

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 مقایسه‌ای تخصصی از چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌های (پارساهای) ثبت شده در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

تاریخ شروع طرح پژوهشی	۱۴۰۰/۱۰/۲۲	تاریخ پایان طرح پژوهشی	۱۴۰۲/۰۶/۲۱
نام و نام خانوادگی مجری طرح پژوهشی	الهام علایی ابوذر		
نام و نام خانوادگی همکاران طرح پژوهشی	علی اصغر حجت‌پناه		

درباره طرح پژوهشی

مجموعه داده‌هایی که به صورت سازمان‌یافته و رایانه‌ای شده گردآوری شده‌اند و برای تجزیه و تحلیل‌های زبان-شناسی مورد استفاده قرار می‌گیرند، پیکره نامیده می‌شود. پیکره‌ها بر اساس زبان به کار رفته در متون تشکیل دهنده آن‌ها به پیکره‌های تک‌زبانه، دوزبانه و چندزبانه تقسیم می‌شوند. پیکره‌های دوزبانه یا چندزبانه با دو نام شناخته می‌شوند: پیکره‌های مقایسه‌ای/تطبیقی و پیکره‌های موازی. پیکره مقایسه‌ای، پیکره‌ای است دوزبانه یا چندزبانه که شامل متونی است مشابه در حوزه‌های موضوعی یکسان. پیکره‌های مقایسه‌ای می‌تواند از متون عمومی ساخته شود که امکانات گوناگونی را برای تحلیل گفتمان، کاربردشناسی، تجزیه و تحلیل ژانرهای متون و زبان‌شناسی اجتماعی فراهم می‌کند؛ نمونه‌هایی از چنین پیکره‌هایی می‌تواند شامل مجموعه مدخل‌های دایرةالمعارف‌ها یا متون ادبی یک دوره زمانی خاص باشد. اما متداول‌ترین نوع پیکره‌های مقایسه‌ای که مخاطبان بسیاری دارد، آن‌هایی هستند که به حوزه(های) تخصصی مربوط می‌شوند و دارای تراکم بالای واژگان و اصطلاحات تخصصی هستند. چنین پیکره‌هایی، پیکره تخصصی مقایسه‌ای نامیده می‌شوند. برخی از کاربردهای چنین پیکره‌هایی شامل پژوهش‌های تطبیقی زبان‌ها، داده‌کاوی، موتورهای جستجوی دوزبانه، پژوهش‌های بین‌زبانی، ترجمه ماشینی، فرهنگ‌نگاری محاسباتی و

بازیابی اطلاعات است. با توجه به کاربرد پیکره‌های مقایسه‌ای در ترجمه ماشینی و تعداد اندک این پیکره‌ها در زبان فارسی، لازم است پیکره‌های مقایسه‌ای با حجم بالای واژگان تخصصی در زبان فارسی ساخته شود. پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) مرکز ثبت و تنها آرشیو ملی اطلاعات پایان‌نامه‌ها، رساله‌ها، و پیشنهاد آن-هاست. بنابراین، منبع غنی چکیده‌های فارسی و انگلیسی پایان‌نامه‌ها و رساله‌ها است که می‌تواند برای ساخت پیکره-های تخصصی مورد استفاده قرار گیرد. در این طرح پژوهشی پیکره تخصصی مقایسه‌ای به نام «پارسا» ساخته شده است و در کنار دو پیکره دیگر (پژوهش‌نامه و پکا) در سامانه پیکره‌های ایرانداک (ساپا) بارگزاری شده است.

روش پژوهش

برای انجام این طرح پژوهشی، پس از بررسی پژوهش‌های انجام‌شده در حوزه ساخت پیکره‌های عمومی و تخصصی و مقایسه‌ای و چالش‌هایی که هنگام ساخت پیکره‌های متنی / نوشتاری با آن‌ها مواجه بوده‌اند، چارچوب اصلی ساخت این پیکره تخصصی مقایسه‌ای (پارسا) مشخص شد که شامل چند مرحله است: نمونه‌گیری، پیش‌پردازش، برچسب‌گذاری و کنترل برچسب‌ها. برای نمونه‌گیری از سامانه ویرایش دو فیلتر مدنظر قرار گرفته شد: فایل‌ها چکیده انگلیسی نیز داشته باشند و حداقل تعداد واژه ۲۰۰ و حداکثر ۲۰۰۰ باشد. حدوداً ۲۹۵ هزار فایل چکیده انتخاب شدند که معادل انگلیسی نیز داشتند. سپس روی داده‌های فارسی پیش‌پردازش‌هایی انجام گرفت (هنجارسازی و واحدسازی)؛ در این مرحله هم از ابزار «هضم» و هم اطلاعات ساختواری فارسی استفاده شد. با این حال بسیاری از واژه‌ها در فایل‌های فارسی به هم چسبیده بودند که به صورت دستی و ماشینی اصلاح شدند و سعی شد تا حد امکان مشکلات ویرایشی در فایل‌های فارسی برطرف گردد. سپس، برچسب‌گذاری اجزای واژگانی کلام (POS) در مورد داده‌های فارسی و انگلیسی انجام گرفت و برچسب‌های داده‌های فارسی کنترل شد و در نهایت پیکره در سامانه پیکره‌های ایرانداک (ساپا) قرار گرفت. پیکره تهیه شده شامل بیش از ۸۹ میلیون واژه فارسی و ۷۹ میلیون واژه انگلیسی است. در کنار ساخت پیکره، کارهایی نیز برای ارتقاء سامانه پیکره‌های ایرانداک انجام گرفت.

