹۸-۱۳۹۷ است. انتهای بازه بر اساس حجم داده‌هایی است که بتوان با استفاده از رایانه‌های شخصی قابل پردازش باشد وگرنه و در صورت وجود رایانه‌هایی با قابلیت‌های بالاتر، امکان ذخیره حجم بیشتری از داده‌ها و بروزرسانی پیکره وجود دارد.

۳-۲-۳. ساختار پایگاه داده پیکره

ساختار پایگاه داده به گونه‌ای طراحی شده که در عین سادگی، امکان گزارش‌گیری سریع از داده‌ها را میسر سازد. ساختار ساده داده‌ها همچنین موجب می‌شود که اطلاعات موجود در پایگاه داده قابل استفاده در سایر پردازش‌ها و نرم‌افزارها نیز باشد.

پایگاه داده نرم‌افزار پیکره، از سه جدول تشکیل شده است. این امکان وجود دارد که تمامی این جداول در یک پایگاه داده واحد ذخیره شوند. اما با هدف افزایش سرعت پردازش وگزارش‌گیری، هر یک از جداول در قالب یک پایگاه داده و در یک فایل مستقل ذخیره‌سازی شده است.

جدول ۱. جداول پایگاه داده پیکره متون زبان طبیعی و مشخصات هر جدول

جداول	محتوا	فایل‌های کتابخانه‌ای/داده‌ای	نام فیلد	کاربرد	نوع
واژه‌ها	واژه‌های پیکره زبانی	Words.db	Rowid	کلید اصلی	Integer
			Word	واژه	رشته یونیکد با طول متغیر حداکثر ۵۰ کاراکتر
			Ferq	تعداد کل تکرار واژه در متون	Integer
محتوا	صفحات داده‌های متنی	Content.db	Rowid	کلید اصلی	Integer
			Title	عنوان سند	رشته یونیکد با طول متغیر نامحدود
			Link	آدرس یکتای اینترنتی	رشته یونیکد با طول متغیر نامحدود
			FaDate	تاریخ انتشار سند	رشته یونیکد با طول متغیر حداکثر ۲۰ کاراکتر
			Content	متن سند	رشته یونیکد با طول متغیر نامحدود
ارجاعات	ارجاع واژه‌ها به صفحات	Refs.db	Rowid	کلید اصلی	Integer
			WordID	شناسه واژه در جدول Words	Integer
			RefBlock	یک بلاک از ارجاعات	Blob

جدول واژه‌ها: این جدول شامل واژه‌های بکار رفته در تمامی متون پیکره است. همچنین تعداد تکرار یا بسامد واژه‌ها نیز در این جدول ذخیره می‌شود.

جدول محتوا : این جدول شامل عناوین محتویات اسناد متنی پیکره، آدرس یکتای اینترنتی، تاریخ انتشار سند و متن محتوای سند است. این جدول بیشترین حجم داده پیکره زبانی را در خود جای داده است.

جدول ارجاعات : در این جدول، ارجاعات واژه‌ها به متن اسناد پیکره نگهداری می‌شود. به عبارت ساده‌تر در این جدول بیان می‌شود که یک واژه در چه اسنادی تکرار شده است.

مشخصات فیلدهای این پایگاه داده و جدول‌های آن در جدول ۱ آمده است.

با توجه به تعداد زیاد ارجاعات و اسناد مورد استفاده قرار گرفته در این نرم‌افزار، از ساختار خاصی (فیلد رِف بلاک[۱]) برای ذخیره‌سازی ارجاعات استفاده شده است که در ادامه این ساختار شرح داده خواهد شد.

---

۱. RefBlock

فیلد رِف بلاک شامل بلوکی از داده‌هاست که ارجاعات به یک واژه در متن‌های مختلف را نشان می‌دهد. از آنجایی که تعداد ارجاعات به یک واژه می‌تواند در مجموعه بزرگی از اسناد اتفاق بیافتد، در پیاده‌سازی این طرح از یک ساختار رشته‌ای برای ذخیره‌سازی ارجاعات استفاده شده است. استفاده از این ساختار دارای مزایای زیر است:

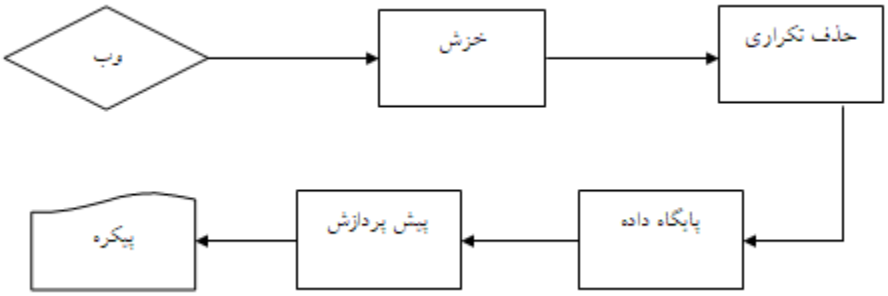
- تعداد رکوردهای پایگاه داده ارجاعات کمتر شده و سر بار[۱] حافظه مصرفی به ازای هر ارجاع به یک واژه کاهش می‌یابد.
- با واکشی هر رکورد از پایگاه داده ارجاعات، تعداد بیشتری از ارجاعات قابل پردازش است. در طراحی پایگاه داده ارجاعات، در هر رِف بلاک تعداد ۱۰۰۰ ارجاع ذخیره می‌شود.

