نشریه پژوهش‌های زبان‌شناسی تطبیقی
سال یازدهم- شماره ۲۲- پاییز و زمستان ۱۴۰۰

معرفی یک پیکرهٔ متنی تخصصی: پیکرهٔ پژوهش‌نامه

الهام علایی‌ابوذر[۱*]
نصراله پاک‌نیت[۲]
علی‌اصغر حجت‌پناه[۳]
مجتبی زالی[۴]
محمدهادی آقالویی آغمیونی[۵]

مقاله پژوهشی

چکیده

بسیاری از پژوهش‌های زبان‌شناسی و برنامه‌ریزی‌های زبانی با استفاده از پیکره‌های زبانی انجام می‌شود. در این پژوهش پیکره‌ای با استفاده از متون مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات ساخته شده است. این پیکره شامل بیش از ۶۰۰ مقاله (بیش از چهار میلیون واژه) است. موضوع این مقاله‌ها کتابداری و اطلاع‌رسانی، علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، زبان‌شناسی، زبان‌شناسی رایانشی، اصطلاح‌شناسی، هستان‌شناسی و سایر حوزه‌های پردازش اطلاعات است. متون مقاله‌ها تخصصی و میان‌رشته‌ای است و برای پردازش‌هایی که مستلزم بهره‌گیری از متون تخصصی است، ارزشمند است. برای ساخت پیکره پس از نمونه‌گیری و وارد کردن داده‌ها در پیکره، فرادادهٔ مقاله‌ها وارد پیکره شد. سپس نرمال‌سازی ماشینی و به دنبال آن برچسب‌گذاری ماشینی (نوعاً برچسب‌گذاری اجزای واژگانی کلام) انجام شد. در نهایت تعداد قابل توجهی از فایل‌های برچسب‌خورده در پیکره به‌صورت رندوم انتخاب شد و الگوهای زبانی برای اصلاح ماشینی و دستی برچسب‌ها استخراج و در پیکره به کار برده شد.

کلیدواژه‌ها: پیکره، نرمال‌سازی، برچسب‌گذاری اجزای واژگانی کلام.

۱- استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)* ✉| alayi@irandoc.ac.ir
۲- استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) ✉| pakniat@irandoc.ac.ir
۳- رئیس اداره سامانه‌های اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) ✉| hojjatpanah@irandoc.ac.ir
۴- پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) ✉| m.zali@irandoc.ac.ir
۵- پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) ✉| aghalouei@irandoc.ac.ir

تاریخ دریافت: ۱۴۰۰/۰۲/۱۵ **تاریخ پذیرش: ۱۴۰۰/۰۵/۲۴**

شناسه دیجیتال (DOI): 10.22084/RJHLL.2021.24209.2136

۱- مقدمه

بسیاری از پژوهش‌های زبان‌شناسی و تصمیم‌گیری‌ها در برنامه‌ریزی زبانی، تنها با استفاده از یک پیکرهٔ زبانی امکان‌پذیر است. برخی از کاربردهای پیکره شامل موارد زیر است:

پردازش زبان طبیعی و درک و بازشناسی گفتار، تبدیل متن به گفتار و برعکس، تدوین فرهنگ‌ها، آموزش و پژوهش، بررسی واژه‌های هم‌آیند در زبان‌های گوناگون، پایشگری زبان برای پیگیری و ردگیری دگرگونی‌های زبانی، ترجمهٔ ماشینی، توسعهٔ مفاهیم و منابع در پیوند با واژگان، نگارش و گسترش مهارت‌های نوشتاری، آموزش و یادگیری زبان با شناخت گویش‌ها و گوناگونی زبانی، پژوهش‌های گوناگون زبان‌شناختی، واکاوی ژانرهای ادبی و پژوهش‌های دستوری است. در استفاده از پیکره در مطالعات زبان‌شناسی سه رویکرد رایج وجود دارد: رویکرد پیکره‌بنیاد[1]: در این رویکرد نظریه مطرح می‌گردد و برای سنجش نظریه از پیکره استفاده می‌شود. رویکرد برگرفته از پیکره[2]: در این رویکرد از پیکره برای بازیابی اطلاعات استفاده می‌شود. رویکرد پیکره‌یار[3]: این رویکرد ترکیبی است از به‌کارگیری روش‌های کمی و کیفی.

پیکره انواع گوناگونی دارد؛ اتکینز و همکاران (۱۹۹۲) انواع پیکره را از چند دیدگاه مختلف بررسی کرده‌اند و بر آن اساس پیکره‌ها را به انواع «متن کامل[4]»، «نمونه‌ای[5]»، «نظارتی[6]»، «بسته[7]»، «باز[8]»، «هم‌زمانی[9]»، «درزمانی[10]»، «عمومی[11]»، «اصطلاح‌شناسی[12]»، «یک‌زبانه[13]»، «دوزبانه[14]»، «چندزبانه[15]»، «منفرد[16]»، «موازی[17]»، «مرکزی[18]»، «پوسته‌ای[19]»، «هسته‌ای[20]» و «پیرامونی[21]» طبقه‌بندی کرده‌اند. در پژوهش حاضر با در نظر گرفتن ابعاد و مؤلفه‌های ساخت پیکرهٔ متنی (علایی و علیدوستی، ۱۳۹۹) و نیز ملاحظات حقوقی، از متون مقاله‌های موجود در «پژوهش‌نامهٔ پردازش و مدیریت اطلاعات» برای ساخت پیکره استفاده

1. corpus-based approach
2. corpus-driven approach
3. corpus-assisted approach
4. whole text
5. samples
6. monitor
7. open
8. closed
9. synchronic
10. diachronic
11. general
12. thesaurus
13. monolingual
14. bilingual
15. multilingual
16. single
17. parallel
18. central
19. cluster
20. nuclear
21. perimeter

شده است. از آنجائی‌که این پیکره از مقاله‌های مجلهٔ پردازش و مدیریت اطلاعات ساخته شده است، محتوای این پیکره عمومی نیست و حاوی متون تخصصی و میان‌رشته‌ای است و از این نظر برای پردازش‌هایی که مستلزم بهره گرفتن از متون تخصصی است، بسیار ارزشمند خواهد بود.

