به نام خدا

پژوهشکده : ارتباطات و فناوری اطلاعات

بررسی پیکره‌ها و ابزارهای پردازش زبان فارسی

پروژه: مدیریت طرح

کد پروژه: ۹۳۳۲۰۱۲

مجری:	**علیرضا یاری**
تهیه‌کننده:	**تیم مدیریت طرح**
کد گزارش:	**P-PD-VAS-SBM-S-006-1.02**
تاریخ ارائه:	**۹۴/۰۴/۱۴**
نسخه/وضعیت:	**اولیه**

شناسنامه گزارش

عنوان: بررسی پیکره‌ها و ابزارهای پردازش زبان فارسی		**شماره نسخه: ۱,۰۲**
کد: P-PD-VAS-SBM-S-006-1.02	**نوع گزارش: فنی**	**تاریخ ارائه گزارش:** ۹۴/۰۴/۱۴
نام پروژه: مدیریت طرح	**نوع پروژه:** پژوهشی - کاربردی	
تاریخ شروع: ۹۳/۱۱/۲۰	**تاریخ پایان:** ۹۶/۱۱/۲۰ (۳۶ ماه)	
نام گروه: طرح جویشگر		
کد پروژه: ۹۳۳۲۰۱۲	**شماره و تاریخ قرارداد:** ۹۳/۱۱/۲۰	
مجری: علیرضا یاری	**ناظر / ناظرین:** کامبیز بدیع، رامین شکری‌پور و روح‌اله رحمانی	
تهیه کننده/ تهیه کنندگان: تیم مدیریت طرح		
نام و نشانی مجری: تهران، انتهای خیابان کارگر شمالی، مرکز تحقیقات مخابرات ایران-کد پستی: ۱۴۳۹۹۵۵۴۷۱ - تلفن: ۱۰-۸۸۰۰۵۵۰۸		
نام و نشانی حمایت کننده: تهران، خیابان شریعتی، وزارت ارتباطات و فناوری اطلاعات		
ملاحظات: ندارد		

چکیده:

در این مستند ابزارهای پردازش زبان فارسی مورد بررسی قرار گرفته اند. در فصل اول از این مستند مترجم‌های ماشینی انگلیسی-فارسی موجود و توسعه یافته در کشور مورد بررسی قرار گرفته‌اند و از جنبه کیفیت رده بندی شده‌اند. در فصل دوم نیز ابزارهای خطایاب و یکسان ساز فارسی مورد مطالعه قرار گرفتند. در فصل سوم لیست پیکره‌های دو زبانه و تک زبانه فارسی-انگلیسی موجود ارائه شده‌اند. ابزارهای تبدیل متن به گفتار و همچنین ابزارهای وردنت موجود در زبان فارسی و موجود در جهان نیز تشریح شده است.

همانگونه که می‌دانیم یکی از اصلی‌ترین سرویس‌های یک جویشگر، سرویس ترجمه ماشینی می‌باشد. همچنین پردازش زبان طبیعی به خصوص زبان فارسی یکی از فعالیت‌های مهم این طرح در نظر گرفته شده‌است. در نتیجه این گزارش با هدف شناسایی وضعیت سرویس‌ها و منابع موجود در این حوزه ارائه شده‌اند. از جمله اهدافی که می‌توان برای این گزارش (در حوزه پردازش زبان طبیعی) در نظر گرفت عبارتند از:

- ✓ بررسی و شناسایی منابع و دادگان موجود در این حوزه

- ✓ تحلیل وضعیت موجود
- ✓ برنامه ریزی‌های آتی
- ✓ شناسایی نواقص و تلاش در جهت رفع آنها در هر بخش
- ✓ تحلیل نیازمندی‌ها در جهت توسعه و ارتقاء سرویس‌ها
- ✓ مقایسه و ارزیابی هریک از ابزارها و سیستم‌های موجود
- ✓ شناسایی شاخص‌ها و معیارهای ارزیابی هر بخش
- ✓ بررسی سطح تقاضا و میزان استقبال کاربران از هریک از سرویس‌ها
- ✓ شناسایی درجه اهمیت هریک از بخش‌ها و الویت بندی آنها

در انتهای هر فصل روال‌ها و شاخص‌های ارزیابی برای هریک ابزارها و سرویس‌های موجود ارائه شده‌است.

کلمات کلیدی: معیارهای ارزیابی، پردازش زبان طبیعی، ترجمه ماشینی، دادگان، وردنت

وضعیت گزارش: نهایی	**زبان گزارش: فارسی**
وضعیت دسترسی: عادی	**تعداد صفحات:** ۱۹

چکیده

در این مستند ابزارهای پردازش زبان فارسی مورد بررسی قرار گرفته اند. در فصل اول از این مستند مترجم‌های ماشینی انگلیسی-فارسی موجود و توسعه یافته در کشور مورد بررسی قرار گرفته‌اند و از جنبه کیفیت رده بندی شده‌اند. در فصل دوم نیز ابزارهای خطایاب و یکسان ساز فارسی مورد مطالعه قرار گرفتند. در فصل سوم لیست پیکره‌های دو زبانه و تک زبانه فارسی-انگلیسی موجود ارائه شده‌اند. ابزارهای تبدیل متن به گفتار و همچنین ابزارهای وردنت موجود در زبان فارسی و موجود در جهان نیز تشریح شده است.