۴-۲-۳. پیش‌پردازش داده‌ها

برای اینکه اطلاعات استخراج شده، قابلیت قرار گرفتن در پایگاه داده پیکره را پیدا کند و برای پردازش اصلی آماده شوند، با کمک نرم‌افزار مدیریت پیکره، پیش‌پردازش می‌شوند. از آنجا که داده‌های این پژوهش از وب‌سایت خبرگزاری‌ها گردآوری شده، ممکن است برخی اطلاعات غیرمرتبط را نیز دربرداشته باشد و چون برای پردازش، تنها متن خبر کارآمد است، در مرحله پیش‌پردازش این موارد شناسایی و حذف می‌شود. تا جملات درست و با مفهوم برای پردازش اصلی تهیه و انتخاب گردد. همچنین با کمک الگوریتم و توابعی، سایر پیش‌پردازش‌ها از جمله نرمال‌سازی، توکن‌بندی، حذف علایم نگارشی و ایست واژگان را انجام می‌شود. بدین ترتیب و با کمک این نرم‌افزار، ساخت پیکره به صورت خودکار و زیر نظر متخصص انجام می‌شود و داده‌ها برای استفاده و تجزیه و تحلیل در پژوهش‌های گوناگون، آماده می‌شود.

---

۱. Overload



نمودار ۲. فرایند ساخت پیکره

از آنجا که پیکره به صورت خودکار تهیه و برای گردآوری داده‌های آن از خزشگر/ربوت اینترنتی استفاده شده، و همچنین برخی متون این پیکره بسیار مفصل و برخی دیگر بسیار موجز و مختصر است؛ از اینرو تنها تعداد کل واژه‌ها/تعداد واژه‌ها- سند به عنوان یک کل در نظر گرفته شده است.

پیش‌فرض الگوریتم تولید رِف بلاک و ساختارهای داده‌ای مورد نیاز، اینست که متن‌های مورد نظر از منابع استخراج و توسط هر تابع جداکننده[۱] به رشته‌ای از کلمات تبدیل شده باشند. برای استخراج متن‌ها می‌توان از روش‌های مختلفی استفاده کرد.

الگوریتم‌های جداکننده متعددی در محیط‌های برنامه‌نویسی وجود دارند. در این طرح نوعی از این الگوریتم برای جداسازی واژه‌ها استفاده شده است.

ساختار داده‌ای ایست واژگان[۲] شامل کلمات و افعال رابطه‌ای و توصیفی است که در تولید و پردازش پیکره زبانی نقشی ندارند زیرا به دلیل استفاده متواتر در متون، پردازش و اعمال آنها در پیکره زبانی موجب افزایش حجم بانک‌های اطلاعاتی و همچنین تاثیر بر آمار سایر واژه‌های زبانی می‌شود. به منظور کاهش بار پایگاه داده، این کلمات در پایگاه داده واژگان ثبت نمی‌شود و تنها در پایگاه اصلی محتوا و متن اسناد قابل مشاهده هستند. ایست واژگانی که در این طرح از بانک اطلاعاتی واژگان حذف شده‌اند، شامل موارد زیر است:

---

۱. Tokenizer

۲. StopWords

جدول ۲. فهرست ایست‌واژگان

آن	اگر	بودم	دارد	شدن	ما	می کنند	هستند
آمد	با	بودند	دارند	شده	من	می گردد	همه
آنرا	باشد	بگردد	داریم	شما	مورد	مگر	همین
آنها	باشند	تو	داشت	شود	می‌باشیم	نباید	هنوز
آنچه	باشیم	تا	داشتن	شوند	می کند	ندارد	و
آیا	باشید	تواند	داشتند	طی	می کنند	نشده	وی
از	باید	حتی	در	که	میداد	نشدم	یا
است	بایست	خواهد	در این	کرد	میدارد	نمی	پس
اکنون	بر	خواهند	دهد	کرده	میشود	نمودیم	چه
اما	برای	خواهیم	دهیم	کردند	میگردد	نمیشدم	چون
اند	بشود	خود	را	کرده اند	می‌برد	نمی‌شدم	چونانکه
او	به	خورد	رسد	کرده‌اند	می‌خواهد	نیز	چونکه
ای	بلکه	داد	رسید	کند	می‌داد	هم	گردد
این	بهتر	دادن	زیرا	کنند	می‌شود	هر	گردید
ایشان	بود	دادند	سپس	کنیم	می کردند	ها	گرفت
اینرا	بودن	داده	شد	لذا	می کنند	های	گشت
گفت	گیرد	گیرند					

در پایگاه داده پیکره، هر واژه‌ای دارای تعدادی ارجاع است که شماره سند و آدرس محل‌های استفاده از آن واژه را در سند دربرمی‌گیرد. ارجاعات به صورت بسته‌هایی از ۱۰۰۰ ارجاع در بانک اطلاعاتی ارجاعات و در فیلد رِف بلاک ذخیره شده‌اند. ساختار داده مورد استفاده مطابق با شکل زیر تعریف و استفاده شده است:

جدول ۳. ساختار داده واژگان پیکره

شماره سند (۳۲ بیت)	تعداد آدرس‌ها (۱۶ بیت)	آدرس ۱	آدرس ۲	...
شماره سند (۳۲ بیت)	**تعداد آدرس‌ها (۱۶ بیت)**	**آدرس ۱**	**آدرس ۲**	**...**
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

اگر تعداد ارجاعات یک واژه در یک رِف بلاک از ۱۰۰۰ آدرس بیشتر شود، یک رکورد رِف بلاک جدید برای ان واژه ایجاد خواهد شد. از آنجایی که نوع داده‌ای رِف بلاک در پایگاه داده از نوع بلاب[۱] (داده‌های بزرگ باینری) تعریف شده است، برای ذخیره‌سازی ارجاعات در این قالب لازم است که ارجاعات، مطابق ساختار فوق در یک نوع داده‌ای با نام مموری استریم[۲] ذخیره‌سازی شوند. بنابراین الگوریتم تبدیل باید به گونه‌ای عمل کند که ابتدا داده‌های مربوط به یک واژه را از یک سند استخراج کند و آن را تبدیل به نوع داده‌ای مموری استریم کند. در صورتی که تعداد ارجاعات از حد مشخصی بیشتر شده (مثلاً ۱۰۰۰ آدرس) مموری استریم به دو شی شکسته می‌شود و مموری استریم جدید در یک رکورد جدید ذخیره می‌شود.

