

به نام خدا



وزارت علوم، تحقیقات و فناوری

پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)

خلاصه گزارش طرح پژوهشی سازمانی

عنوان طرح پژوهشی

ساخت پیکره از کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

تاریخ شروع طرح پژوهشی	۱۴۰۰/۰۷/۱۴	تاریخ پایان طرح پژوهشی	۱۴۰۱/۰۹/۱۳
نام و نام خانوادگی مجری طرح پژوهشی	الهام علایی ابوذر		
نام و نام خانوادگی همکاران طرح پژوهشی	علی اصغر حجت پناه		

درباره طرح پژوهشی

پیکره‌های زبانی کاربردهای گوناگون دارند، و می‌توان گفت، بسیاری از پژوهش‌های زبان‌شناسی و تصمیم‌گیری‌ها در برنامه‌ریزی زبانی، با استفاده از پیکره‌های زبانی انجام می‌شود. پیکره انواع گوناگونی دارند مانند پیکره‌های عمومی، تخصصی، مقایسه‌ای، موازی و غیره. تا کنون پیکره‌های عمومی بسیاری در زبان‌های گوناگون، از جمله زبان فارسی، ساخته شده است. اما تعداد پیکره‌های تخصصی که برای اهداف خاص و پردازش‌های خاص زبانی استفاده می‌شود، به اندازه پیکره‌های عمومی نیست. در طرح پژوهشی پیشین اینجانب (علایی و همکاران: ۱۴۰۰) تحت عنوان «ساخت پیکره متنی از مقاله‌های پژوهش‌نامه پردازش و مدیریت اطلاعات»، سامانه‌ای ساخته شد به نام «سامانه پیکره‌های ایرانداک» که «پیکره پژوهش‌نامه» در این سامانه قرار گرفت. در این طرح پژوهشی، پیکره تخصصی دیگری به سامانه اضافه شده است تا امکانات بیشتری برای جامعه علمی برای پژوهش‌هایی که مستلزم بهره‌گرفتن از پیکره‌های تخصصی است، فراهم شود؛ این پیکره از متون کتاب‌های دیجیتال ایرانداک ساخته شده است.

هدف اصلی:

- ساخت و ایجاد پیکره کتاب‌های دیجیتال ایرانداک (پکا) در سامانه پیکره‌های ایرانداک

اهداف فرعی:

- شناسایی ابزارهای کارآمد پیش‌پردازش متون
- شناسایی قابلیت‌های مورد نیاز برای افزودن به سامانه پیکره‌های ایرانداک

روش پژوهش

این پژوهش در یک گام انجام شده است.

پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) از سال ۱۳۴۸ تا سال ۱۴۰۰، تعداد ۲۸۱ کتاب منتشر کرده است. در فرآیند نمونه-گیری، با توجه به محتوای متون کتاب‌ها و تاریخ انتشار کتاب‌ها، از ۲۸۱ کتاب دیجیتال ایرانداک، ۶۵ کتاب برای وارد شدن در پیکره انتخاب شدند. کتاب‌ها از نظر تکراری نبودن بررسی شدند. بسیاری از کتاب‌های انتخاب شده برای پیکره، به صورت فایل پی.دی.اف بودند که از طریق ابزار اُ.سی.آر به فایل‌های ورد تبدیل شدند و سپس به صورت دستی و ماشینی تمیز و مرتب کردن متون و اصلاحات نگارشی روی متون انجام گرفت و برای مراحل بعدی آماده شدند. مراحل ساخت این پیکره به صورت زیر بوده است:

- نمونه‌گیری
- تهیه فراداده‌ها
- فرآیند تبدیل فایل‌های پی.دی.اف به ورد از طریق استفاده از ابزار اُ.سی.آر
- تمیز و مرتب کردن فایل‌های خروجی
- هنجارسازی (نرمال‌سازی) داده‌ها
- برچسب‌گذاری داده‌ها (برچسب‌گذاری اجزای واژگانی کلام)
- کنترل دستی برچسب‌ها
- ارتقاء سامانه پیکره‌های ایرانداک (سایا)