۲- مرور پژوهش‌های پیشین ساخت پیکرهٔ زبانی

پیکره‌های زبانی و پژوهش‌های پیکره – بنیاد زبان در زبان‌شناسی پیکره‌ای نقشی کلیدی دارند. ساخت پیکره‌ها کاری پیچیده، زمان‌بر، دارای گام‌های گوناگون و میان‌رشته است. بنابراین برای کارایی و اثربخشی در ساخت آنها نیاز است که پیش از آغاز کار، امکان‌سنجی انجام شود. امکان‌سنجی؛ گام‌ها، هزینه‌ها، نیروی انسانی، حقوق مادی و معنوی و مانند آنها را برای یک پروژه، بررسی و به مدیریت پروژه کمک می‌کند تا آن را با آمادگی و آینده‌نگری بیشتری به سرانجام برساند. در این راستا و به منظور ارائهٔ مدلی برای امکان‌سنجی ساخت پیکره، علایی و علیدوستی (۱۳۹۹) پس از بررسی پژوهش‌های پیشین، شاخص‌های کلی ساخت پیکره‌های زبانی را استخراج کرده‌اند. سپس، فرآیند ساخت پیکره را به تفصیل مورد بررسی قرار داده‌اند. پس از بررسی پژوهش‌های پیشین در زمینهٔ امکان‌سنجی، بر پایهٔ اطلاعات به‌دست آمده از مطالعات امکان‌سنجی، بررسی شاخص‌های ساخت پیکره و فرآیند ساخت پیکره، ابعاد و مؤلفه‌های امکان‌سنجی ساخت پیکره استخراج شده است و مدلی کلی برای ساخت پیکره پیشنهاد شده است. سپس با روش دلفی و در دو دور اعتباریابی، مدل پایانی به‌دست آمده است. این مدل دارای هفت بُعد فنی، اقتصادی، زمان‌بندی، قانونی حقوقی، عملیاتی، تأمین مالی، و بازاریابی و روی‌هم‌رفته ۳۳ مؤلفه است. واینه (۲۰۰۵) نیز نظرات نویسندگان و پژوهشگران سرشناس حوزهٔ ساخت پیکره را ارائه می‌دهد. این پژوهشگران پیشنهادات کاربردی برای ساخت پیکره (از نوع داده‌های مورد استفاده از پیکره تا انواع و چگونگی حاشیه‌نویسی) ارائه می‌کنند. در این کتاب، به زبان ساده مراحل ساخت پیکره توضیح داده شده است. بی‌جن‌خان و همکاران (۲۰۱۱) به بررسی موضوعاتی می‌پردازند که هنگام ساخت پیکره باید مدنظر قرار گیرند. پیکره‌ای که آنها معرفی کرده‌اند شامل ۳۵۰۵۸ فایل متنی است که هر متن حداقل ۱۰۰۰ واژه دارد. برای حاشیه‌نویسی از دستورالعمل EAGLES و روش نیمه‌خودکار استفاده شده است. همچنین آنها در ساخت این پیکره، به هم‌نگاره‌ها و ساخت اضافه نیز توجه داشته‌اند. ساخت پیکرهٔ متنی برای زبان فارسی با دشواری‌هایی همراه است؛ قیومی و همکاران (۲۰۱۳) به بررسی برخی از این موارد پرداخته‌اند. آنها معتقدند منبع مشکلات می‌تواند مربوط به نظام نوشتاری فارسی، تلفیق نظام نوشتاری فارسی با خط عربی، شیوهٔ تایپ‌کردن تایپیست‌ها و سایر موارد باشد. بخش زیادی از این مشکلات به‌صورت خودکار و از طریق برنامه‌نویسی یا به‌صورت دستی حل شده است. به‌عنوان نمونه، دربارهٔ کدهای مربوط به حروف، از همان یونیکدهای عربی استفاده شده است و برخی از یونیکدها که در عربی موجود نیست (مانند یونیکد مربوط به حروف «گ» یا «ژ») به یونیکدهای قبلی اضافه شده است. فاصله‌گذاری درست بین ریشه و وندها که معمولاً تایپیست‌ها رعایت نمی‌کنند، خود باعث مشکل در پردازش متن می‌شود؛ به‌عنوان مثال «کتاب‌ها»، «کتاب ها» و «کتابها» این‌ها در حقیقت باید یک واژه در نظر گرفته

شوند، اما اگر بین «کتاب» و «ها» یک فاصله (و نه نیم‌فاصله) بیفتد، در پردازش دو واژه در نظر گرفته می‌شوند. این مشکل به این صورت حل شده است که شکل‌های مختلف یک کلمه که در نتیجهٔ اتصال وند به ریشه/ ستاک ایجاد شده‌اند، یک شکل در نظر گرفته شوند و همگی با نیم‌فاصله فرض شوند؛ این عمل نیز به‌صورت نیمه‌خودکار و از طریق برنامه‌نویسی انجام شده است. اما مشکل مربوط به اجزای کلمات مرکب که معمولاً با ساخت اضافه به هم متصل می‌شوند، را نمی‌توان از طریق مذکور حل کرد. زیرا وندها محدود هستند و در برنامه‌نویسی می‌توان وندهای تصریفی و اشتقاقی را لحاظ کرد تا هر جا به ستاک/ ریشه متصل شده باشند، نیم‌فاصله در نظر گرفته شود، اما در مورد کلمات مرکب چون نمی‌توان تعداد ثابتی برای اجزای آنها در نظر گرفت، نمی‌توان چنین برنامه‌نویسی‌ای انجام داد. در این حالت باید شکل‌های مختلف کلمه توسط تحلیلگر ساخت‌واژی تشخیص داده شوند. پیکره‌ها می‌توانند آوایی، تخصصی، موازی، متنی و غیره باشد. به‌عنوان نمونه، در فصلی از کتاب «آواشناسی پیکره‌ای» دوراند و همکاران (۲۰۱۴) فرآیند ساخت پیکره‌های آوایی را توضیح می‌دهند. آنها پس از تعریف پیکرهٔ آوایی، به بررسی اجزای مهم ساخت چنین پیکره‌هایی می‌پردازند؛ این بخش‌ها شامل: ذخیرهٔ پیکره، چگونگی به اشتراک‌گذاشتن و استفاده مجدد از پیکره، نمایندگی و اندازهٔ پیکره، انتخاب دادهٔ خام و حاشیه‌نویسی پیکره است. کلود توریدا (۲۰۱۶) ساخت پیکرهٔ تخصصی[1] و تهیهٔ فهرست واژگان حاشیه‌نویسی شده و مبتنی بر فراوانی واژگان در پیکره را گام‌به‌گام توضیح می‌دهد. وی از پروژهٔ «آموزش زبان انگلیسی برای اهداف دانشگاهی» که در دانشگاهی در خاورمیانه توسعه یافته است، نمونه‌هایی را ذکر می‌کند. مراحل ساخت پیکرهٔ تخصصی از نظر کلود توریدا شامل این موارد است: انتخاب متون آموزشی، حذف واژگانی که قاموسی[2] نیستند (مانند حروف اضافه، حروف ربط و...)، تجزیه و تحلیل متن با استفاده از نرم‌افزار AntConc، ایجاد فهرست فراوانی واژه‌ها، توسعهٔ فهرست واژگان حاشیه‌نویسی شده که خود شامل این موارد است: تعیین مقولهٔ نحوی (POS) واژگان موجود در فهرست، اضافه کردن تعریف واژگان، باهم‌آیی واژگان و نمونه‌ای از جمله‌ای که واژه در آن به‌کار رفته است. از جمله تلاش‌هایی که در زمینهٔ ساخت پیکره‌های دوزبانه فارسی- انگلیسی انجام گرفته است، می‌توان به پیکرهٔ دشتبانی و همکاران (۱۳۹۳) اشاره کرد که پیکره‌ای است دو زبانه در حوزهٔ فاوا (حوزهٔ فناوری ارتباطات و اطلاعات). این پیکره به‌صورت خودکار ساخته شده است و منابع آن، اسناد تخصصی حوزهٔ فاوا است. در این پژوهش، نرم‌افزاری برای ساخت پیکره طراحی شده است که هزینه و مدت زمان ساخت پیکره را کاهش می‌دهد. علاوه بر این، نرم‌افزار ارائه شده قابلیت مدیریت پیکره را برای کاربران فراهم می‌کند. سیستم مدیریت پیکره دارای دو بخش اصلی است که بخش اول مربوط به ساخت پیکره است و بخش دوم مربوط به استخراج اطلاعات از پیکره است. نرم‌افزاری که برای مدیریت پیکره ایجاد شده است، حاشیه‌نویسی اسناد، جستجو در پیکره و تصحیح خطا را آسان می‌کند. قبل از شروع فرآیند پردازش اصلی، هر سند توسط نرم‌افزار پیش‌پردازش می‌شود؛ این کار به منظور انتخاب جمله‌های درست و معنی‌دار برای پردازش اصلی است؛ علاوه