برای تعریف ارجاعات به یک واژه در یک سند، کلاسی از با نام ارجاعات تعریف شده است.

نیو آی دی[۳]، شناسه سند و آدرس[۴]، نشانی (آفست یا فاصله به تعداد کاراکتر)های واژه مورد نظر در آن سند است که در یک لیست ذخیره شده‌اند، برای مجموعه اسناد پردازش‌شده از کلاسی با نام جدول واژه[۵] استفاده می‌شود.

در این کلاس فهرستی از واژه‌ها وجود دارد که به ازای هر واژه مجموعه‌ای از کلاس‌های ارجاعات (شماره سند به همراه آدرس‌های تکرار واژه در سند) تولید خواهد شد.

با فرض اینکه مجموعه کلمات یک سند توسط الگوریتم جداکننده‌ای مانند سپریت ترم[۶] تجزیه شده و به صورت یک لیست ساده، مثلاً آرایه‌ای از کلمات تی اِل[۷] باز گردنداند می‌شود.

دو تابع مهم دیگر توابع تبدیل ساختار داده‌ای از نوع جدول واژه به نوع مموری استریم است. از آنجایی که نوع داده‌ها مموری استریم مستقیماً قابل ذخیره‌سازی در فیلدهایی از نوع بلاب در

---

۱. Blob

۲. Memory Stream

۳. NewID

٤ Addrs

٥. WordTable

٦. Separate Term

۷. TL

بانک اطلاعاتی است، این توابع می‌تواند ارتباط کامل بانک اطلاعاتی و الگوریتم‌های کدنویسی به زبان سی را برای ارجاعات پایگاه داده پیکره زبانی، برقرار سازد.

تابع شمارشگر مموری استریم نیز یکی از توابع مهم و مفید کتابخانه نرم‌افزار پیکره است. به کمک این تابع می‌توان تعداد ارجاعات موجود در یک شی مموری استریم که بر پایه ساختار رِف بلاک ساخته شده باشد را مشخص کرد.

روش خواندن داده‌ها از پایگاه داده نرم‌افزار پیکره در قالب الگوریتمی اعمال می‌شود. فرض بر آنست که داده‌ها، با توجه به گزارش‌گیری با زبان اِس کیو اِل در شی رِفز ریدر[۱] به برنامه بازگردانده شده‌اند. این شی دربرگیرنده تمام رکوردهای جدول ارجاعات مرتبط با یک واژه است.

همچنین فرض دیگر آن است که داده‌های واکشی شده از پایگاه داده باید در شی رفز دیتا[۲] از نوع جدول داده[۳] ذخیره شوند تا قابل پردازش در هر برنامه به زبان سی باشند.

به کمک مجموعه الگوریتم‌های شرح داده شده می‌توان اطلاعات پایگاه داده پیکره را توسعه داد یا داده‌های موجود در آن را بازیابی نمود. بدیهی است که سایر الگوریتم‌های مورد نیاز با توجه به کاربرد نرم‌افزارهای استفاده‌کننده از داده‌های پیکره زبانی طراحی می‌گردند.

۵-۲-۳. جستجو و استخراج اطلاعات از پیکره

با توجه به تخمینی که در مورد حجم این منابع برآورد می‌شود، نیاز به ابزاری برای مدیریت اطلاعات آن پیکره وجود دارد که نخستین آن، طراحی یک موتور جستجو است که بتواند مجموعه حجیم متون را در کل پایگاه داده، جستجو و اطلاعات لازم را سریع‌تر و آسان‌تر از پیکره استخراج کند و استفاده از آن با کمک رایانه‌های معمول امکان‌پذیر باشد. نرم‌افزار طراحی شده ناظر پیکره، افزون بر جستجوها، قابلیت‌های دیگری خواهد داشت که استفاده از پیکره را تسهیل کند. از جمله آنها می‌توان به ارائه فهرست واژگان پیکره و بسامد آنها، ارجاعات مربوط به صفحات منبع داده‌ها متنی شامل آدرس اینترنتی و تاریخ انتشار سند و... اشاره کرد.

---

۱. Refs Reader

۲. RefsData

۳. Data Table

۶-۲-۳. نحوه اجرا و فایل‌های مهم نرم‌افزار پیکره نما

نرم‌افزار از مسیر [Corpus Viewer [bin Release فایل Corpus Viewer.exe اجرا می‌شود. برای این منظور، تمامی فایل‌های کتابخانه‌ای که در این مسیر قرار دارند، مورد استفاده قرار می‌گیرند. بنابراین وجود تمامی فایل‌های این مسیر برای اجرای صحیح نرم‌افزار ضروی است.

چنانچه پیشتر هم اشاره شد (نگاه کنید به جدول ۱)، داده‌های نرم‌افزار در فایل‌های زیر ذخیره شده‌اند:

