به نام خدا

وزارت علوم، تحقیقات، و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران

(ایرانداک)

خلاصهٔ گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

ساخت پیکرهٔ متنی از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

تاریخ شروع طرح پژوهشی	۱۳۹۸/۱۰/۲۸	**تاریخ پایان طرح پژوهشی**	۱۳۹۹/۱۲/۱۱

نام و نام خانوادگی مجری طرح پژوهشی	الهام علایی ابوذر
نام و نام خانوادگی همکاران طرح پژوهشی	نصرالله پاک‌نیت، علی‌اصغر حجت‌پناه، مجتبی زالی، محمدهادی آقالوی

دربارهٔ طرح پژوهشی

پیکره، مجموعه‌ای نظام‌مند، رایانه‌ای شده، معتبر و موثق از زبان است که برای مطالعات زبان‌شناسی مورد استفاده قرار می‌گیرد. هدف از پژوهش حاضر ساخت پیکره‌ای با استفاده از متون مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات بوده است. تهیهٔ یک پیکرهٔ زبانی مبتنی بر داده‌های علمی ایرانداک از دو نظر حائز اهمیت است: ۱- ایرانداک را در شمار پژوهشگاه‌ها و دانشگاه‌هایی قرار می‌دهد که پیکرهٔ زبانی مفید (مبتنی بر داده‌های علمی معتبر) تهیه کرده‌اند ۲- می‌توان بسیاری از پژوهش‌های حوزهٔ پردازش متن (مانند انواع طبقه‌بندی داده‌ها، استخراج اطلاعات گوناگون از متن و غیره) با استفاده از این پیکره انجام داد. در این پژوهش به این پرسش‌ها پرداخته شده است: چه حجمی از مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات را می‌توان در پیکره وارد کرد؟ از چه ابزارهایی می‌توان برای پیش‌پردازش متون فارسی در پیکره استفاده کرد؟ و از چه ابزارهایی می‌توان برای حاشیه-نویسی استفاده کرد و چگونه می‌توان برچسب‌گذاری ماشینی را کنترل کرد؟

روش پژوهش

در مرحلهٔ نمونه‌گیری، متن ۶۷۷ 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 از طریق کاربرگه ثبت مقالات وارد پیکره شد. پس از وارد کردن داده‌ها و فراداده‌ها در پیکرهٔ متنی ایرانداک، از طریق خروجی گرفتن از پایگاه داده SQLSERVER داده‌ها به برنامهٔ نرمال‌سازی و سپس، برچسب‌گذاری منتقل شد. ابتدا نرمال‌سازی به صورت ماشینی انجام شد. برای یکدست کردن انواع «ی» و «ک» و «کسره اضافه روی «ه» (مانند «هٔ») از دستورات TSQL در برنامه پایگاه داده SQKServer نیز استفاده شد. یکی از پرکاربردترین نوع برچسب‌گذاری، برچسب‌گذاری اجزای واژگانی کلام(POS) است. پس از ساخت پیکرهٔ متنی ایرانداک، با توجه به کاربرد برچسب اجزای واژگانی کلام در پردازش متن، این نوع برچسب‌گذاری هم به پیکره اضافه شد. در این راستا تصمیم گرفته شد ابتدا از یکی از ابزارهای آماده برای برچسب‌گذاری ماشینی اجزای واژگانی کلام استفاده شود و سپس، برچسب‌ها دستی کنترل شوند. ابزار استفاده شده برای برچسب‌گذاری ماشینی، ابزار «هضم» است. فهرست برچسب-ها، شامل برچسب‌هایی است که در مقالهٔ بی‌جن‌خان و همکاران (۲۰۱۱) معرفی شده است. در نهایت، برچسب‌ها چک شدند و تا حد امکان از طریق انجام اصلاحات روی کدهای ابزارهای نرمال‌سازی و کنترل دستی، برچسب-های غلط اصلاح شدند.