همانگونه که می‌دانیم یکی از اصلی‌ترین سرویس‌های یک جویشگر، سرویس ترجمه ماشینی می‌باشد. همچنین پردازش زبان طبیعی به خصوص زبان فارسی یکی از فعالیت‌های مهم این طرح در نظر گرفته شده-است. در نتیجه این گزارش با هدف شناسایی وضعیت سرویس‌ها و منابع موجود در این حوزه ارائه شده‌اند. از جمله اهدافی که می‌توان برای این گزارش (در حوزه پردازش زبان طبیعی) در نظر گرفت عبارتند از:

- ✓ بررسی و شناسایی منابع و دادگان موجود در این حوزه
- ✓ تحلیل وضعیت موجود
- ✓ برنامه ریزی‌های آتی
- ✓ شناسایی نواقص و تلاش در جهت رفع آنها در هر بخش
- ✓ تحلیل نیازمندی‌ها در جهت توسعه و ارتقاء سرویس‌ها
- ✓ مقایسه و ارزیابی هریک از ابزارها و سیستم‌های موجود
- ✓ شناسایی شاخص‌ها و معیارهای ارزیابی هر بخش
- ✓ بررسی سطح تقاضا و میزان استقبال کاربران از هریک از سرویس‌ها
- ✓ شناسایی درجه اهمیت هریک از بخش‌ها و الویت بندی آنها

در انتهای هر فصل روال‌ها و شاخص‌های ارزیابی برای هریک ابزارها و سرویس‌های موجود ارائه شده‌است.

اطلاعات مرتبط

مستندات مرتبط

شماره مستند	نوع مستند	نام مستند

تغییرات اعمال‌شده در نسخه‌های پیشین

شماره نسخه	تاریخ	تغییرات اعمال شده
۱,۰۰	۹۴/۰۴/۱۴	آماده سازی گزارش
۱,۰۲	۹۴/۰۶/۲۲	چکیده و نتیجه گیری باتوجه به فیدبک ناظرین اصلاح شد.

تأییدکنندگان

	نام و نام خانوادگی	تاریخ	امضاء	ملاحظات
مجری پروژه	علیرضا یاری			
تهیه کننده/ تهیه کنندگان	تیم مدیریت طرح			
ناظر پروژه	کامبیز بدیع، رامین شکری‌پور و روح‌اله رحمانی			
مدیر گروه				
مسئول مستندات پژوهشکده	مانا روزی طلب			
رئیس پژوهشکده/ معاون پژوهشی	علیرضا یاری			

سرفصل مطالب

فصل اول: ماشین ترجمه

ترجمه ماشینی در اصل ترجمه‌ای است که از طریق رایانه و بدون دخالت انسان انجام پذیرد. ترجمه ماشینی از جمله اولین اهداف مورد نظر در علوم رایانه و بخصوص در حوزه هوش ماشینی به حساب می‌آید و سابقه آن به حدود نیم قرن پیش از این باز می‌گردد. نخستین ترجمه‌ای که بطور کامل توسط کامپیوتر انجام شد، ترجمه متنی بود از زبان انگلیسی به زبان روسی .گر چه از آن زمان تا کنون فناوری ترجمه ماشینی رشد زیادی داشته‌است، هنوز هم نقص‌های فراوانی را داراست. اصولاً چون کامپیوترها نمی‌توانند مانند انسان هوشمند باشند، ترجمه‌ای هم که توسط آنها انجام شود، ترجمه کاملی نخواهد بود. نمی‌توان انتظار داشت که با استفاده از یک نرم‌افزار مترجم، هر متنی به آسانی ترجمه شود. نرم‌افزارهای مترجم، در بهترین حالت، عمل ترجمه را با دقتی در حدود ۷۰ درصد انجام می‌دهند. برای به دست آوردن نتیجه بهتر، لازم است قبل و بعد از ترجمه، مقداری ویرایش روی متن انجام شود.

در طی چند دهه اخیر و همزمان با گسترش و پیشرفت زبان‌شناسی رایانه‌ای، در بسیاری از کشورهای جهان، تلاشهای همه‌جانبه و پیگیر در جهت ترجمه متون از طریق کامپیوتر انجام گرفته، و حاصل کار با توجه به تنگناها، محدودیت‌ها، و مسائل خاص ترجمه درخور توجه‌است. در بعضی از زمینه‌ها حاصل کار واقعاً رضایت‌بخش است، ولی، در برخی موارد نتایج به دست آمده را علی‌رغم قابل فهم بودن، باید ویراستاری کرد. طبیعتاً نوع متن و میزان پیچیدگی آن اهمیت زیادی در نتیجه کار دارد. روش های توسعه مترجم ماشینی عبارتند از:

- ✓ روش آماری
- ✓ روش قاعده‌مند

- ✓ روش بین زبانی
- ✓ روش مبتنی بر فرهنگ لغت
- ✓ روش مبتنی بر مثال
- ✓ ترجمه ماشینی پیوندی

در زیر برخی از سامانه‌های فعال نام برده شده‌است.

۱-۱ محصول ۱: سامانه ترجمه ترگمان

شرح:

سامانه ترجمه ماشینی مورد نظر از سال ۱۳۹۰ تا کنون در قالب دو پروژه درحال توسعه می باشد. پروژه اول با عنوان "طراحی و پیاده‌سازی الگوریتم‌های ترجمه ماشینی آماری انگلیسی به فارسی" بوده است و در سال ۱۳۹۱ به اتمام رسید. هدف اصلی از این پروژه، توسعه زیرساخت ها و الگوریتم های ترجمه ماشینی تخصصی برای زبان فارسی در جهت تولید ماشین ترجمه بومی بوده است. پس از اتمام پروژه مذکور سامانه ترجمه ترگمان به عنوان یک سرویس در بخش ترجمه موتور جست و جوی پارسی جو مورد استفاده قرار گرفت. در سال ۱۳۹۲ پروژه «طراحی و توسعه یک سامانه نوین ترجمه ماشینی به منظور ارتقاء و بهبود سامانه های موجود ترجمه ماشینی فارسی انگلیسی» با هدف توسعه کمی و کیفی این محصول تعریف گردید.