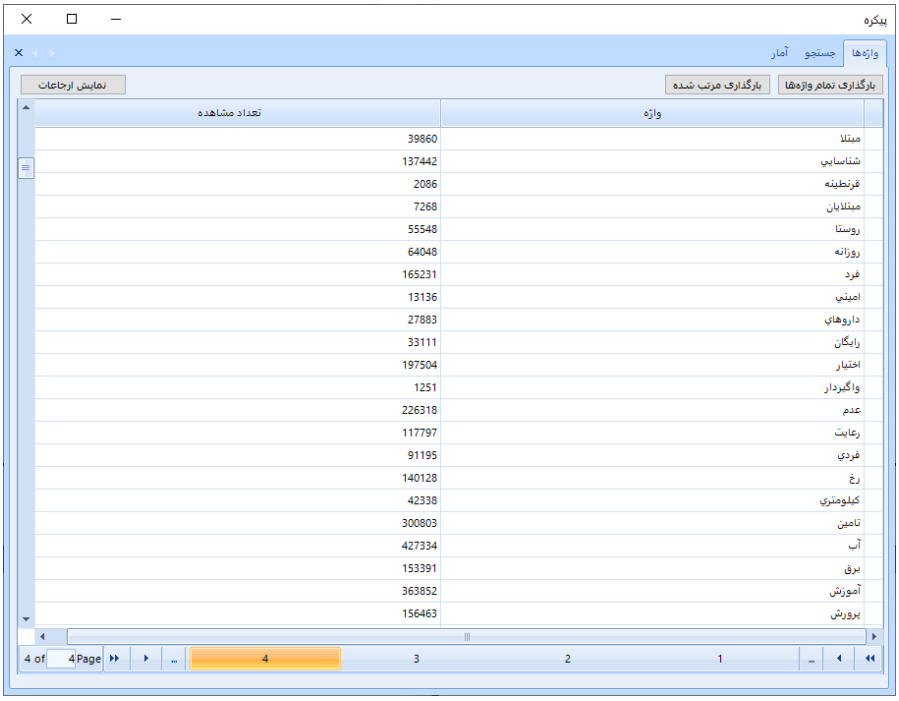
۱. جدول واژگان
۲. جدول محتوا
۳. جدول ارجاعات

تمامی فایل‌های داده‌ای با ساختار پایگاه داده اِس کیو اِل لایت، نسخه ۳ سازگار هستند. برای بازخوانی یا هر گونه اعمال تغییر در داده‌های پایگاه‌های داده می‌توان از نرم‌افزارهای ویرایش پایگاه داده اِس کیو اِل لایت استفاده نمود. برای این پژوهش از نرم‌افزار حرفه‌ای اِس کیو لایت[۱] نسخه۵/۳ استفاده شده است.

۷-۲-۳. شرح محیط نرم‌افزار

در طراحی نرم‌افزار سعی شده که از پیچیدگی‌های کاربری اجتناب گردد و داده‌ها و جستجو بر روی آنها به راحتی در اختیار کاربرد قرار گیرد. پنجره ورودی نرم‌افزار به شکل زیر است:

۱. SQLite Expert Professional



شکل ۱ . پنجره ورودی نرم‌افزار

پنجره اصلی نرم‌افزار دارای سه برگه: «واژه‌ها»، «جستجو» و «آمار» است.

برگه واژه‌ها: در برگه واژه‌ها، جدولی از واژه‌های موجود در پایگاه داده پیکره نمایش داده می‌شود. این جدول دارای دو ستون است. ستون اول به خود واژه اختصاص دارد و ستون دوم بسامد تکرار واژه در مجموعه متن‌های پیکره نمایش داده شده است.

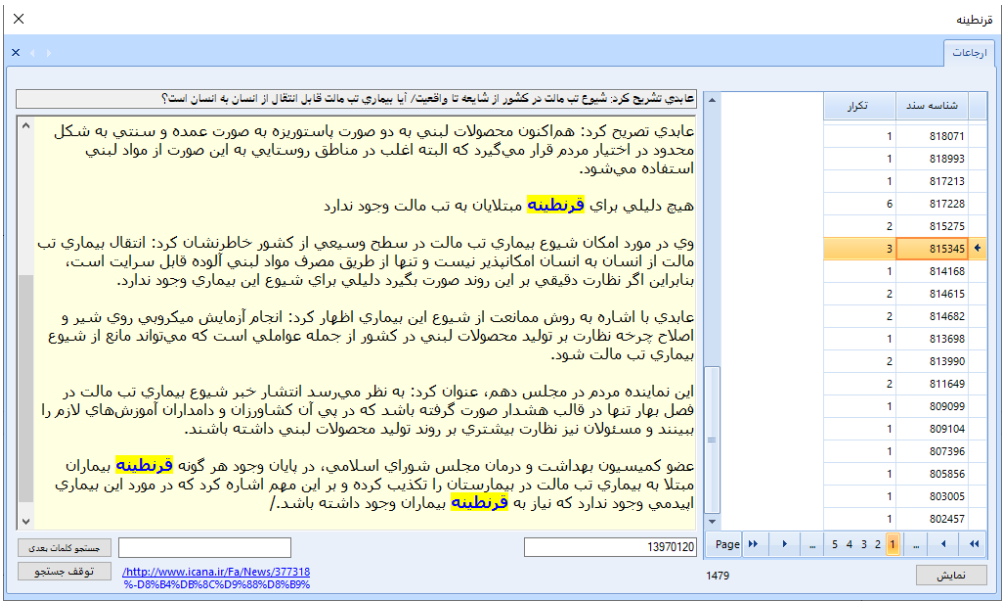
در هنگام اجرای اولیه نرم‌افزار، ۲۰۰۰ واژه اول پایگاه داده در قالب چهار صفحه ۵۰۰ واژه‌ای به نمایش در می‌آید. علت محدودسازی تعداد در نمایش اولیه، افزایش سرعت بارگذاری داده‌ها و کاهش زمانی انتظار کاربر برای بارگذاری نرم‌افزار است. ضمن آنکه بارگذاری تمامی داده‌ها در حافظه بر روی رایانه‌هایی که حافظه رَم[1] کافی نداشته باشند، باعث صرفه بخش عمده‌ای از حافظه رایانه و کندی اجرای سایر نرم‌افزارها خواهد شد. کاربران در صورت نیاز به بارگذاری تمامی

---

[1]. RAM

داده‌ها می‌توانند از کلیدهای «بارگذاری تمام واژه‌ها» و «بارگذاری مرتب شده» استفاده نمایند که اجرای این دستورات مدت زمان قابل توجهی به طول خواهد انجامید.