یافته‌های طرح پژوهشی

پیکره تهیه شده شامل بیش از ۸۹ میلیون واژه فارسی و ۷۹ میلیون واژه انگلیسی است. برچسب‌های فارسی همراه با فراوانی آن‌ها در این پیکره به صورت زیر است:

ADV	682723
ADVe	96534
AJ	7617693
AJe	3106973
CL	119125
CONJ	6731744
CONJe	6797
DET	1733948
DETe	174172
INT	685
N	17383371
Ne	22483499
NUM	3380242
NUMe	94391
P	10276735
Pe	1416302
POSTP	620247
PRO	824305
PROe	8912
PUNC	4363936
RES	1563942
RESe	34642
V	6283020

تعداد واژه‌های محتوایی (فعل، اسم، صفت، قید)، ۵۷۶۵۳۸۱۳ است و تعداد واژه‌های دستوری به همراه اعداد و علائم سجاوندی شامل ۳۱۳۵۰۱۲۵ است. بن‌واژه‌های فارسی نیز استخراج شد؛ تعداد بن‌واژه‌ها ۴۱۰۶۴ است.

برچسب‌های انگلیسی به همراه فراوانی آن‌ها در این پیکره به صورت زیر است:

'	27743
#	5413
\$	28344
(	707427
)	721314

,	3373793
.	2965628
:	305307
``	34742
CC	3504811
CD	1707777
DT	6916770
EX	95839
FW	53067
IN	10179166
JJ	7551971
JJR	179109
JJS	135049
LS	6
MD	320426
NN	15833944
NNP	4678139
NNPS	27827
NNS	6025180
PDT	21819
POS	2164
PRP	527942
PRP\$	324839
RB	1403802
RBR	67710
RBS	93102
RP	38301
SYM	1039
TO	1349688
UH	3412
VB	1295483
VBD	1647568
VBG	1448143
VCN	2533053
VBP	1040006
VBZ	1646600
WDT	337079
WP	46615
WP\$	4472
WRB	57361

تعداد واژه‌های محتوایی متون انگلیسی (فعل، اسم، صفت، قید)، ۴۵۶۰۶۶۸۶ است و تعداد واژه‌های دستوری به همراه اعداد و علائم سجاوندی شامل ۳۳۶۶۲۳۰۴ است. بن‌واژه‌های انگلیسی نیز استخراج شد؛ تعداد بن‌واژه‌ها ۱۲۹۳۷ است.

نتیجه‌گیری و پیشنهادها

در پژوهش حاضر یک پیکره مقایسه‌ای تخصصی ساخته شده است که شامل متون فارسی و انگلیسی چکیده‌های پایان‌نامه‌ها و رساله‌های ثبت شده در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. این پیکره شامل بیش از ۸۹ میلیون واژه فارسی و ۷۹ میلیون واژه انگلیسی است. محتوای این پیکره عمومی نیست و حاوی متون بسیار تخصصی در حوزه‌های موضوعی کلان مانند علوم اجتماعی، علوم انسانی و هنر، فنی و مهندسی و رشته‌های مربوط به این حوزه‌ها است و از این نظر برای پردازش‌های زبانی که مستلزم بهره‌گرفتن از متون تخصصی است، بسیار ارزشمند است.

پیشنهادهای:

- با توجه به کاربرد برجسب‌های گفتمانی در بسیاری از حوزه‌های پردازش زبان طبیعی، مانند تحلیل احساس، استخراج اطلاعات، پرسش و پاسخ، خلاصه‌سازی و ترجمه ماشینی، می‌توان برجسب‌گذاری گفتمانی، مخصوصاً از نوع نگرشی را نیز به پیکره‌های تخصصی موجود در سامانه پیکره‌های ایرانداک (ساپا) اضافه نمود.

- با توجه به نیاز تشخیص افعال مرکب در ترجمه ماشینی مبتنی بر پیکره، می‌توان ابتدا از میان رویکردهای رایج برای تشخیص افعال مرکب در زبان فارسی، یکی را انتخاب نمود و در پیکره‌های متنی تخصصی ایرانداک به جست‌وجوی افعال مرکب (جدانشدنی و جداشدنی) پرداخت و برجسب «فعل مرکب/Compound Verb (CV) را به آن‌ها اختصاص داد.

- تبدیل پیکره مقایسه‌ای (پارسا) به پیکره موازی

- با توجه به کاربرد برجسب معنایی در بسیاری از حوزه‌های پردازش زبان طبیعی، مانند استخراج اطلاعات،

پرسش و پاسخ، خلاصه‌سازی و ترجمه ماشینی، می‌توان برچسب‌گذاری نقش معنایی را نیز به پیکره متنی
ایرانداک افزود.