طی آزمایشات و ارزیابی های رسمی انجام گرفته، مترجم ترگمان در حال حاضر با کیفیت ترین مترجم انگلیسی-فارسی و فارسی-انگلیسی در بین سایر ماشین های ترجمه موجود می باشد. همچنین این مترجم بصورت میانگین روزانه ۷۰۰۰۰ به درخواست ترجمه کاربران پاسخ داده و مورد استقبال کاربران قرار گرفته است. (www.targoman.com)

متولی:

توسعه دهنده این سامانه آزمایشگاه پردازش زبان طبیعی دانشگاه امیرکبیر می‌باشد. این پروژه با حمایت پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران) اجرا شده‌است.

۱-۲ محصول ۲: سامانه ترجمه فرآزین

شرح:

مترجم فرآزین در پروژه " توسعه سرویس ترجمه ماشینی انگلیسی-فارسی" در سال ۱۳۹۰ طراحی و پیاده سازی شده و در سال ۱۳۹۲ به صورت رسمی به بهره برداری رسید. این سامانه با معماری مبتنی بر قانون (Rule Base) توسعه یافته است. از جنبه کیفیت ترجمه این مترجم در سطح مترجم گوگل بوده و به نسبت

دارای کیفیت مطلوبی می‌باشد. از جمله قابلیت‌های این مترجم فرهنگ واژگان غنی، سرعت ترجمه مطلوب، ترجمه فایل‌ها با فرمت‌های گوناگون و ارسال نتایج از طریق پست الکترونیکی به متقاضیان است.

این مترجم قادر به ترجمه متون تخصصی در حوزه‌های منتخب می‌باشد.

این سامانه در حال حاضر از طریق آدرس زیر در دسترس می‌باشد:

www.faraazin.ir

متولی:

توسعه دهنده این سامانه آزمایشگاه پردازش زبان طبیعی دانشگاه تهران می‌باشد. این پروژه با حمایت پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران) اجرا شده‌است.

۱-۳ محصول ۳: مترجم پارس

شرح:

این مترجم نیز با رویکرد مبتنی بر قاون (Rule Base) توسعه یافته و بر پایه بانک‌های درختی زبان فارسی و انگلیسی آموزش دیده‌است. از جمله قابلیت‌های این مترجم سرعت ترجمه مطلوب و فرهنگ لغات غنی می‌باشد. همچنین این سامانه توانایی ترجمه متون در حوزه‌های تخصصی متعددی را دارا بوده و ترجمه‌ها را متناسب به این حوزه‌ها ارائه می‌دهد. جهت ترجمه در این مترجم انگلیسی به فارسی می‌باشد.

این مترجم از طریق آدرس زیر در دسترس می‌باشد:

http://parstranslator.com/

متولی:

توسعه دهنده این سامانه انتشارات مبنا رایانه پژواک (مترجم پارس) می‌باشد.

۱-۴ محصول ۴: سامانه ترجمه دبیرخانه شورای عالی اطلاع رسانی

شرح:

این مترجم براساس روش آماری توسعه یافته است. در آموزش این مترجم از پیکره‌های موازی و دوزبانه انگلیسی-فارسی استفاده شده است. حوزه پیکره آموزش ادبیات داستانی می‌باشد و به گونه‌ای این مترجم نیز در حوزه ادبیات داستانی توسعه یافته است.

این مترجم از طریق آدرس زیر در دسترس می‌باشد:

http://www.machinetranslation.ir/

متولی:

توسعه دهنده این سامانه دبیرخانه شورای عالی اطلاع رسانی می‌باشد.

۱-۵ شاخص‌های ارزیابی سامانه‌های ترجمه ماشینی

مترجم‌های ماشینی از جنبه‌های گوناگونی مورد ارزیابی قرار می‌گیرند و برای هریک معیارها و شاخص‌های متعددی موجود می‌باشد. از جمله شاخص‌های ازریابی این سامانه‌ها عبارتند از:

الف) دقت: معیارهای ارزیابی خودکار

ب) آمار بازدید و تعداد کاربران

ج) کاربرپسندی و میزان استقبال

د) سرعت پاسخگویی

ر) پایداری و دسترس پذیری

معیارهای ارزیابی خودکار موجود در این ارزیابی دقت و کیفیت ترجمه به دسته های زیر تقسیم می‌گردند و شاخص‌ترین معیار بکار رفته در هر دسته نیز ارائه شده است.

- ✓ معیار ارزیابی مبتنی بر سنجش شباهت N-Gram: معیار BLEU
- ✓ معیار ارزیابی مبتنی بر سنجش میزان خطا ترجمه: معیار TER
- ✓ معیار ارزیابی به همراه اطلاعات اضافه: معیار METEOR

همچنین بخش آمار بازدید نیز از طریق موارد زیر مورد ارزیابی و سنجش قرار می‌گیرد:

- ✓ تعداد کلمات ترجمه شده
- ✓ تعداد درخواست ترجمه ارسالی
- ✓ تعداد کاربران بازدید کننده
- ✓ تعداد کاربران بازگشتی

۱-۶ مقایسه سامانه‌های ترجمه موجود

طی ارزیابی‌های انجام گرفته از جنبه کیفیت این سامانه‌ها برحسب معیار BLEU مورد سنجش قرار گرفتند.