1. specialized corpus
2. content words

بر آن، در صورتی‌که در سند کاراکترهای بی‌معنی وجود داشته باشد، در فرآیند پیش‌پردازش از سند حذف می‌شوند. از جمله کارهایی که این سیستم انجام می‌دهد می‌توان به این موارد اشاره کرد: ویرایش پیکره، اضافه کردن متن‌های جدید به پیکره، اندیس‌گذاری و حاشیه‌نویسی پیکره. بخش دوم، یک موتور جستجوی پیکره است که برای مدیریت مجموعه بزرگی از متون طراحی شده است. پردازش متون به کمک سیستمی انجام شد که دارای این بخش‌ها است: طبقه‌بند[1] برای پذیرش اسناد حوزهٔ فاوا، برچسب‌گذار نقش دستوری واژگان[2] و تجزیه‌کننده[3] برای اسناد فارسی و یک تجزیه‌کننده، برچسب‌گذار نقش دستوری واژگان و ریشه‌یاب[4] برای اسناد انگلیسی است. اسناد حوزهٔ فاوا به کمک این سیستم حاشیه‌نویسی می‌شوند و اطلاعات پردازش شدهٔ اسناد در پایگاه داده پیکره ذخیره می‌شوند. مهم‌ترین مرحلهٔ ساخت پیکره‌های چندزبانی، ترازبندی داده‌های پیکره است. در این پروژه روشی برای ترازبندی جمله‌های پیکرهٔ فارسی تخصصی حوزه فاوا و جملات انگلیسی پیکرهٔ تخصصی حوزه فاوا ارائه شده است. الگوریتم پیشنهادی آنها از مدل ترجمهٔ کلمه‌به‌کلمه و تکنیک بلندترین زیردنبالهٔ مشترک[5] برای ترازبندی استفاده می‌کند و در نهایت امتیاز نشان دهندهٔ شباهت دو جمله، محاسبه می‌شود و اطلاعات مربوط به نگاشت جمله‌های دو مجموعهٔ انگلیسی و فارسی در پایگاه دادهٔ پیکره، ذخیره می‌گردد.

از پیکره‌های مهم زبان فارسی می‌توان موارد زیر را نام برد: پیکرهٔ متنی زبان فارسی (بی‌جن‌خان و همکاران، ۲۰۱۱)، پایگاه دادگان زبان فارسی (عاصی، ۱۳۸۴)، پیکرهٔ واژگان نحوی و معنایی افعال مرکب فارسی (نسخه ۱٫۰)[6] (سامولین و فقیری، ۲۰۱۳)، فارس‌دات تلفنی (بی‌جن‌خان و همکاران، ۲۰۰۳)، دادگان دایفونی فارسی (اسلامی و همکاران، ۱۳۸۸)، فارس‌نت (شمس‌فرد و همکاران، ۲۰۱۰)، پیکرهٔ وابستگی نحوی زبان فارسی (رسولی و همکاران، ۲۰۱۳)، پیکرهٔ واژگان زایای زبان فارسی (اسلامی و همکاران، ۱۳۸۳)، پیکرهٔ همشهری (آل‌احمد و همکاران، ۲۰۰۹)، پیکره چندزبانه رایانامه‌ها (دهقانی و همکاران، ۲۰۱۳) و پیکرهٔ نقش‌های معنایی زبان فارسی (میرزایی و مولودی، ۱۳۹۳) است.

۳- روش ساخت پیکرهٔ متنی

نمونه‌گیری

در طراحی کلی پیکره، ملاحظاتی از قبیل نوع متون مورد استفاده، تعداد متون، انتخاب متون خاص، طول نمونه‌های متون و... وجود دارد که هر کدام مستلزم تصمیم‌گیری دربارهٔ نمونه‌گیری[7] است. تصمیماتی که در

1. classifier
2. pos tagger
3. parser
4. stemmer
5. longest common subsequence (LCS). هدف این روش، مقایسهٔ دو رشته و پیدا کردن شباهت بین آنهاست.
6. Perspred
7. sampling

این زمینه گرفته می‌شود بر مناسب بودن پیکره برای انواع تجزیه و تحلیل زبانی تأثیر می‌گذارد. نمونه‌گیری در واقع عمل انتخاب متون مربوط به هر ژانر با توجه به هدف تهیهٔ پیکره است. برخی از معیارهایی که بر اساس آنها نمونه‌گیری صورت می‌پذیرد شامل این موارد است: شکل متن[1] (گفتاری/ نوشتاری/ الکترونیکی)، نوع متن (کتاب/ مجله/ نامه)، حوزهٔ متن (آکادمیک/ عمومی)، زبان متن (زبان‌ها یا گونه‌های زبانی پیکره) و مکان متن (به‌عنوان مثال، انگلیسی بریتانیا باشد یا استرالیا) است. داده‌های پیکرهٔ متنی از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات تهیه شده است؛ بنابراین، گونهٔ زبانی متون موجود در مجله، گونهٔ نوشتاری و رسمی است و حوزهٔ متون، آکادمیک است. نمونه‌گیری به این صورت انجام شد: جامعهٔ آماری، مجموعه مقاله‌های موجود در پژوهش‌نامهٔ پردازش و مدیریت اطلاعات است که شامل ۱۰۸۹ مقاله می‌شود؛ این مقاله‌ها در بازه زمانی سال ۱۳۵۱ تا ۱۳۹۹ در این مجله چاپ شده‌اند. متن بیش از ۶۰۰ مقاله که فایل ورد[2] آنها موجود بود، وارد پیکره شد. بنابراین، روش نمونه‌گیری خاصی برای انتخاب متون پیکره، به‌کار برده نشده است و هر آنچه موجود بوده است، وارد پیکره شده است. موضوع این مقاله‌ها شامل کتابداری و اطلاع‌رسانی، علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، مدیریت اطلاعات، مدیریت دانش، زبان‌شناسی، زبان‌شناسی رایانشی، اصطلاح‌شناسی، هستان‌شناسی و سایر حوزه‌های مرتبط با پردازش اطلاعات است. یکی از راههای مرتب کردن انواع فایل، استفاده از اطلاعات توصیفی است. این داده‌های توصیفی، فراداده[3] نامیده می‌شود و بسته به نوع فایل، متفاوت است. اطلاعات به این صورت وارد پیکره شده است که فراداده از طریق کاربرگه ثبت مقالات وارد شده است. سپس از طریق خروجی گرفتن از پایگاه داده SQL Server به برنامهٔ نرمال‌سازی و برچسب‌گذاری منتقل شده و مجدد اطلاعات در پایگاه داده ذخیره شده است. برای پیشگیری از خراب شدن داده‌ها (به هم ریختن فونت‌ها) هنگام انتقال دادن داده‌ها به پایگاه داده، بخش بزرگی از داده‌ها را به‌صورت دستی در پایگاه داده وارد کردیم؛ این کار از طریق ایجاد صفحه‌ای که در آن می‌توانستیم فراداده و متن مقاله‌ها را دستی وارد کنیم، انجام شد. اطلاعات توصیفی موجود در همهٔ مقاله‌های پیکره شامل عنوان مقاله، موضوع مقاله، نام نویسنده و متن مقاله در قسمت مربوطه وارد شد.

