نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

طراحی و ساخت پیکره‌ی متنی برای حوزه‌ی تخصصی فاوا

شکوفه دشتبانی[۱]، محرم منصوری‌زاده[۲] و محمد نصیری[۳]
sh.dashtbani@basu.ac.ir[۱]
mansoorm@basu.ac.ir [۲]
m.nassiri@basu.ac.ir [۳]

چکیده

در زبانشناسی پیکره انباره‌ای از داده‌های متنی است که برای اهداف مختلفی مثل مطالعات فرهنگی یک زبان خاص، مطالعه تغییرات یک زبان با گذشت زمان، پروژه‌های پردازش زبان‌های طبیعی، پروژه‌هایی که مربوط به حوزه‌ی زبان شناسی است و غیره، ایجاد می شوند. در این مقاله تمرکز ما بر طراحی و ساخت پیکره‌ی دو زبانه‌ی فارسی-انگلیسی حوزه‌ی فاوا است. این پیکره به صورت خودکار ساخته شده است و منابع آن اسناد تخصصی حوزه‌ی فاوا است. ما نرم‌افزاری برای ساخت پیکره طراحی کرده‌ایم که هزینه و مدت زمان ساخت پیکره را کاهش می-دهد علاوه بر این نرم‌افزار ارائه شده قابلیت مدیریت پیکره را نیز برای کاربران فراهم می‌کند. از ویژگی‌های پیکره‌ی ساخته شده فراهم کردن یک مجموعه متمرکز از اسناد تخصصی است که می‌تواند در پروژه‌های مختلف حوزه‌ی فاوا استفاده شود. در انتها آماری از وضعیت فعلی پیکره‌ی فاوا ارائه می‌کنیم.

کلمات کلیدی

پیکره حوزه‌ی فاوا، مدیریت پیکره ، زبانشناسی رایانشی

۱- مقدمه

در علم زبان شناسی، شاخه‌ای به نام زبان شناسی پیکره‌ای [1] وجود دارد که هدف آن مطالعات علمی بر روی زبان طبیعی است. در تحقیقات و مطالعات زبان شناسی که بر روی یک حوزه‌ی خاص صورت می‌گیرد، لازم است که داده‌های حوزه‌ای که مورد مطالعه قرار گرفته است از نمونه‌های طبیعی باشند. پس از گردآوری این مجموعه می توان آن را در حوزه‌ی مورد مطالعه برای تحلیل و توصیف زبان استفاده کرد. به مجموعه‌ای از داده-های متنی پیکره[۱] می‌گویند. پیکره‌ها سفارش‌های مختلفی را می توانند بپذیرند، بدین معنا که می توانند حاوی داده‌های گفتاری، مقوله دستوری خاص مانند فعل، محاورات عامیانه و غیره باشند. در صورتی که مجموعه‌ای از داده‌های قابل اطمینان برای محققان فراهم باشد، نتایج دقیق تری حاصل خواهد شد. به عبارت دیگر پیکره مجموعه‌ای از داده‌های متنی سازمان یافته است. پیکره ممکن است که حاوی متن‌ها، نقل قول‌ها، فهرست‌ها و یا حتی لغات باشد. پیکره‌ها برای اهداف مختلفی ایجاد می شوند، تمامی پروژه‌های پردازش زبان‌های طبیعی، پروژه‌هایی که مربوط به حوزه‌ی زبان شناسی هستند و ... از پیکره استفاده می‌کنند. علاوه بر این ، برای ایجاد پایگاه داده‌های زبانی، از پیکره‌های آن زبان استفاده می گردد.

تاکنون پیکره‌های زیادی به زبان انگلیسی ایجاد شده‌اند. از جمله پیکره‌های انگلیسی معروف می‌توان پیکره‌ی Penn [2] را نام برد. این پیکره یک پیکره حاشیه نویسی[۲] شده مشهور است و هنوز هم دارای خطاهای نشانه‌گذاری است که رفع نشده است. این پیکره حاوی ۴,۵ میلیون کلمه به زبان انگلیسی آمریکایی است. پیکره‌ی آکسفورد[3] (OEC) یک پیکره بسیار حجیم به زبان انگلیسی است. این پیکره برای ساخت فرهنگ لغت آکسفورد استفاده شده است. پیکره Brown [4] نیز به زبان انگلیسی است که در دانشگاه brown تهیه شده است و حاوی ۱ میلیون لغت است. از جمله پیکره‌های معروف دیگر در زبان انگلیسی می‌توان پیکره BNC[۳][5] را نام برد که به زبان انگلیسی بریتانیایی است و حاوی ۴۰۰۰ داده‌ی متنی و داده‌ی صوتی است. لغات پیکره بیشتر از ۱۰۰ میلیون کلمه است. داده‌های آن مربوط به قرن ۲۰ به بعد است. آخرین نسخه این پیکره در سال ۲۰۰۷ منتشر شده است. ۹۰ درصد از داده‌های این پیکره، داده‌های نوشتاری و ۱۰ درصد آن هم داده‌های صوتی هستند. BNC یکی از پیکره‌های مهم زبان انگلیسی است. پیکره‌های دیگری از قبیل COCA [6] ، ANC [7] و غیره نیز ایجاد شده‌اند. برای زبان فارسی پیکره دکتر محمود