از جنبه کیفیت ترتیب این مترجم‌ها به شرح زیر می‌باشد:

۱. مترجم ترگمان
۲. مترجم گوگل
۳. مترجم فرآزین
۴. مترجم پارس
۵. دبیرخانه شورای عالی اطلاع رسانی

۲ فصل دوم: خطایاب

هدف از توسعه این ابزارها تشخیص و رفع خطاهای معنایی و نحوی جملات فارسی می‌باشد. خطاهای معنایی و واژگانی به خطاهایی گفته می‌شود که به شیوه املائی و نگارشی لغات برمی‌گردد. تحلیل نحوی نیز از حوزه‌هایی است که در همه‌ی زبان‌های امروز دنیا مورد توجه قرار گرفته است. توجه تنها به واژگان بدون در نظر گرفتن جایگاه دستوری آنها به هیچ وجه برای تحلیل و خطایابی یک متن کافی نیست. به عنوان مثال جمله‌ی "امروز دارد باران می‌بارم" از لحاظ واژگانی کاملاً درست ۱ است، ولی از لحاظ نحوی نادرست است. در واقع در جمله‌ی اخیر واژه‌ی "میبارم" یک واژه‌ی سردرگُم است. واژه‌ی سردرگم در واقع کلماتی هستند که از لحاظ املایی درست ولی از لحاظ نحوی در جای نادرستی قرار گرفته‌اند. اگر جمله‌ی "سیب مرا خورد" حتی از لحاظ نحوی درست است ولی از لحاظ معنایی نادرست است. با ساز و کارهای موجود در بسیاری از خطایاب‌های نحوی میتوان به صورت محدود به خطایابی معنایی نیز پرداخت. جمله‌ی "رئیس جمهور امریکا یک کودک شش‌ساله است" از لحاظ معنایی نیز درست است ولی در حوزه‌ی عمل چنین چیزی با توجه به قوانین و محدودیت سنی ریاست جمهوری نادرست می‌شود. با وجود اینکه برای خطایابی کامل باید وارد حوزه‌ی معنا و کاربرد شد؛ نیاز است که با استفاده از روش‌های مرسوم از لحاظ نحوی جملات را مورد بررسی قرار داد. این روش‌ها وابسته به دستور زبان خاصی و زبان خاصی نیستند.

۲-۱ محصول ۱: ابزار ویراستیار

شرح:

نرم‌افزار «ویراستیار» افزونه‌ای برای مایکروسافت وُرد است که برای استفاده کاربران فارسی‌زبان طراحی شده است. از قابلیت‌های ویراستیار می‌توان به اصلاح خطاهای املایی، اشتباهات ویرایشی و نشانه‌گذاری، و نیز استانداردسازی متون فارسی اشاره کرد. غلطیاب املایی ویراستیار از کارایی بالا و سرعت مناسبی برخوردار است. از موارد کارکرد غلطیاب می‌توان به موارد زیر اشاره کرد:

- ✓ اصلاح املای واژه‌ها
- ✓ ارائه‌ی لیستی از واژه‌های صحیح پیشنهادی
- ✓ اصلاح انواع غلط‌های فاصله‌گذاری
- ✓ چسبیدن واژه‌های متوالی به هم
- ✓ درج فاصله‌ی اشتباه میان کلمه
- ✓ اصلاح کاربرد نابجای فاصله به جای شبه‌فاصله

- ✓ تلفیق درج اشتباه فاصله و چسبیدن واژه‌های متوالی
- ✓ تشخیص و اصلاح واژه‌های با پسوند
- ✓ تشخیص و اصلاح تکرار متوالی کلمه
- ✓ اصلاح غلط‌های ناشی از هم‌آوایی
- ✓ امکان افزودن واژه‌های جدید به واژه‌نامه
- ✓ امکان اصلاح یک مورد غلط به طور یکباره در کل متن
- ✓ امکان نادیده گرفتن یک مورد غلط و عدم اصلاح آن
- ✓ امکان نادیده گرفتن یک مورد غلط و عدم اصلاح آن در کل متن

متولی:

این ابزار در شورای عالی اطلاع رسانی توسعه یافته است.

۲-۲ محصول ۲: ابزار وفا (واریسگر)

شرح:

وارسیگر وفا به صورت افزونه بر روی Microsoft Word نصب می شود. این خطایاب قابلیت تشخیص خطاهای لغوی، دستور زبانی و معنایی را داراست. خطای لغوی آن دسته از خطاست که بر اثر اشتباه در تایپ بوجود می آید. به عنوان مثال تایپ اشتباه "تریقت" به جای "طریقت ." خطاهای نحوی (گرامری) نیز که در سطح جمله و ارتباط کلمات با یکدیگر مطرحند، شامل خطاهایی چون عدم تطابق (مثلا بین فاعل و فعل)، عدم رعایت ترتیب بکارگیری کلمات (صفت قبل از موصوف) و بطورکلی هرگونه نوشتاری که با قواعد دستوری فارسی ناسازگار باشد، می شود. خطاهای معنایی به به کارگیری نادرست کلمه در جملات مربوط می‌شود. به عنوان مثال، "سازمان ملت متحد" که در آن کلمه "ملل" به اشتباه "ملت" تایپ شده است. (کلمات خطادار معنایی از لحاظ نوشتاری صحیح می باشد)

ویژگی های اصلی این ابزار عبارتند از:

- تشخیص و تصحیح خطاهای تایپی، دستوری و معنایی
- قابلیت نصب بر روی ویرایشگر متداول word
- قابلیت یادگیری و ارتقاء عملکرد به صورت خودکار

- دقیق و کارآمد
- رایگان

متولی:

این نرم افزار محصول موسسه تحقیقات ارتباطات و فناوری اطلاعات ایران می‌باشد.