نرمال‌سازی داده‌ها

در ساخت پیکرهٔ متنی، پس از انتخاب متونی که در پیکرهٔ متنی مورد استفاده قرار می‌گیرند، باید پیش‌پردازش‌هایی روی متن انجام شود که به اصطلاح «نرمال‌سازی متن[4]» نامیده می‌شود. این پیش‌پردازش‌ها در متون فارسی می‌تواند شامل، یک‌دست کردن فاصله‌ها و نشانه‌گذاری‌های درون متن، یکسان کردن یونیکد کاراکترهای استفاده شده در متون (مانند انواع «ی»، «ک»، همزه و...)، یکسان کردن روش اتصال وندهای گوناگون به ستاک، اصلاح غلط‌های املایی، ارتباط دادن کلمات چندواژه‌ای و یکسان

1. mode
2. word
3. metadata
4. normalization

در نظر گرفتن آنها و... باشد. پس از واردکردن داده‌ها در پیکرهٔ متنی، نرمال‌سازی ابتدا به‌صورت ماشینی انجام شد. همچنین برای یکدست کردن انواع «ی» و «ک» و «کسره اضافه روی «ه»/ «ه» در حالت اضافی («هٔ»)، از دستورات TSQL در برنامه پایگاه داده SQL Server نیز استفاده شده است (جدول ۱).

جدول ۱: یونیکدهای «ی»، «ک» و «ه»

حروف عربی با یونیکد	ی ۱۶۱۰	ک ۱۶۰۳	هٔ ۱۵۷۷
حروف فارسی با یونیکد	ی ۱۷۴۰	ک ۱۷۰۵	ه ۱۶۰۷

حاشیه‌نویسی داده‌ها

حاشیه‌نویسی، اضافه کردن اطلاعات زبانی توضیحی – تفسیری به پیکره است. برخی از پژوهشگران مانند سینکلر (۲۰۰۴، نقل از لیچ: ۲۰۰۴)، ترجیح می‌دهند وارد مقولهٔ حاشیه‌نویسی پیکره نشوند؛ زیرا معتقد هستند پیکرهٔ حاشیه‌نویسی نشده، پیکره‌ای خالص است و برای مطالعات زبان‌شناسی چنین پیکره‌هایی را ترجیح می‌دهند. این پژوهشگران به حاشیه‌نویسی اعتماد ندارند و با شک به آن می‌نگرند و معتقد هستند حاشیه‌نویسی نمی‌تواند بدون خطا باشد. اما اکثر پژوهشگران این حوزه بر این باورند که حاشیه‌نویسی، پیکره را بسیار مفیدتر و غنی‌تر از پیکرهٔ خام می‌کند؛ از نظر آنها حاشیه‌نویسی، ارزشی مضاعف به پیکره می‌دهد. به‌عنوان مثال، عمل «برچسب‌گذاری اجزای واژگانی کلام»[1] که برای «پیکرهٔ براون[2]» انجام شده است، در سطح گسترده‌ای توزیع شده و مورد استفادهٔ افرادی قرار گرفته است که در راستای پژوهش‌های خود نیازمند برچسب‌گذاری اجزای واژگانی کلام یا پیکرهٔ برچسب‌گذاری شده بودند. غیر از برچسب‌گذاری اجزای واژگانی کلام، حاشیه‌نویسی‌های دیگری نیز وجود دارد که به سطوح مختلف تجزیه و تحلیل‌های زبانی یک پیکره مربوط می‌شود. به برخی از رایج‌ترین آنها اشاره می‌کنیم (لیچ: ۲۰۰۴):

– حاشیه‌نویسی آوایی[3]: اضافه کردن اطلاعاتی دربارهٔ نحوهٔ تلفظ یک کلمه در پیکرهٔ گفتاری یا اضافه کردن مشخصه‌های زبرزنجیری[4] (مانند تکیه[5]، آهنگ کلام[6]، وقفه[7] و...) به کلمات تشکیل‌دهندهٔ یک پیکره است.

– حاشیه‌نویسی نحوی[8]: اضافه کردن اطلاعاتی دربارهٔ نحوهٔ تجزیهٔ یک جمله به عبارت‌ها و اجزای تشکیل‌دهندهٔ آن است.

1. part of speech (POS) tagging
2. Brown corpus
3. phonetic annotation
4. prosodic
5. stress
6. intonation
7. pause
8. syntactic annotation

– حاشیه‌نویسی معنایی[1]: اضافه کردن اطلاعاتی دربارهٔ مقوله/ طبقه‌بندی معنایی کلمات پیکره است. به‌عنوان مثال، صورت نوشتاری <cricket> در انگلیسی، فارغ از صورت نوشتاری یا تلفظی یکسان، هم در طبقه‌بندی انواع ورزش (نوعی ورزش) قرار می‌گیرد و هم در طبقه‌بندی دیگر و متفاوتی تحت عنوان حشرات (نوعی حشره).

– حاشیه‌نویسی کاربردی[2]: اضافه کردن اطلاعاتی دربارهٔ کنش گفتار[3] که در مکالمه اتفاق می‌افتد. به‌عنوان مثال، پاره گفتار <okay> در انگلیسی، در موقعیت‌های مختلف می‌تواند به منزلهٔ اقرار و تصدیق، درخواست بازخورد، پذیرفتن یا نشانهٔ شروع مرحلهٔ جدیدی از بحث باشد.

– حاشیه‌نویسی گفتمانی[4]: اضافه کردن اطلاعاتی دربارهٔ ارتباطات ارجاعی[5] در متن است. به‌عنوان مثال، ارتباط دادن ضمیر <them> و مرجع آن <the horses> در جملهٔ زیر:

I`ll saddle the horses and bring them round.

– حاشیه‌نویسی سبکی[6]: اضافه کردن اطلاعاتی دربارهٔ نمود گفتار (گفتار مستقیم[7]، گفتار غیرمستقیم[8] و...)

– حاشیه‌نویسی واژگانی[9]: اضافه کردن اطلاعاتی مانند ریشهٔ کلمه.

یکی از رایج‌ترین نوع حاشیه‌نویسی پیکره، اضافه کردن برچسب‌ها به کلمات پیکره است؛ این برچسب‌ها می‌تواند نشان‌دهندهٔ مقولهٔ کلمات (اسم، فعل، صفت و...) باشد که «برچسب‌گذاری اجزای واژگانی کلام» نامیده می‌شود. برچسب‌گذاری اجزای واژگانی کلام، عملی کاربردی در بسیاری از حوزه‌های پیشرفته‌تر پردازش زبان طبیعی[10] از جمله ترجمهٔ ماشینی، خطایاب، تبدیل متن به گفتار، بازیابی اطلاعات، موتورهای جستجو و کمک به مدل‌های آماری است (مگردومیان: ۲۰۰۴). در این پژوهش، پس از ساخت پیکرهٔ متنی، با توجه به کاربرد برچسب اجزای واژگانی کلام در پردازش متن، تصمیم گرفته شد این نوع برچسب هم به پیکره اضافه شود. در این راستا ابتدا از یکی از ابزارهای آماده برای برچسب‌گذاری ماشینی اجزای واژگانی کلام، ابزار «هضم»، استفاده شد. «هضم» ابزاری است برای پردازش زبان فارسی در پایتون (sobhe/hazm). ویژگی‌های این ابزار شامل موارد زیر است

– تمیز و مرتب کردن متن
– قطیع جمله‌ها و واژه‌ها
– ریشه‌یابی واژه‌ها

1. semantic annotation
2. pragmatic annotation
3. speech act
4. discourse annotation
5. anaphoric
6. stylistic annotation
7. direct speech
8. indirect speech
9. lexical annotation
10. natural language processing

- تحلیل صرفی جمله
- تجزیه نحوی جمله
- واسط استفاده از داده‌های زبان فارسی
- سازگاری با بسته NLTK
- پشتیبانی از پایتون نسخه ۲ و ۳
- تست مداوم کدها

فهرست برچسب‌ها در «هضم»، شامل برچسب‌هایی است که در مقالهٔ بی‌جن‌خان و همکاران (۲۰۱۱) معرفی شده است (جدول ۲).