با انتخاب هر واژه می‌توان از کلید «نمایش ارجاعات» استفاده نمود. در صورت استفاده از این گزینه، پنجره‌ای از ارجاعات واژه مورد نظر نمایش داده می‌شود:



شکل ۲ . پنجره ارجاعات

اجزای پنجره ارجاعات

در پنجره ارجاعات، تمامی متن‌هایی که واژه مورد بررسی در آنها به کار رفته است نمایش داده می‌شود. در جدول سمت چپ فهرست اسنادی که از واژه مورد جستجو در آنها استفاده شده است قرار دارد. در مقابل هر سند تعداد تکرار واژه مورد جستجو نمایش داده شده است. در سمت راست پنجره، متن سند نمایش داده می‌شود. در متن واژه مورد جستجو با رنگ متفاوت نمایش داده شده است. همچنین آدرس اینترنتی محل استخراج متن سند و تاریخ انتشار در زیر متن نمایش داده شده‌اند.

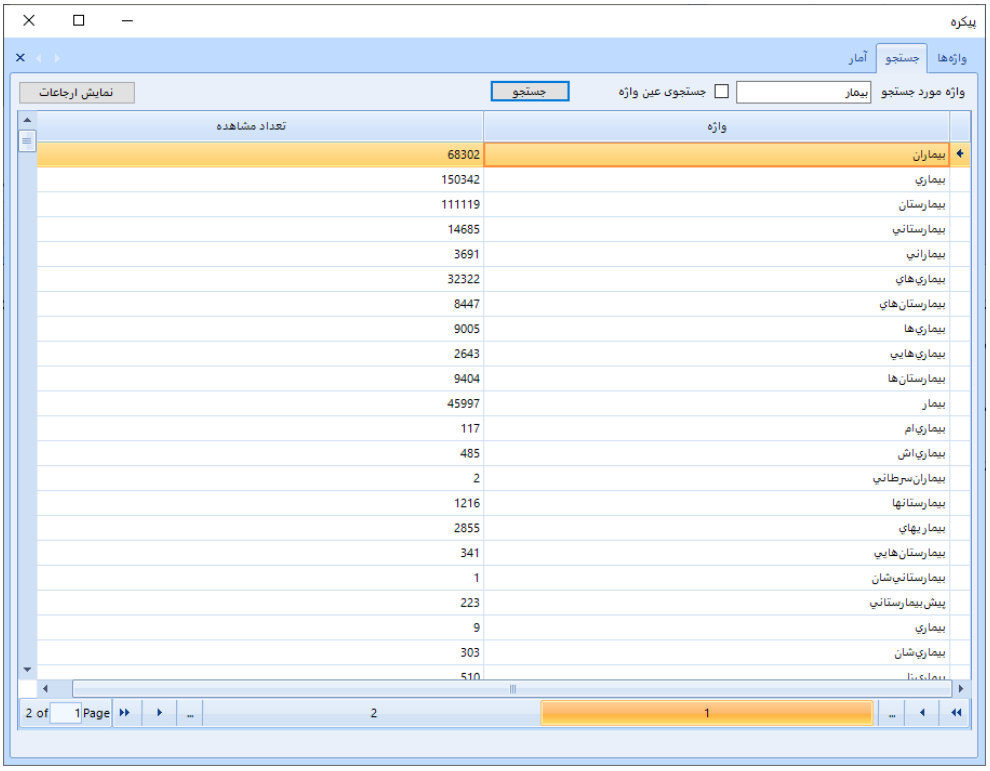
امکان جستجو کلمات بعد از واژه اصلی نیز در این پنجره ایجاد شده است. به عنوان مثال با جستجوی واژه «سنتی» در ارجاعات مربوط به واژه درمان نتایج زیر حاصل می‌شود:

شکل ۳. پنجره جستجوی واژه در ارجاعات مربوط

در زیر جدول سمت چپ آمار تعداد کلمات بعدی یافته شده به نسبت تعداد کل اسناد نمایش داده می‌شود. همچنین واژه دوم جستجو در اسناد یافته شده به رنگ متفاوت نمایش داده می‌شود.

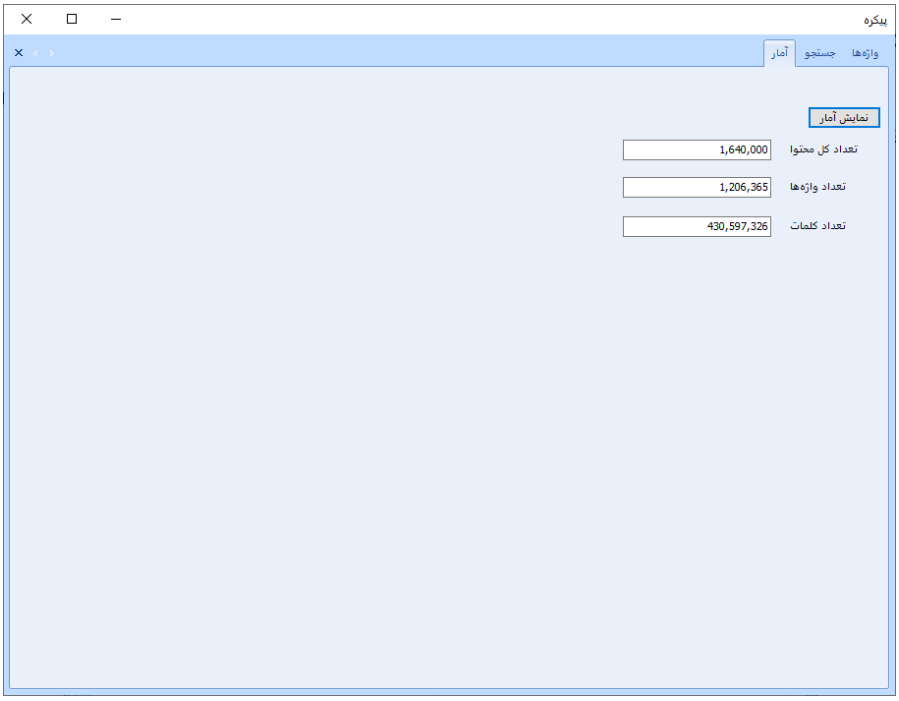
برگه جستجو: در این برگه امکان جستجو یک واژه در کل پایگاه داده وجود دارد. پس از ورود واژه یا بخشی از واژه مورد نظر و زدن کلید جستجو، فهرستی از رکوردهایی که شامل آن واژه باشند نمایش داده می‌شود. در صورتی که گزینه «جستجوی عین واژه» انتخاب شود، تنها رکورد مربوط به واژه مورد جستجو نمایش داده می‌شود و واژه‌های مشابه نمایش داده نمی‌شود.