۲-۳ شاخص‌های ارزیابی ابزارهای خطایاب

شاخص‌های ارزیابی این ابزارها عبارتند از:

الف) دقت

ب) پوشش کلیه موارد و خطاها

ج) کیفیت و قدرت فرهنگ لغات

د) پایگاه داده جامع و کافی

ر) سرعت پاسخگویی

معمولاً دو معیار عمده و مرسوم برای آزمون خطایاب‌های نحوی وجود دارد. این دو ملاک دقت و فراخوانی نام دارند. دقت، برابر است با تعداد خطاهایی که اعلام شده‌اند و واقعاً خطا هستند؛ تقسیم بر تعداد خطاهایی که سامانه پیدا کرده است. فراخوانی هم برابر است با تعداد خطاهایی که اعلام شده‌اند و واقعاً خطا هستند؛ تقسیم بر تعداد خطاهای واقعی در متن. از نظر کاربران بالا بودن فراخوانی از اهمیت بیشتری برخوردار است.

۲-۴ مقایسه ابزارهای خطایاب موجود

ابزار ویراستیار هم اکنون نیز در حال بروز رسانی می‌باشد اما ابزار خطایاب وفا چند سالی است که بروز رسانی نشده است. به همین دلیل از جنبه دقت و کیفیت ابزار ویراستیار در رتبه بالاتری می‌باشد.

فصل سوم: پیکره زبانی

۲-۵ پیکره های تک زبانه و دوزبانه

وجود پیکره های زبانی در امر پردازش زبان یکی از موارد لازم می‌باشد. این پیکره‌ها در آموزش ابزارهای پردازش زبان طبیعی به کار گرفته می‌شوند. برخی از پیکره‌های فارسی موجود در این زمینه عبارتند از :

۱)پیکره حوزه اخبار

۲)پیکره میزان (ادبیات داستانی)

۳)پیکره رمان و زیرنویس فیلم

۴)پیکره بیجن‌خان

۵)پیکره همشهری

هریک از این پیکره‌ها در حوزه‌های مختلفی توسعه یافته‌اند.

۲-۵-۱ پیکره حوزه اخبار مرکز تحقیقات مخابرات ایران

این پیکره به حجم ۱۰ میلیون لغت بوده و همانگونه که از نام آن بر می‌آید این پیکره در حوزه اخبار می‌باشد. این پیکره دارای دقت مناسی بوده و در آموزش سامانه‌های ترجمه ماشینی انگلیسی-فراسی از آن استفاده شده است. فرآیند تولید این پیکره بدین صورت می‌باشد که ابتدا متن انگلیسی این پیکره از سایت‌ها خبرگزاری استخراج و پالایش شده و در مرحله این متون به مترجمین انسانی ارائه شده است. متون ترجمه شده فارسی نیز پس از پالایش و اصلاح به صورت تراز شده متون انگلیسی (مبدا) قرار گرفته شده‌اند.

۲-۵-۲ پیکره میزان

این پیکره توسط شورای عالی اطلاع رسانی گردآوری شده است و دارای حجمی در حدود ۱۵ میلیون لغت می‌باشد. این پیکره نیز دو زبانه (انگلیسی-فارسی) بوده و ترجمه هر متن انگلیسی متناظرا در مقابل آن وجود دارد. این پیکره در حوزه داستان (رمان) و ادبیات داستانی می‌باشد.

۲-۵-۳ پیکره بی جن خان

این پیکره بصورت برچسب گذاری شده بوده و در امور پردازش زبان طبیعی کاربردهای فراوانی دارد. حوزه انتخابی این پیکره نیز اخبار و متون عامیانه می‌باشد. این پیکره به موضوعات و حوزه‌های متفاوتی دسته بندی شده‌اند که تعداد این موضوعات ۴۳۰۰ مورد می‌باشد. این پیکره در دانشگاه تهران توسعه یافته است.

۲-۶ بانک های درختی

این بانک‌ها پیکره درختی مجموعه‌ای از جملات فارسی است که در آنها روابط نحوی کلمات بر مبنای تقش دستوری آنها مشخص شده است. کاربرد اصلی این پیکره‌ها توسعه سامانه‌های ترجمه ماشینی مبتنی بر قانون می‌باشند. از جمله بانک‌های درختی موجود در زبان فارسی عبارتند از:

۲-۶-۱ دادگان درختی فارسی در چارچوب دستور ساخت سازه‌ای هسته‌بنیان دبیرخانه شورای عالی اطلاع رسانی

دادگان درختی فارسی در چارچوب دستور ساخت سازه‌ای هسته‌بنیان (HPSG) مجموعه‌ای است شامل بیش از ۱۰۰۰ جمله برچسب‌خورده با اطلاعات نحوی. ازجمله ویژگی‌های دستور ساخت سازه‌ای هسته‌بنیان این است که علاوه بر ارائه توصیف ساختاری سلسله مراتبی سازه‌ها، دانش واژگانی واژه‌ها مورد استفاده قرار می‌گیرد، و روابط بین واژه‌های یک سازه به طور واضح و صریح مشخص می‌گردد.

۲-۶-۲ دادگان درختی فارسی آزمایشگاه پردازش زبان طبیعی دانشگاه تهران

این دادگان در این آزمایشگاه توسعه یافته و در سامانه ترجمه فرآزین نیز مورد استفاده واقع شده‌است.