جدول ۲: فهرست برچسب‌ها

POS tag	Tag Name	مثال
N	Noun	کشاورز، خانه، کتاب
PREP	Preposition	از، در، برای
PUNC	Punction	نقطه، ویرگول، علامت سوال
AJ	Adjective	زیبا، اجتماعی، بزرگ
V	Verb	می‌تواند، خورد، شمرد
CON	Conjunction	که
NUM	Number	۵۰،۹۰ ،۱۲
PRO	Pronoun	من، تو، ایشان
DET	Determiner	این، آن، هر
ADV	Adverb	یقیناً، خوب، بسیار
POSTP	Postposition	را
RES	Residual	?
CL	Classifier	نوع، دست، چنین
INT	Interjection	ای، یا

شایان ذکر است که در فهرست برچسب‌ها در مقالهٔ بی‌جن‌خان و همکاران (۲۰۱۱) کسره اضافه نمایش داده نشده است؛ «هضم» به منظور نمایش کسره اضافه، هرگاه پس از واژه‌ای، کسرهٔ اضافه وجود داشته، کنار برچسب، e گذاشته است که نمایانگر کسرهٔ اضافه است. به‌عنوان نمونه، برچسب عبارت «شناسایی ساختار محتوایی مطالعات» به صورت: شناسایی (Ne) ساختار (Ne) محتوایی (ADJe) مطالعات (N) است.

کنترل دستی برچسب‌ها

کنترل دستی برچسب‌ها به این صورت انجام گرفت که ابتدا متن ۵۰ مقاله به‌عنوان نمونه انتخاب شد و برچسب‌ها به دقت مورد مطالعه قرار گرفت و به اصلاح برچسب‌های غلط پرداخته شد. به منظور اصلاح برچسب‌های غلط در پیکره، امکان ویرایش در صفحهٔ هر مقاله فراهم شد، به این صورت که تک به تک روی هر واژه‌ای که برچسب غلط داشت، کلیک می‌شد و پنجره‌ای باز می‌شد که در آن امکان اصلاح املایی واژه (اگر نیاز داشت) و برچسب آن فراهم می‌شد (شکل ۱).

شکل ۱: پنجره اصلاح برچسب‌ها

الگوی خطاها

ضمن اصلاح برچسب‌های غلط، به بررسی خطای موجود در برچسب‌گذاری نیز پرداخته شد تا اگر الگویی برای خطاها مشاهده می‌شود، بتوان به‌صورت ماشینی اصلاح کرد. برخی از این خطاها را می‌توان در دسته‌بندی زیر قرار داد:

- الگوی ۱: در بسیاری موارد واژهٔ «می‌توان» و «نمی‌توان» به دلیل ختم شدن به «ن» که نشانهٔ مصدر/اسم است، یا به دلیل این‌که نویسنده (در مقاله) بین پیشوند تصریفی «می-/نمی-» و «توان»، فاصله گذاشته است، به اشتباه برچسب اسم (N) دارد.
- به‌کارگیری الگوی ۱: هر جا تکواژهای «می-» و «نمی-» جدا از تکواژ بعد است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/ تکواژ بعد از آن متصل گردد.
- الگوی ۲: در بسیاری موارد پسوند جمع «-ها»، پسوند جمع «-ها» همراه با کسرهٔ اضافه («-های») و پسوند جمع «-ها» همراه با کسرهٔ اضافه و «-ی» نکره («هایی»)، با فاصله به تکواژ قبل از خود وصل شده است (مانند «کتاب ها، کتاب های، کتاب هایی) و این مسئله باعث شده است این واژه‌ها دو

برچسب داشته باشند (به‌عنوان نمونه واژهٔ «کتاب» برچسب اسم /N/ و پسوند تصریفی «ها» برچسب اسم /N/ یا صفت/ADJ/).

- به‌کارگیری الگوی ۲: هر جا تکواژ «-ها»، «-های» و «-هایی» جدا از تکواژ قبل است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/ تکواژ قبل از آن متصل گردد.
- الگوی ۳: تکواژ «-ای» در عباراتی همچون «کتابخانه ای»، در بعضی مواقع از تکواژ قبل از خود جدا افتاده است.
- به‌کارگیری الگوی ۳: هر جا تکواژ «-ی» و «-ای» جدا از تکواژ قبل است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/ تکواژ قبل از آن متصل گردد.
- الگوی ۴: در بسیاری موارد واژه‌هایی که به پسوندهای «-ساز/سازی» (مانند «نمایه‌سازی»)، «رسان/رسانی» (مانند «اطلاع‌رسانی»)، «-گیر/گیری» (مانند «نتیجه‌گیری»)، «-دار/داری» (مانند «معنی‌داری»)، «-شناس/شناسی» (مانند «زبان‌شناسی»)، «-سنج/سنجی» (مانند «علم‌سنجی»)، «-گذار/گذاری» (مانند «برچسب‌گذاری»)، «-پذیر/پذیری» (مانند «ریسک‌پذیری»)، «-بند/بندی» (مانند «طبقه‌بندی»)، «-افزار/افزاری» (مانند «نرم‌افزار»، «-نویس/نویسی» (مانند «برنامه‌نویسی»)، «-یاب/یابی» (مانند «دستیابی»)، «-آور/آوری» (مانند «جمع‌آوری»)، «-جو/جویی» (مانند «جست‌وجو» و....» ختم می‌شوند، به دلیل فاصله‌ای که نویسنده بین تکواژها گذاشته است، به اشتباه دو برچسب (یا بیشتر) گرفته‌اند و پسوندها برچسب اسم (N) یا فعل (V) ممکن است گرفته باشند.
- به‌کارگیری الگوی ۴: هر جا تکواژهای «-ساز/سازی»، «-گیر/گیری»، «-شناس/شناسی»، «-بند/بندی»، «-سان/رسانی»، «-سنج/سنجی»، «-گذار/گذاری»، «-پذیر/پذیری»، «-نویس/نویسی»، «-افزار/افزاری»، «-یاب/یابی»، «-دار/داری»، «-آور/آوری»، «-جو/جویی»، جدا از تکواژ قبل است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/ تکواژ قبل از آن متصل گردد.
- الگوی ۵: واژهٔ «داده» به دلیل حوزهٔ موضوعی مجلهٔ پردازش و مدیریت اطلاعات، بسیار در متون گوناگون تکرار شده است و در برخی موارد، به جای این‌که برچسب «اسم (N)» بگیرد، برچسب «فعل» (V) گرفته است.
- به‌کارگیری الگوی ۵: این مورد به صورت دستی کنترل و اصلاح شد.
- الگوی ۶: تکواژ «-تر» و «ترین» در برخی موارد از تکواژ قبل از خود جدا شده است، در نتیجه یک واژه دو برچسب گرفته است.
- به‌کارگیری الگوی ۶: هر جا تکواژ «-تر»، «-ترین» جدا از تکواژ قبل است و یک واژهٔ مجزا در نظر گرفته شده است، با نیم‌فاصله به واژه/تکواژ قبل از آن متصل گردد.

برای اصلاح برچسب‌ها از فهرست وندها و واژه‌بست‌های فارسی (جدول ۳)، برگرفته از مجموعه ۱۲ مقالهٔ دکتر علی‌اشرف صادقی تحت عنوان «*شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر ۱ تا ۱۲ : ۱۳۷۲-*

۱۳۷۰» و خسرو کشانی (در بحث اشتقاق) و لازار: ۱۳۸۹ و قطره: ۱۳۸۶ (در بحث تصریف) استفاده شده است.