در صورت انتخاب هر یک از واژه‌های پیدا شده و فشار کلید «نمایش ارجاعات»، صفحه ارجاعات مربوط به واژه مورد جستجو نمایش داده خواهد شد.



شکل ۴ . پنجره ارجاعات مربوط به واژه مورد جستجو

برگه آمار: در این برگه آمار واژه‌ها و ارجاعات پایگاه داده استخراج و نمایش اده می‌شوند. با توجه به حجم قابل توجه پایگاه داده و تعداد رکوردهای ذخیره شده در آن، استخراج آمار، به خصوص در رایانه‌هایی با مشخصات سخت‌افزاری ضعیف، بسیار زمان‌بر و طولانی خواهد بود. تعداد کل اسناد محتوا، تعداد کل واژه‌ها و مجموع کلمات ذخیره شده در پایگاه داده مطابق با تصویر زیر است. قابل ذکر است که افعال و کلمات رابطه‌ای متون مورد استفاده در محاسبه این آمار پیکره لحاظ نشده است.



شکل ۵ . پنجره آمار واژه‌ها و ارجاعات پایگاه داده

٤. تجزیه و تحلیل یافته‌ها

برخی از داده‌های آماری بدست آمده از پایگاه داده پیکره به شرح زیر است:

- تاریخ انتشار جدیدترین سند: ۱۳۹۷/۱۱/۰۶
- تاریخ انتشار قدیمی‌ترین سند: ۱۳۸۹/۱۲/۱۰
- بیشترین بسامد استفاده از یک واژه: واژه «ایران» با ۲۶۱۹۲۶۵ ارجاع
- تعداد کل محتواهای تمام‌متن در بانک اطلاعاتی: ۱۶۴۰۰۰۰ سند
- تعداد واژه‌های استفاده شده در متن اسناد: ۱۲۰۶۳۶۵ واژه
- تعداد کل کلمات استفاده شده در متن اسناد: ۴۳۰۵۹۷۳۲۶ کلمه

۵. بحث و نتیجه‌گیری

در این مقاله معرفی و شیوه ساخت پیکره‌ای خودکار به تفصیل مورد بحث قرار گرفت. پژوهش‌ها نشان داده که ساخت پیکره در مقیاس بزرگ ساده نیست و به زمان هزینه زیادی نیاز دارد. بنابریان برای کارایی و اثربخشی بیشتر آنها، باید در مرحله ساخت پیکره به نکاتی توجه نمود و آنها را در جریان ساخت پیکره‌ها لحاظ کرد. از جمله آنها می‌توان به روش‌های خودکار تهیه پیکره اشاره کرد.

مطالعه حاضر نشان داد که در زمینه ساخت و تهیه پیکره می‌توان از ابزار و شیوه‌های گوناگون بهره گرفت که بیش از همه هدف پژوهش، ساخت و استفاده از آن را توجیه می‌کند. با این حال، بدیهی است که تکیه صرف بر آن کافی نیست و می‌بایست در هر دوره زمانی و با توجه به کارکردها، از جنبه‌های گوناگون بررسی و به‌روزرسانی شود.

امروزه به کارگیری و توجه به پیکره‌ها و زبانشناسی پیکره‌ای برای تحلیل منابع در حوزه‌های مختلف و به ویژه مطالعات زبانشناختی، پژوهش‌های زبانی را اعتبار دیگری بخشیده است و مزایای استفاده از پیکره‌ها در تحلیل‌ها و استدلال‌های زبانی همچون: بهره‌گیری از حجم زیاد داده‌ها، گردآوری نظام‌مند و صرفه‌جویی در زمان سبب شده که با کمک آنها بتوان اطلاعات نهفته در متون را شناسایی و استخراج کرد و اهداف پژوهشی گوناگونی را دنبال نمود (عاصی و قندی، ۱۳۹۴؛ میرزایی و صفری، ۱۳۹۴؛ صفری، ۱۳۹۵؛ قدردوست نخچی و دیگران، ۱۳۹۵).

از آنجا که این پیکره حاوی مقالات و متون خبری، خبرگزاری‌های فارسی زبان است، یک پیکره عمومی زبان طبیعی به شمار می‌رود و برای پردازش‌هایی که مستلزم بهره‌جویی و استفاده از متون زبان طبیعی است، مناسب و ارزشمند خواهد بود.

فهرست منابع

افراشی، آزیتا، مصطفی عاصی، و کامیار جولایی. ۱۳۹۴. استعاره‌های مفهومی در زبان فارسی؛ تحلیلی شناختی و پیکره‌مدار. *زبان‌شناخت* ۶ (۲): ۳۹-۶۱.

بحرانی، محمد، حسین صامتی، نازیلا حافظی، و سعیده ممتازی. ۱۳۸۶، اسفند ۱۹-۲۱. خوشه‌بندی خودکار کلمات بر اساس مقوله‌های نحوی برای سیستم‌های بازشناسی گفتار پیوسته فارسی. مقاله ارائه شده در *سیزدهمین کنفرانس ملی انجمن کامپیوتر ایران*، جزیره کیش، ایران.

پاول، رونالد ار. ۱۳۸۹. *روش‌های اساسی پژوهش برای کتابداران*. (نجلا حریری، مترجم). [تهران]: آثار نفیس.