یافته‌های طرح پژوهشی

یکی از روش‌های گردآوری منسجم داده‌ها، تهیه پیکره‌های زبانی است. تا کنون پیکره‌های عمومی بسیاری در زبان‌های گوناگون، از جمله زبان فارسی، ساخته شده است. اما تعداد پیکره‌های تخصصی که برای اهداف خاص و پردازش‌های خاص زبانی استفاده می‌شود، به اندازه پیکره‌های عمومی نیست. بنابراین، ساخت پیکره‌هایی که شامل متون تخصصی و میان‌رشته‌ای است، بسیار کاربردی است. در این طرح پژوهشی پیکره‌ای ساخته شده است که محتوای آن متون کتاب‌های دیجیتال پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. برای ساخت این پیکره، پس از طی فرآیند نمونه‌گیری ۶۵ کتاب انتخاب شدند و فراداده مربوط به کتاب‌ها (عنوان مقاله، نام نویسنده/نویسندگان و موضوع مقاله) به طور جداگانه تهیه شد؛ نزدیک به نیمی از کتاب‌ها به صورت فایل پی.دی.اف (PDF) بودند که از طریق استفاده از ابزار اُ.سی.آر (OCR) به فایل ورد (WORD) تبدیل شدند و برای وارد شدن در پیکره ویرایش‌های نگارشی و ماشینی روی آن‌ها انجام گرفت. برای هنجارسازی داده‌های وارد شده در پیکره از ابزار هضم، اطلاعات ساختاری زبان فارسی (وندهای اشتقاقی و تصریفی و واژه‌بست‌ها) و برنامه‌نویسی ماشینی استفاده شد. پس از هنجارسازی، ابتدا برچسب‌گذاری ماشینی داده‌ها (برچسب‌گذاری اجزای واژگانی کلام (POS)) انجام شد و در نهایت برچسب‌ها به صورت دستی نیز کنترل شدند و اطلاعات در پایگاه داده ذخیره شد و پیکره در سامانه پیکره‌های ایرانداک (سایا) قرار گرفت. این پیکره شامل ۳۳۲۹۲۸۱ واژه است که حوزه‌های میان‌رشته‌ای کتابداری، مدیریت، زبان‌شناسی، مدیریت اطلاعات و غیره را پوشش می‌دهد.

نتیجه‌گیری و پیشنهادها

تا کنون پیکره‌های عمومی بسیاری در زبان‌های گوناگون، از جمله زبان فارسی، ساخته شده است. اما تعداد پیکره‌های تخصصی که برای اهداف خاص و پردازش‌های خاص زبانی استفاده می‌شود، به اندازه پیکره‌های عمومی نیست. بنابراین، ساخت پیکره‌هایی که شامل متون تخصصی و میان‌رشته‌ای است، بسیار ارزشمند است. گزارش این طرح پژوهشی شامل پنج فصل است: پس از مقدمه، در فصل دوم به بررسی پژوهش‌هایی پرداخته شد که در حوزه ساخت پیکره انجام شده است. در این پژوهش‌ها، چکیده‌ای از تجربیات پژوهشگران این حوزه درباره ساخت پیکره‌ها و چالش‌هایی که هنگام ساخت پیکره‌های متنی یا نوشتاری با آن‌ها مواجه شده‌اند، ارائه شده است. برخی دیگر از این پژوهش‌ها شامل توصیف انواع روش‌های ساخت پیکره، نوع پیکره ساخته شده و حاشیه‌نویسی‌های استفاده شده در پیکره است. در بخش انتهایی این فصل به معرفی برخی از پیکره‌های زبان نیز فارسی پرداخته شد. در فصل سوم ابتدا به چگونگی فرآیند نمونه‌گیری پرداخته شد. سپس، تهیه فراداده‌ها توضیح داده شد. در ادامه به فرآیند تبدیل فایل‌های پی.دی.اف به ورد از طریق استفاده از ابزار اُ.سی.آر پرداخته شد و در نهایت چالش‌های این روش و چگونگی تمیز و مرتب کردن فایل خروجی پرداخته شد و در فصل چهارم، چگونگی هنجارسازی داده‌های وارد شده در پیکره توضیح داده شد که در این راستا از ابزار هضم، اطلاعات ساختوازی زبان فارسی (وندهای اشتقاقی و تصریفی و واژه‌بست‌ها) و برنامه‌نویسی ماشینی استفاده شد. سپس مفهوم «حاشیه‌نویسی» توضیح داده شد و نوع حاشیه‌نویسی/برچسب‌گذاری استفاده شده در پیکره و ابزار استفاده شده برای این منظور، معرفی شد و در نهایت نحوه کنترل دستی برچسب‌ها توضیح داده شد.