جدول ۳: فهرست وندها و واژه‌بست‌های فارسی

وندها	نوع وند	مثال (بر اساس ترتیب فهرست وندها)
ها، ان، ین، ات، ی، ه، ش، تر/ترین، ت/اد/د/ید/ست، ان، می/ب/ن/م،	پیشوند یا پسوند تصریفی	کتاب‌ها، درختان، معلمین، درجات، کتابی (در عبارت «کتابی خوب»)، دختره، آشش (در جمله «آشش خوبه»)، خوب‌ترین/بهتر، کشت/ایستاد/خورد/رسید/آراست، رساند (رس+ان+د)، می‌گفت/بروم/نرو/مشنو
ام/ای/ه/س/ست//ایم/اید/اند (صورت‌های وابسته فعل بودن)، و (صورت پی‌بستی «را»)، و (حرف هم‌پایگی یا عطف)، م (صورت پی‌بستی «هم»)، ها و صورت محاوره آن (آ) (برای تأکید بعد از فعل می‌آید)، م/ت/ش/مان/تان/شان (ضمایر پی‌بستی/پیوسته)	واژه‌بست	زنده‌ام/زنده‌ای/زنده‌ست/زنده‌ایم/زنده‌اید/زنده‌اند، «و» در جمله »منو می‌بری؟» (من را می‌بری؟)، «و »در «من و تو»، «م» در «منم میام» (من هم می‌آیم)، اومدم آ، مدادم/مدادت/مدادمان/مدادتان/مدادشان
با، بی، نا، هم، گاه، ستان، کده، ئیه، ک، دان/دانی، زار، آباد، ی، سار، بار، لاخ، ان، بان، چی، دار، ی، یار، گر، ساز، دوز، باف، پز، ه، ک، ئیه، ی، گان، واره، ئینه، ک، چه، ی، ه، ا، یه، ی، ن، ا، کار، بندی، گری، کاری، بازی، ش، ار، ه، مان، ان، ئیت، مند، ور/اور/وار، ه، نده، ان، ا، ار، ی، گار، ند، ه، ی، ین، ئینه، گان، گانه، وند، ئیه، و، گین،، گون، فام، آسا، سان، وار، وش، انه، سا، وی، باره، آگین، آنی	پیشوند یا پسوند اشتقاقی	باادب، بیکار، ناکام، همکار، نشکن، دانشگاه، کردستان، دانشکده، جمشیدیه، قلهک، نمکدان/سگ‌دانی، کشتزار، حسن‌آباد، پنچری، کوهسار، جویبار، سنگ‌لاخ، دیلمان، نگهبان، کافه‌چی، صندوق‌دار، بازاری، استادیار، آهنگر، آهنگ‌ساز، کفش‌دوز، فرش‌باف، شیرینی‌پز، پایه، عروسک، فطریه، گوشی، ناوگان، جشنواره، گنجینه، طفلک، تاریخچه، مامانی، همشیره، آتوسا، خیریه، آزادی، رفتن، ژرفا، فراموش‌کار، طبقه‌بندی، وحشی‌گری، کتک‌کاری، کاغذبازی، آرایش، خریدار، خنده، زایمان، آشتی‌کنان، حساسیت، ثروتمند، بارور/سزاوار، همه‌کاره، فرساینده، لرزان، توانا، پرستار، شکاری، خواستگار، خوشایند، پخته، دنده‌ای، آهنین، دیرینه، خدایگان، پنج‌گانه، شهروند، تحریریه، اخمو، غمگین، دنده‌ای، آهنین، دیرینه، خدایگان، پنج‌گانه، شهروند، تحریریه، اخمو، غمگین، گندم‌گون، گل‌فام، رعدآساف گربه‌سانان، دیوانه‌وار، مهوش، دلیرانه، پریسا، دنیوی، شکمباره، زهرآگین، عصبانی،
گرا، نما، طلب، پرست، سنج، بر، پذیر، خیز، شناس، دان، نشین، خواه، دار،	پسوندواره فعلی	سنت‌گرا، سال‌نما، استقلال‌طلب، خداپرست، فشارسنج، تب‌بر، امکان‌پذیر، نفت‌خیز، هواشناس، ریاضیدان، کارگرنشین،

گیر، زن، کش، بند، خوار، فروش، گو، پوش، خوان، آوار، جو، انداز، رو، کن، آمیز، کوب، انگیز، تاب، چین، نویس، پرور، شکن، بخش، پیچ، افکن، ریز، آرا، بر، گداز، بار، نواز، پرواز، نشین، افشان، کش، گرد، گذار، تراش، جوش، کن، مال، یاب، اندیش، زا، شمار، افروز، پر، گشا، رس، گستر، خواب، نورد، پسند، دزد، ران، فریب، نام، نگار، افزا، آموز، پاش، بو، دوش، شو، آفرین، تراز، خند، سرا، آویز، دو، چران، نوش، ربا، پیوند، تاز، رسان، روب، انجام، پیرا، خا، دم، گزین، آزار، بیز، خر، فرسا، ورز، آگند، باش، ترس، چسب، اندوز، چرخ، زدا، گسار، توان، چر، آزما، آشوب، پران، توز، خراش، شتاب، گوار، آشام		مشروطه‌خواه، امانتدار، دندانگیر، چشمک‌زن، زجرکش، ماست‌بند، گوشتخوار، دستفروش، حقیقتگو، ژنده‌پوش، آوازه‌خوان، سودآور، حقیقت‌جو، چشم‌انداز، تندرو، خردکن، تمسخرآمیز، دیوارکوب، رقت‌انگیز، بازتاب، دورچین، خوشنویس، شهیدپرور، دندان‌شکن، فرح‌بخش، طناب‌پیچ، مردافکن، سرریز، زینت‌آرا، بیرون‌بر، جانگداز، سبک‌بار، گوش‌نواز، بلندپرواز، خوش‌نشین، زرافشان، دخترکش، بیابان‌گرد، خدمتگذار، پیکرتراش، خودجوش، پوست‌کن، لگدمال، طلایاب، دوراندیش، بیماری‌زا، گاه‌شمار، جنگ‌افروز، ؟، مشکل‌گشا، فریادرس، مهرگستر، خوش‌خواب، دریانورد، مشکل‌پسند، بچه‌دزد، لوکوموتیوران، دلفریب، خوش‌نام، چهره‌نگار، روح‌افزا، دست‌آموز، نمک‌پاش، خوش‌بو، ماردوش، تاشو، جان‌آفرین، هم‌تراز، پوزخند، ترانه‌سرا، رخت‌آویز، سگ‌دو، گاوچران، دردنوش، آهن‌ربا، ؟، یکه‌تاز، پیام‌رسان، خاک‌روب، سرانجام، چمن‌پیرا، شکرخا، آتشین‌دم، خلوت‌گزین، مردم‌آزار، آردبیزف بزخر، طاقت‌فرسا، دادورز، کج‌آگند، شادباش، خداترس، دلچسب، مال‌اندوز، تک‌چرخ، گندزدا، غم‌گسار، کم‌توان، علف‌چر، بخت‌آزما، دل‌آشوب، مگس‌پران، کینه‌توز، گوش‌خراش، ؟، خوش‌گوار، خون‌آشام
ی، اً، ک، کی، ه، ان، انه، وار	پسوند قیدساز	آخر عمری/آمدنی/برگشتنی، اخیراً/دائما، کم‌کمک/نم‌نمک، یواشکی/هول‌هولکی، امروزه، سحرگاهان، پیروزمندانه، دیوانه‌وار
بر، در، باز، فرا، فرو	پیشوندهایی که در معنی فعل تغییر ایجاد می‌کند	برداشت، درافتاد، بازداشت، فراگرفت، فرورفت