دشتبانی، شکوفه، محرم منصوری‌زاده، و محمد نصیری. ۱۳۹۱، شهریور ۱۵-۱۶. طراحی و ساخت پیکره‌ی متنی برای حوزه تخصصی فاوا. مقاله ارائه شده در *نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی*، سمنان.

دفتری‌نژاد، الهه. ۱۳۸۵. فرآیند مارکوف، الگوی احتمالاتی رفع ابهام در زبان‌شناسی رایانه‌ای. *علوم انسانی دانشگاه الزهرا (س)* ۱۶-۱۷ (۶۳-۶۴): ۱۰۷-۱۳۹.

ذوالفقار، زهره، طیبه موسوی میانگاه، بلقیس روشن، و امیررضا وکیلی‌فرد. ۱۳۹۹. بررسی تکنیک‌های بهبود عملکرد روش‌های بسامدشماری پیکره‌بنیاد در استخراج خودکار واژگان (مورد مطالعه: واژگان پایه علوم پزشکی). *پژوهش‌نامه پردازش و مدیریت اطلاعات* ۳۵ (۴): ۱۰۳۹-۱۰۶۴.

رباطی، زهرا. ۱۳۹۳. *دسته‌بندی اخبار فارسی با استفاده از تکنیک‌های هوش مصنوعی*. پایان‌نامه کارشناسی‌ارشد، دانشگاه صنعتی شاهرود، [شاهرود].

رضایی‌پناه، امیر، و سمیه شوکتی مقرب. ۱۳۹۵. تحلیل پیکره‌بنیاد مدارهای هویت در سند استراتژی امنیت ملی ۲۰۱۵ بریتانیا. در *مجموعه مقالات دومین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۶۹-۹۱. تهران: نشر نویسه پارسی.

روحانیان، مرتضی، مصطفی صالحی، علی درزی، و وحید رنجبر. ۱۳۹۹. تحلیل احساس در رسانه‌های اجتماعی فارسی با رویکرد شبکه عصبی پیچشی. *مهندسی برق و مهندسی کامپیوتر ایران* ۱۸ (۱): ۵۹-۶۶.

سلامی، مریم، زهرا سادات جلالی، مریم پاکدامن نائینی، و محمد علائی آرانی. ۱۳۹۴. تحلیل محتوای مقالات علوم پزشکی بر اساس مطالعه پیکره زبانی. *مدیریت اطلاعات سلامت* ۱۲ (۵): ۵۹۵-۶۰۷.

شهشهانی، مهسا، مهدی محسنی، آزاده شاکری، و هشام فیلی. ۱۳۹۸. پیکره برچسب خورده موجودیت‌های اسمی زبان فارسی. *پردازش علایم و داده‌ها* ۱۶ (۱): ۹۱-۱۰۹.

صفری، سعید. ۱۳۹۴. از زبان‌شناسی پیکره‌ای تا پیکره زبان‌آموز. در *مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۱۳۱-۱۵۲. تهران: نشر نویسه پارسی.

صفری، سعید. ۱۳۹۵. پیکره زبان‌آموز: مبانی، روش‌شناسی، الگوی طراحی و تولید. در *مجموعه مقالات دومین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۹۳-۱۲۳. تهران: نشر نویسه پارسی.

عاصی، مصطفی، و سعیده قندی. ۱۳۹۴. پایگاه داده‌های زبان فارسی و پیکره تاریخی آن. در *مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۱۹۳-۲۲۰. تهران: نشر نویسه پارسی.

علایی ابوذر، الهام؛ نصراله پاک‌نیت، علی‌اصغر حجت‌پناه، مجتبی زالی، و محمدهادی آقالویی آغمیونی. ۱۴۰۰. معرفی یک پیکره متنی تخصصی: پیکره پژوهش‌نامه. *پژوهش‌های زبان‌شناسی تطبیقی* ۱۱ (۲۲): ۲۷۱-۲۸۹.

قدردوست نخچی، سعیده، ندا پورمرتضی خامنه، پری‌ناز دادرس، و سلیمه زمانی. ۱۳۹۵. بررسی پیکره‌بنیاد مقوله قید. در *مجموعه مقالات دومین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۱۴۷-۱۶۵. تهران: نشر نویسه پارسی.

کامیابی گل، عطیه، الهام اخلاقی باقوجری، احسان عسگریان، و هانیه حبیبی. ۱۳۹۷. استخراج اطلاعات از پیکره زبانی: معرفی پیکره مقاله‌های علمی پژوهشی دانشگاه فردوسی مشهد. *کتابداری و اطلاع‌رسانی* ۲۱ (۲): ۳-۲۵.

مظاهری، ویدا، و چنگیز دل‌آرا. ۱۳۹۸، مرداد. استخراج اطلاعات از وب‌سایت‌های خبری با استفاده از روش مبتنی بر آنتولوژی. مقاله ارائه شده مقاله ارائه شده در *هفتمین کنفرانس ملی علوم و مهندسی کامپیوتر وفناوری اطلاعات*، مازندران، ایران.

میرزائی، آزاده، و پگاه صفری. ۱۳۹۴. ساختِ واژه- متن‌های تخصصی و عمومی زبان فارسی بر اساس بسامدگیری واژه‌های نقشی و محتوایی. در *مجموعه مقالات نخستین همایش ملی زبان‌شناسی پیکره‌ای*، ویراسته آزاده میرزایی، ۱۷۵-۱۹۱. تهران: نشر نویسه پارسی.

میرزائی، آزاده، و امیرسعید مولودی. ۱۳۹۳. نخستین پیکره نقش‌های معنایی زبان فارسی. *علم زبان* ۲ (۳): ۲۹-۴۷.

نظارات، امین؛ طیبه موسوی میانگاه. ۱۳۹۰. طراحی و پیاده‌سازی یک سامانه بازیابی اطلاعات دو زبانه با استفاده از پیکره‌های زبانی. *پژوهش‌نامه پردازش و مدیریت اطلاعات*، ویژه‌نامه ذخیره، *بازیابی و مدیریت اطلاعات*: ۱۹۷-۲۱۲.