یافته‌های طرح پژوهشی

انجام بسیاری از پژوهش‌های زبان‌شناسی و تصمیم‌گیری‌ها در برنامه‌ریزهای زبانی، تنها با استفاده از یک پیکرهٔ زبانی امکان‌پذیر است. در پژوهش حاضر پیکره‌ای ساخته شده است که شامل متون مقاله‌های موجود در پژوهش-نامهٔ پردازش و مدیریت اطلاعات است. ۶۷۷ مقاله و ۴۷۷۴۷۹۰ واژه. موضوع این مقاله‌ها شامل کتابداری و اطلاع-رسانی، علم اطلاعات و دانش‌شناسی، فناوری اطلاعات، مدیریت اطلاعات، مدیریت دانش، زبان‌شناسی، زبان‌شناسی رایانشی، اصطلاح‌شناسی، هستان‌شناسی و سایر حوزه‌های مرتبط با پردازش اطلاعات است. بنابراین، محتوای این پیکره عمومی نیست و حاوی متون بسیار تخصصی و میان‌رشته‌ای است و از این نظر برای پردازش‌هایی که مستلزم بهره‌گرفتن از متون تخصصی است، بسیار ارزشمند است. پس از نمونه‌گیری و وارد کردن متون مقاله‌ها در پیکره، فرادادهٔ مربوط به مقاله‌ها (مانند عنوان مقاله، نام نویسنده/نویسندگان و موضوع مقاله) نیز از طریق کاربرگه ثبت

مقالات وارد پیکره شد. سپس، ابتدا نرمال‌سازی ماشینی و به دنبال آن، برچسب‌گذاری ماشینی (نوعاً برچسب‌گذاری اجزای واژگانی کلام (POS)) انجام شد و اطلاعات در پایگاه داده ذخیره شده است و در نهایت، برچسب‌ها دستی کنترل شده‌اند. همچنین در این طرح پژوهشی، در کنار ساخت پیکره، سامانه‌ای طراحی شد به نام «سامانهٔ پیکره‌های ایرانداک» که پیکرهٔ متنی ایرانداک در این سامانه قرار گرفت و در آینده می‌توان پیکره‌های دیگری را نیز در این سامانه قرار داد.

نتیجه‌گیری و پیشنهادها

در پژوهش حاضر پیکره‌ای ساخته شده است که شامل متون مقاله‌های موجود در پژوهش‌نامهٔ پردازش و مدیریت اطلاعات است. بیش از ۶۰۰ مقاله و بیش از چهار میلیون واژه. این متون هم نرمالایز شدند و هم برچسب‌گذاری اجزای واژگانی کلام در مورد آن‌ها انجام گرفت و در نهایت برچسب‌ها دستی کنترل شدند و با استفاده از استخراج الگوهایی برای خطاهای برچسب‌گذاری، برچسب‌های غلط، ماشینی و دستی اصلاح شدند. در این پژوهش پیشنهادات زیر برای پژوهش‌های آینده ارائه گردید:

- با توجه به نیاز تشخیص افعال مرکب در ترجمهٔ ماشینی مبتنی بر پیکره، می‌توان ابتدا از میان رویکردهای رایج برای تشخیص افعال مرکب در زبان فارسی، یکی را انتخاب نمود و در پیکرهٔ متنی ایرانداک به جست‌وجوی افعال مرکب (جدانشدنی و جداشدنی) پرداخت و برچسب «فعل مرکب/ Compound Verb (CV) را به آن‌ها اختصاص داد.
- با توجه به کاربرد برچسب معنایی در بسیاری از حوزه‌های پردازش زبان طبیعی، مانند استخراج اطلاعات، پرسش و پاسخ، خلاصه‌سازی و ترجمهٔ ماشینی، می‌توان برچسب‌گذاری نقش معنایی را نیز به پیکرهٔ متنی ایرانداک افزود.
- با توجه به موجود بودن چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و مقاله‌های پژوهش‌نامهٔ پردازش و مدیریت اطلاعات در ایرانداک، می‌توان با بررسی ملاحظات حقوقی، پیکره‌ای موازی حاوی چکیده‌های فارسی و انگلیسی مقاله‌ها یا پایان‌نامه‌ها ساخت و به سامانهٔ پیکره‌های ایرانداک اضافه کرد.