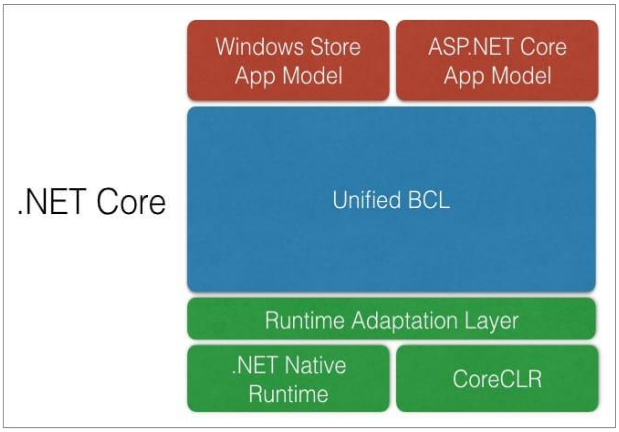
پس از این مرحله، متون مقاله‌های ویرایش شده، مجددا برچسب‌گذاری شد و در نهایت، برچسب‌ها مجدداً دستی کنترل شدند. به این ترتیب تا حد امکان برچسب‌های نادرست، اصلاح شدند. البته کماکان درصـدی از خطا وجود دارد. به‌عنوان نمونه، متن برچسب‌گذاری‌شدهٔ زیر (شکل ۲) نشان می‌دهد، در هر ۳۰۰ واژه، ممکن است، چهار یا پنج خطای برچسب‌گذاری وجود داشته باشد.

شکل ۲: نمونه‌ای از متن برچسب‌گذاری شده در پیکره

پس از ساخت پیکره، پیکره در سامانه‌ای قرار داده شد که دارای ویژگی‌های فنی و قابلیت‌های زیر است:

چارچوب اصلی سامانه

چارچوب اصلی سامانه، Net Core. (دات نت کور) می‌باشد: پلت‌فرمی است چند منظوره برای توسعه و برنامه‌نویسی که توسط مایکروسافت و مجموعه Net. در GitHub تهیه شده است. چارچوب Net Core. کاملاً cross-platform طراحی شده است، به شکلی که در ویندوز، لینوکس و مک قابل استفاده بوده است. همچنین برای دستگاه‌های مختلف، فضاهای ابری و سناریوهای مرتبط با اینترنت اشیاء کاربرد دارد. شکل ۲، چارچوب اصلی Net Core. را نشان می‌دهد.



شکل ۲: چارچوب اصلی Net Core

زبان‌های برنامه‌نویسی

در طراحی و پیاده‌سازی سامانه از زبان‌های برنامه‌نویسی زیر استفاده شده است:

- C# (سی‌شارپ[1]) در این سامانه برای توسعهٔ سمت سرور (Back End) از این زبان برنامه‌نویسی استفاده شده است. این زبان در بطن چارچوب Net Core. قرار دارد.
- HTML[2] (اچ‌تی‌ام‌ال)
- CSS[3] (سی‌اس‌اس)
- Javascript: در سرتاسر سامانه به‌صورت گسترده از کدهای JavaScript استفاده شده است. این زبان به همراه چارچوب[4] Jquery و Angular بخش اصلی کدهای Client Side (سمت کاربر) را تشکیل می‌دهد.

تکنولوژی‌های استفاده شده در سامانه

در سامانه از تکنولوژی‌های زیر برای توسعه و پیاده‌سازی استفاده شده است:

- Jquery
- AngularJs
- Bootstrap
- Linq[5]
- Entity Frameword
- SQL server

سامانه از طریق تکنولوژی Entity Framework در Net Core. از امکانات SQL Server استفاده می‌کند. در واقع این تکنولوژی استفاده از دستورات SQL را از طریق زبان Linq در خود کد سامانه امکان‌پذیر می‌کند.

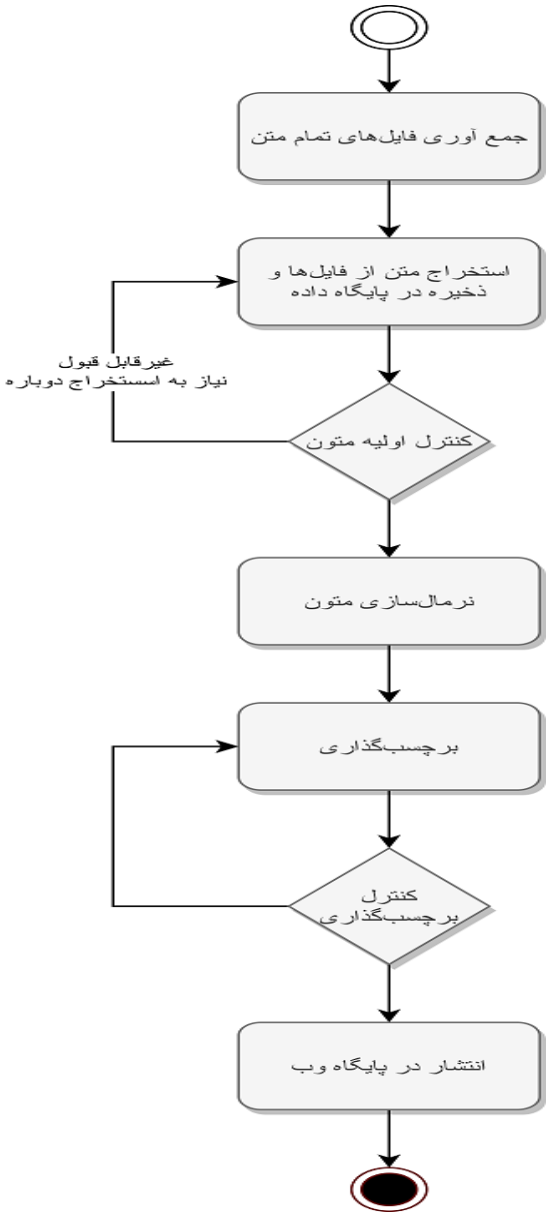
معماری سامانه

- توسعه MVC

الگوی MVC مخفف سه کلمه Model (مدل)، View (نمایشگر) و controller (کنترل‌کننده) است. در واقع MVC بر روی معماری‌های چندلایه‌ای برای تفکیک بخش‌های مختلف برنامه (بخش‌های منطقی برنامه مانند داده‌ها، مجوزها، کنترل صحت داده‌ها و لایه‌های مرتبط با کاربر نهایی) قرار می‌گیرد.

شکل ۳، نمودار جریانی (block diagram) از ماجول‌های ساخت پیکره را نشان می‌دهد.

1. C sharp
2. Hyper Text Markup Language (HTML)
3. Cascading Style Sheets (CSS)
4. framework
5. Language-Integrated Query



شکل ۳: نمودار جریانی (block diagram) از ماجول‌های ساخت پیکره

نتیجه‌گیری

در این پژوهش، پیکرهٔ متنی ساخته شد که محتوای آن متون مقاله‌های موجود در «پـژوهش‌نامـهٔ پـردازش و مدیریت اطلاعات» است. در مرحلهٔ نمونه‌گیری، متن بیش از۶۰۰ مقاله که فایل ورد (word) آنها موجود بود،

وارد پیکره شد. علاوه بر وارد کردن متن مقاله‌ها در پیکره، فرادادهٔ مربوط به هر مقاله (شامل عنوان مقاله، نام نویسنده/ نویسندگان و موضوع مقاله) نیز وارد پیکره شد. پس از وارد کردن داده‌ها و فراداده‌ها در پیکرهٔ متنی، داده‌ها به برنامهٔ نرمال‌سازی و سپس، برچسب‌گذاری منتقل شد. ابتدا نرمال‌سازی به‌صورت ماشینی انجام شد. همچنین برای یکدست کردن انواع «ی» و «ک» و «کسره اضافه روی «ه» (مانند «هٔ») از دستورات TSQL در برنامه پایگاه داده SQL Server استفاده شده است. پس از ساخت پیکرهٔ متنی، با توجه به کاربرد برچسب اجزای واژگانی کلام در پردازش متن، این نوع برچسب‌گذاری هم به پیکره اضافه شد. در این راستا ابتدا از یکی از ابزارهای آماده برای برچسب‌گذاری ماشینی اجزای واژگانی کلام (ابزار «هضم») استفاده شد و سپس، برچسب‌ها دستی کنترل شوند. در نهایت، برچسب‌ها کنترل شدند و تا حد امکان از طریق انجام اصلاحات روی کدهای ابزارهای نرمال‌سازی و کنترل دستی، برچسب‌های غلط اصلاح شدند.