نظری، مریم. ۱۳۹۲. گسست دانشی در پژوهش‌های مولد چگونه رصد می‌شود؟ پیشنهاد ترسیم دو نقشه: نقشه دانش و نقشه پژوهش. *تحقیقات کتابداری و اطلاع‌رسانی دانشگاهی* ۴۷ (۱): ۲۷-۴۸.

Bennett, Gena R. 2010. *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. [Michigan]: University of Michigan Press.

Kilgarriff, Adam, and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. Computational Linguistics 29 (3): 333-347.

Kofax. 2016. Kofax Kapow. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwid7NjGuJ__AhVgSfEDHeNbAGIQFnoECAwQAQ&url=https%3A%2F%2Fcobwebb.com%2Fwp-content%2Fuploads%2F2021%2F11%2Fds-kofax-kapow-en.pdf&usg=AOvVaw2aAIEADX7lGrhnmULWN85g (دسترسی در ۱۴۰۱/۱۲/۱۶).

Kofax. 2017. Kofax Kapow. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjck7GEs5__AhV1RvEDHdJ9B6cQFnoECAwQAQ&url=https%3A%2F%2Fbpas.dk%2Fwp-content%2Fuploads%2F2018%2F02%2FKapow-datasheet.pdf&usg=AOvVaw0Oxy4hId6MjYY_a0D-MAGm (دسترسی در ۱۴۰۱/۱۲/۱۶).

Li, Qin, Shaobo Li, Sen Zhang, Jie Hu, and Jianjun Hu. 2019. A Review of Text Corpus-Based Tourism Big Data Mining. *Applied Sciences* 9: 3300.

Liu, Vinci, and James R. Curran. 2006, April 3-7. Web Text Corpus for Natural Language Processing. Paper presented at *11th Conference of EACL: The European Chapter of the Association for Computational Linguistics*. Trento, Italy.

Mihalcea, Rada, Courtney Corley, Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *AAAI'06: Proceeding of the 21st National Conference on Artificial Intelligence*, (Vol.1, P:775-780).

Pustejovsky, James, Sabine Bergler, Peter Anick. (1993). Lexical Semantic Techniques for Corpus. *Computational Linguistics* 19 (2): 331-358.

Sabeti, Behnam, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobasti, S.H.E. Mortazavi Najafabadi, Amir Vaheb. 2018. MirasText: An Automatically Generated Text Corpus for Persian. In *Proceedings of the Eleventh International Conferenece on Language Resources and Evaluation (LREC 2018)*, 1174-1177. Japan: European Language Resources Association (ELRA).

Sokolova, Marina, Victoria Bobicev. 2018. Corpus Statistics in Text Classification of Online Data. *Arxiv*: 1803.06390.

Verma, Parul, and Brijesh Khandelwal. 2019. Word Embeddings and Its Application in Deep Learning. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 8 (11): 337-341.

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات
X xx ۱۳ | دورهٔ xx | شمارهٔ x

Natural Language Text Corpus: Design, Construction and Management

Hamideh Asadi *
PhD Candidate in Library and Information Science-Information Retrieval; University of Tehran; Tehran, Iran.
Asadi1366@gmail.com

Nader Naghshineh
Associated Professor in Library and Information Science; University of Tehran; Tehran, Iran.
nnaghsh@ut.ac.ir

Moluk Sadat Hosseini Beheshti
Associated Professor in Terminology & Ontology Research Group, Iranian Research Institue for Information Science and Technology (IranDoc), Tehran, Iran.
beheshti@irandoc.ac.ir

Abstract

Aim: Considering the role of corpora in various fields of study and the need to construct a general corpus to increase efficiency and effectiveness in processes that require the extraction/use of natural language text, the purpose of this study is to focus on design and automatic construction of natural language text corpus and software for its management.

In this research, a technology-based method has been used to build a monolingual composition in Persian language

Methodology: In this research, a technology-based method has been used to construct a monolingual corpus in Persian language. This corpus is produced automatically by collecting web data and its sources are news texts included in Persian language news agencies.

Findings: In the study, a corpus of natural language texts in Persian language was made. Due to the automaticity of the construction process, software is needed to manage it both in the construction stage and in the information extraction stage, which was designed, construct and implemented in this study.

Result: The construction of general corpus of natural language texts is used for various research purposes, and the proposed method and the use of introduced tools in this study can facilitate the construction of corpus. Also, software design for corpus management will save time and cost of construction and will provide the possibility of extracting information from it.

Keywords: Corpus, Data Set, Natural Language Processing, NLP, Corpus Linguistic, Artificial Intelligence

حمیده اسدی: دانشجوی دکتری رشته علم اطلاعات و دانش‌شناسی با گرایش بازیابی اطلاعات از دانشگاه تهران است. حوزه‌های روش‌شناسی پژوهش، بازیابی اطلاعات و علم‌سنجی از جمله علایق پژوهشی وی است.

نادر نقشینه: دارای مدرک تحصیلی دکتری در رشته علم اطلاعات و دانش‌شناسی از دانشگاه تهران است. ایشان هم‌اکنون دانشیار گروه علم اطلاعات و دانش‌شناسی دانشگاه تهران است. حوزه‌های فناوری اطلاعات، داده‌کاوی، سیبرنتیک از جمله علایق پژوهشی وی است.

ملوک‌السادات حسینی بهشتی: درای مدرک تحصیلی دکتری در رشته زبانشناسی با گرایش همگانی از دانشگاه تهران است. ایشان دانشیار پژوهشکده علوم اطلاعات، گروه اصطلاح‌شناسی و هستان‌شناسی پژوهشگاه علوم و فناوری اطلاعات (ایرانداک) است. مطالعه اصطلاح‌شناسی، مدیریت دانش، مدیریت اطلاعات، سازماندهی اطلاعات و پردازش زبان طبیعی از جمله علایق پژوهشی وی است.