۲-۶-۳ شاخص‌های ارزیابی بانک‌های درختی

این بانک‌های درختی باید دارای ویژگی‌های زیر بوده که بتوان از آنها در امر پردازش زبان طبیعی استفاد ه نمود. این ویژگی‌های عبارتند از:

الف) حجم مناسب

ب) پوشش تعداد قواعد دستوری مناسب

ج) حداقل خطا

۲-۷ پیکره‌های زبانی حوزه صوت

۲-۷-۱ گروه پردازش صوت پژوهشگاه پردازش هوشمند علائم

- دادگان استاندارد گفتار رسمی میکروفنی زبان فارسی- فارس دات(FarsDat)

- دادگان استاندارد گفتار محاوره ای تلفنی زبان فارسی- فارس دات تلفنی((TFarsDat) به صورت مونولوگ)
- دادگان بزرگ گفتار رسمی میکروفنی زبان فارسی- فارس دات میکروفنی بزرگ.
- دادگان بزرگ گفتار محاوره ای تلفنی زبان فارسی- فارس دات تلفنی بزرگ(به صورت دیالوگ)
- پیکره متنی زبان فارسی (Farsi Electronic Text Corpus)همراه با برچسب.
- واژگان زایای زبان فارسی.

۸-۲ شاخص ها و معیارهای ارزیابی پیکره‌های زبانی

به منظور ارزیابی و سنجش پیکره‌ها روش‌ها و معیارهای متخلفی وجود دارد اما در مجموع یک پیکره مناسب باید دارای شرایط زیر باشد:

الف) حجم مناسب

ب) پوشش موضوعات و حوزه های مختلف

ج) حداقل نویز و خطا

د) دایره لغات مناسب و کافی

۹-۲ پایگاه مرجع دادگان زبان فارسی

۱. معرفی

از سال ۹۱ این درگاه با رویکرد تجمیع خدمات دادگانی زبان فارسی راه‌اندازی شد. آدرس اینترنتی: www.dadegan.ir

دادگان زیرساختی ابزارهای هوشمند فارسی در لایه‌های مختلف پردازش زبان از جمله: مفهوم، گفتمان، معنی، نحو، صرف که حاصل بیش از ۵۰نفر-سال فعالیت دانش آموختگان(زبان‌شناسی، نرم‌افزار و هوش مصنوعی)نیز تهیه شده است که از طریق درگاه قابل دسترس برای عموم می باشد.

۱. متولی

مرکز تحقیقات کامپیوتری علوم اسلامی (نور تهران) با مشارکت فرهنگستان

۳ فصل چهارم: ابزارهای پایه کفتار

۳-۱ ابزارهای تبدیل متن به گفتار (TTS)

به منظور ایجاد ارتباط دوطرفه بین کامپیوتر و انسان، کامپیوتر علاوه بر توانایی تشخیص گفتار، بایستی توانایی بیان و صحبت کردن را نیز داشته باشد. این مساله توانایی خواندن متون و نامه‌های الکترونیکی و یا بیان اعلام‌ها و پاسخ‌ها را برای ماشین‌ها فراهم می‌سازد.

۳-۲ محصول ۱: ابزار آریانا

شرح:

از جمله ویژگی‌های این محصول می‌توان به موارد زیر اشاره نمود:

- ✓ دقت بسیار بالا
- ✓ گفتار به صورت مفهوم و کاملا شنوا
- ✓ بیان جملات با صدای طبیعی
- ✓ تهیه فایل صوتی از متن
- ✓ قابلیت تشخیص و بیان جملات به صورت طبیعی و پیوسته
- ✓ بیان جملات به صورت صحیح با توجه به گرامر زبان فارسی
- ✓ بیان صحیح کلمات با توجه به نقش آنها در جمله
- ✓ شامل کلمات پرکاربرد زبان فارسی

این ابزار کاربردهای فراوانی دارد که می‌توان به خواندن متون برای نابینایان و کم‌بینایان، خواندن کتاب‌های الکترونیکی، خواندن اخبار و متون موجود در سایت‌های اینترنتی، خواندن (پاراف) نامه‌ها برای مدیران درنرم‌افزارهای اتوماسیون، خواندن گزارش‌ها برای کارشناسان درنرم‌افزارهای اتوماسیون، بیان توضیحات در مورد محصولات شرکت‌ها از پشت تلفن، ارایه خدمات آسانتر به کاربران توسط بانک‌ها و خواندن متون شخصی (یادداشت‌ها، وبلاگ‌ها، مقالات) اشاره نمود.

در نتیجه کاربران اصلی این ابزار عبارتند از:

- نابینایان، کم‌بینایان، معلولین جسمی
- مدیران، کارشناسان، در کلیه سازمان‌ها، بانک‌ها، شرکت‌ها و ادارات دولتی و غیردولتی
- پزشکان در بیمارستان‌ها، مراکز رادیولوژیMRI ، CTو مطب‌های شخصی

- وکلا، قضات و افراد حقوقی در دفاتر حقوقی، دادگستری‌ها و دادگاه‌ها
- اساتید دانشگاه‌ها و دانشجویان برای ارایه گزارشات و مقالات

متولی:

متولی این ابزار شرکت عصرگویش‌پرداز می‌باشد.

۳-۳ محصول۲: ابزار پارس‌خوان

شرح:

این ابزار با اهداف خواندن متون فارسی و تبدیل آنها به صدای مفهوم توسعه یافته است. از ویژگی‌های اصلی این ابزار می‌توان به موارد زیر اشاره نمود:

- پارس‌خوان یک سیستم Persian Text To Speech کاملاً رایگان و منبع‌باز است و تحت قوانین GNU/GPLمنتشر می‌شود.
- امکان خواندن متن فارسی با صدای مفهوم بدون نیاز به علامت‌گذاری متن
- امکان تهیه فایل صوتی با فرمت‌های wav و mp۳ از متن
- امکان خواندن متون با لحن گفتاری
- امکان خواندن متون انگلیسی
- امکان تنظیم سرعت خواندن
- سازگاری با انواع ویندوزها

متولی:

توسعه دهنده این محصول تیم نرم افزاری کانون فرهنگی آفتابگردان می‌باشد.