منابع

- اسلامی، محرم؛ شریفی آتشگاه، مسعود؛ علیزاده لمجیری، صدیقه؛ زندی، طاهره (۱۳۸۳). «واژگان زایای زبان فارسی»، *مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه.*
- اسلامی، محرم؛ شیخ‌زادگان، جواد؛ احمدی‌نیا، زهرا؛ بهرامی راد، علی (۱۳۸۸). «مراحل و نحوهٔ تهیهٔ دادگان‌های صوتی هجایی و دایفونی برای سامانهٔ تبدیل متن به گفتار فارسی»، *دوفصل‌نامه علمی-پژوهشی پردازش علائم و داده‌ها*، ۲: ۳-۱۲.
- دشتبانی، شکوفه؛ منصوری‌زاده، محرم؛ نصیری، محمد (۱۳۹۳). «پیکرهٔ متنی تطبیقی فارسی-انگلیسی حوزهٔ تخصصی فاوا»، *پژوهش‌های زبان‌شناسی تطبیقی*. سال چهارم. ۸: ۱۴۱-۱۲۱.
- صادقی، علی‌اشرف (۱۳۷۰). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱)»، *نشر دانش*، شمارهٔ ۶۴.
- صادقی، علی‌اشرف (۱۳۷۰). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۲)»، *نشر دانش*، شمارهٔ ۶۵.
- صادقی، علی‌اشرف (۱۳۷۰). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۳)»، *نشر دانش*، شمارهٔ ۶۷.
- صادقی، علی‌اشرف (۱۳۷۱). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۴)»، *نشر دانش*، شمارهٔ ۶۹.
- صادقی، علی‌اشرف (۱۳۷۱). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۵)»، نشر دانش، شمارهٔ ۷۰.
- صادقی، علی‌اشرف (۱۳۷۱). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۶)»، *نشر دانش*، شمارهٔ ۷۱.
- صادقی، علی‌اشرف (۱۳۷۱). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۷ »، *نشر دانش*، شمارهٔ ۷۲.
- صادقی، علی‌اشرف (۱۳۷۱). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۸) »، *نشر دانش*، شمارهٔ ۷۴.
- صادقی، علی‌اشرف (۱۳۷۲). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۹)»، *نشر دانش*، شمارهٔ ۷۵.
- صادقی، علی‌اشرف (۱۳۷۲). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۰)»، *نشر دانش*، شمارهٔ ۷۶.
- صادقی، علی‌اشرف (۱۳۷۲). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۱)»، *نشر دانش*، شمارهٔ ۷۷.

- صادقی، علی‌اشرف (۱۳۷۲). «شیوه‌ها و امکانات واژه‌سازی در زبان فارسی معاصر (۱۲)»، *نشر دانش*، شمارهٔ ۷۹ و ۸۰.
- عاصی، مصطفی (۱۳۸۴). «گزارش کوتاهی از شکل‌گیری پایگاه داده‌های زبان فارسی در اینترنت»، *مجله پژوهشگران شماره ۲، صفحه ۱۳.*
- علایی، الهام؛ علیدوستی، سیروس (۱۳۹۹). «ساخت پیکرهٔ متنی: طراحی مدل امکان‌سنجی»، *پژوهش‌های زبان-شناسی تطبیقی*، سال دهم. ۲۰: ۳۰۹-۲۷۹.
- قطره، فریبا (۱۳۸۶). «مشخصه‌های تصریفی در زبان فارسی امروز»، *دستور*. ۳: ۸۱-۵۲.
- کشانی، خسرو (۱۳۷۱). *اشتقاق پسوندی در زبان فارسی امروز*، تهران، مرکز نشر دانشگاهی.
- لازار، ژیلبر (۱۳۸۹). *دستور زبان فارسی معاصر*. ترجمهٔ مهستی بحرینی و توضیحات و حواشی هرمز میلانیان، تهران، انتشارات هرمس. چاپ دوم.
- میرزایی، آزاده؛ مولودی، امیرسعید (۱۳۹۳). «نخستین پیکرهٔ نقش‌های معنایی در زبان فارسی»، *علم زبان*، ۲ (۳): ۴۸-۲۹.
- هضم برای پردازش زبان فارسی در پایتون: https://www.sobhe.ir/hazm/
- AleAhmad, A; Amiri, H; Darrudi, E; Rahgozar, M; Oroumchian. F (2009). *"Hamshahri: A Standard Persian Text Collection", Knowledge-Based Systems*, Elsevier, Dubai, 22(5): 382–387.
- Atkins, S; Clear, J; Ostler, N (1992). "Corpus design criteria", *Literary and Linguistic Computing*. 7 (1): 1-16
- Bijankhan, M.; Sheykhzadegan, J; Bahrani, M; Ghayoomi, M (2011). "Lesson from building a Persian written corpus: Peykare", *Language resources and evolution* 45 (2): 143-164. Springer.
- Bijankhan, M.; Sheykhzadegan, J; Roohani, M. R; Zarrintare, R; Ghasemi, S. Z; Ghasedi, M. E (2003). "Tfarsdat - The Telephone Farsi Speech Database", In Proceeding of *EUROSPEECH*, 1525-1528, Geneva, Switzerland.
- Claude Toriida, M (2016). "Steps for creating specialized corpus and developing an annotated frequence-based vocabulary list", *TESL Canada journal/ revue TESL du Canada* 34 (11): 87-105.
- Durand, J; Gut, U; Kristoffersen, G (2014). *The handbook of corpus phonology*, Oxford University Press (OUP).
- Ghayoomi, M.; Momtazi, S; Bijankhan, M (2013). "A study of corpus development for Persian", *International journal on Asian language processing* 20 (1): 17-33.
- Leech, G (2004). *Developing Linguistic Corpora: a Guide to Good Practice*. *adding linguistic annotaion. Edited by Martin Wynne .ahds.literature*, languages and linguistics. The Oxford Text Archive.
- Megerdoomian, K (2004). "Developing a Persian part-of-speech tagger", Proceedings of *the 1st workshop on Persian language and computer*. 99-105.
- Rasooli, M.; Kouhestani, M.; Moloodi, A (2013). "Development of a Persian Syntactic Dependency Treebank", *In The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*: 306-314. Atlanta, USA.

- Samvelian, P; Faghiri, P (2013). "Introducing PersPred, A Syntactic and Semantic Database for Persian Complex Predicates", In Proceedings of *the 9th Workshop on Multiword Expressions, Atlanta, Georgia, USA. Association for Computational Linguistics*, 11-20.
- Shamsfard, M.; Hesabi, A; Fadaei, H; Mansoory, N; Famian, A; Bagherbeigi, S; Fekri, E; et al (2010). "Semi-Automatic Development of Farsnet; the Persian Wordnet", Proceedings of *5th Global WordNet Conference* (GWA). Mumbai, India.
- Waynne, M (2005). *Developing linguistic corpora: a guide to good practice.* Oxbow books. Literary and linguistic computing 22 (1).