۳-۴ محصول ۳: ابزار گویا

شرح:

نرم افزار تبدیل متن به گفتار گویا با کیفیت صدای انسانی ابزار بسیار مناسبی برای ایجاد خروجی صوتی همزمان و یا تولید فایل‌های صوتی می باشد.

متولی:

این ابزار از طریق شرکت دانش بنیان پکتوس توسعه یافته است.

۳-۴-۱شاخص‌های ارزیابی ابزارهای تبدیل‌گر متن به گفتار

این ابزارها را از جنبه‌های گوناگون مورد ارزیابی و سنجش قرار می‌دهند. شاخص‌های مورد نظر عبارتند از:

- دقت بالا
- گفتار به صورت مفهوم و کاملا شنوا
- پایگاه داده با حجم بالا و پوشش کلمات پرکاربرد
- بیان جملات با صدای طبیعی
- تهیه فایل صوتی از متن
- قابلیت تشخیص و بیان جملات به صورت طبیعی و پیوسته
- بیان جملات به صورت صحیح با توجه به گرامر زبان فارسی

۳-۴-۲ مقایسه ابزارهای تبدیل متن به گفتار فارسی

این ابزارها با ارزیابی انسانی از جنبه کیفیت و دقت مورد بررسی قرار گرفت. نتایج بدست آمده بدین صورت می‌باشد که این هریک از این ابزارها در زمینه و حوزه خاصی با دقت‌تر می‌باشند در نتیجه نمی‌توان مقایسه دقیقی بین آنها انجام داد.

۳-۵ سامانه بهسازی کیفیت گفتار

شرح:

هدف از طراحی و پیاده‌سازی این سامانه، بهبود کیفیت فایل‌هایی است که کیفیت شنیداری آنها بر اثر نوفه جمع‌شونده محیطی تخریب شده و از قابلیت فهم پایینی برخوردار هستند. فرض می‌شود که فایل‌های ورودی این سامانه به‌صورت تک‌کاناله و فقط با استفاده از یک میکروفن ضبط شده‌اند.

ورودی و خروجی سامانه:

ورودی سامانه بهسازی کیفیت گفتار، در حالت استاندارد یک فایل صوتی با فرمت .wav است، ولی می‌تواند فایل‌های صوتی با فرمت‌های دیگر را نیز (در صورت موجود بودن دیکودر آن) قبول کند (همچون . mp۳، mp۴. و ...). البته ذکر این نکته لازم است که برخی از فشرده‌سازها با اعمال تغییرات طیفی غیریکنواخت و ناهمگون، کار بهسازی گفتار را مشکل می‌کنند. بنابراین بهتر است که در صورت امکان فایل‌های نویزی فشرده نگردد.

مشخصات سامانه:

۱. سامانه بهبود کیفیت گفتار دارای مشخّصات زیر است:
۲. کاهش انواع نویز بدون توجّه به مشخّصات آماری و توزیع طیفی آن
۳. کاهش نویزهای ایستان (نویزهای پایدار با تغییرات آرام در طول زمان)
۴. کاهش نویزهای ناایستان (نویزهای ناپایدار با تغییرات نسبتاً سریع در طول زمان)
۵. تنظیم میزان کاهش نویز در فایل خروجی (برای کنترل میزان تخریب فایل بهبودیافته)
۶. امکان انجام فیلترینگ در یک باند خاص
۷. کاهش نویزهای ضربه‌ای (Impulse Noise) در سیگنال صوتی
۸. امکان حذف انعکاس (Reverberation) از سیگنال صوتی (این قابلیت می‌تواند در آینده به نرم‌افزار اضافه گردد).
۹. امکان اعمال Equalizer با میزان تقویت یا تضعیف دلخواه در باندهای فرکانسی مختلف (این قابلیت می‌تواند در آینده به نرم‌افزار اضافه گردد).
۱۰. امکان عملکرد سامانه به‌صورت Online برای بهبود کیفیت گفتار به‌صورت برخط و بلادرنگ (این قابلیت می‌تواند در آینده به نرم‌افزار اضافه گردد.)

کاربردها:

۱- افزایش قابلیت فهم کلام در فایل‌های گفتاری تخریب‌شده.

۲- کاهش میزان خستگی گوش در هنگام گوش دادن به فایل‌های بسیار نویزی.

۳- افزایش کیفیت آرشیوهای صوتی و سخنرانی‌های قدیمی که در شرایط نامناسب ضبط‌شده و یا اینکه کیفیت آنها به مرور زمان اُفت پیدا کرده است.

۴- افزایش دقّت سامانه‌های شناسایی گوینده و گفتار برای شناسایی فایل‌های نویزی و بی‌کیفیت.

۵- افزایش قابلیت فهم و کاهش میزان خستگی گوش در کانال‌های ارتباطی همچون HF که معمولاً کیفیت مناسبی ندارند.

متولی

پژوهشگاه پردازش هوشمند علائم

۴ فصل پنجم: ابزارهای پایه پردازش زبان فارسی

۴-۱ محصول ۱: پارسی‌پرداز

شرح:

این ابزار در بسیاری از پروژه‌های مرتبط با زبان‌شناسی رایانشی قابل بکارگیری خواهد بود، به نحوی که با دریافت متن خام فارسی، پردازش‌های زبان فارسی را از پایین‌ترین لایه پردازش زبان طبیعی یعنی لایه لغوی آغاز و تا لایه‌های بالاتر مانند لایه دستور و معنا ادامه می‌دهد. این نرم افزار قادر است ترکیبی از یکسان‌سازی، قطعه‌بندی، برچسب‌گذار جزء کلام، تحلیل‌گر ساختواژی شامل ریشه‌یاب و لم‌یاب، تجزیه‌گر نحوی وابستگی و در نهایت برچسب‌گذار نقوش معنایی(SRL)، را اجرا کند.

متولی

ابزار پارسی‌پرداز توسط اعضای پروژه سامانه پرسش و پاسخ قرآنی در پژوهشگاه ارتباطات و فناوری اطلاعات تولید و پیاده‌سازی شده است.

۴-۲ محصول ۲: استپ-۱

معرفی

این ابزار دو سطح از پردازش‌های زبانی یعنی لایه لغوی و لایه مورفولوژی را پوشش می‌دهد و شامل دو ابزار قطعه بند و ریشه‌یاب می‌باشد.

متولی

آزمایشگاه پردازش زبان دانشگاه شهید بهشتی

۴-۳ ابزارهای زبانی موسسه نور

شرح:

ابزارهای زبانی پایه‌ای که توسط این موسسه برای فعالیت‌های دادگانی فارسی تهیه و ارائه شده است شامل موارد زیر است:

- ابزارهای برچسب‌زنی روابط صرفی، نحوی و معنایی

- ابزارهی خطایابی خودکار و قاعده محور برچسب‌های نحوی
- ابزارهای جستجوی نحوی
- ابزارهای گزاره‌های معانی و گزارش اختلاف
- ابزار جستجوی معنایی
- ابزارهای برچسب‌زنی NER

متولی

موسسه نور

۴-۴ هضم

شرح:

این بسته نرم‌افزاری که سازگار با بسته NLTK می‌باشد. شامل مجموعه‌ای از ابزارهای پردازش زبان فارسی است که به زبان پایتون تهیه و به صورت متن‌باز ارائه شده است.

ویژگی‌ها

- تمیز و مرتب کردن متن
- تقطیع جمله‌ها و واژه‌ها
- ریشه‌یابی واژه‌ها
- تحلیل صرفی جمله
- تجزیه نحوی جمله
- واسط استفاده از داده‌های زبان فارسی
- سازگاری با بسته NLTK
- پشتیبانی از پایتون نسخه ۲ و ۳

متولی

گروه سبحه

۵-۴ شاخص‌های ارزیابی ابزارهای پایه پردازش زبان

- پوشش
- دقت

فصل هفتم: وردنت

وردنت نام عمومی است که بر شبکه های واژگانی مختلفی برای بسیاری زبان‌های جهان اطلاق می شود. این شبکه ها عموماً در نقش واژهستان شناسی و یا واژگان معنایی محاسباتی در خدمت سیستم‌های هوشمند دانش پایه و معناگرا قرار دارند. وردنت انگلیسی یا شبکه واژگانی پرینستون (PWN) نخستین بار توسط جرج میلر و همکارانش در سال ۱۹۸۶ در آزمایشگاه علوم شناختی دانشگاه پرینستون براساس واژگان ذهنی و در حوزه‌ی پژوهش‌های روانشناسی زبان طراحی و ایجاد شد. وردنت در واقع شبکه‌ای معنایی از بیش از یکصدهزار مفهومی است که بوسیله روابط معنایی به هم مرتبطند. هر مفهوم نشان‌دهنده‌ی مجموعه‌ای انتزاعی از عناصری می باشد که بر اساس مختصه‌های مشترکشان یک گروه را تشکیل می دهند. پس از آن شبکه‌های واژگانی بسیاری برای زبان‌های دیگر طراحی و به وردنت انگلیسی نگاشته شدند که از این میان می‌توان به وردنت زبان های اروپایی، زبان‌های منطقه بالکان، عربی، هندی، دانمارکی، آفریقایی و ... اشاره نمود. هریک از این شبکه های جدید با ساختار و ویژگی های خاص زبان مورد پوشش خود طراحی و ارائه شده اند. هم اکنون برای بیش از ۴۰ زبان دنیا وردنت طراحی شده و یا در دست طراحی است. علاوه بر طرح‌های عظیم وردنت اروپایی و وردنت بالکان، نام زبانهایی بسیاری مانند عربی، دانمارکی، آلمانی، زبان‌های بانتو (بَکه واژگانی آفریقایی)؛ باسک، کاتالان، چینی، ژاپنی، کره ای، سوئدی، عبری، ایسلندی، هندی و کانادا و ماراثی، اوریه، مولداویایی، روسی، اسلووناییایی، تامیلی، تایلندی و حتی سانسکریت در فهرست انجمن جهانی وردنت به چشم می خورد که بعضی از آنها گویشوران بسیار کمتری نسبت به فارسی دارند و حتی تعداد کاربران اینترنت در آن کشورها با ایران قابل مقایسه نیست.

۵-۱ وردنت‌های مطرح جهان

۵-۱-۱ شبکه‌واژگانی انگلیسی - وردنت پرینستون

در سال ۱۹۷۸، گروهی از زبان‌شناسان و روان‌شناسان دانشگاه پرینستون به سرپرستی جرج ا . میلر بر پایه یافته‌های متعدد حوزه روان‌شناسی زبان درباره واژگان ذهنی انسان، طراحی یک پایگاه داده واژگانی را آغاز کردند.

حاصل تلاش‌های این گروه در قالب پایگاه داده‌هایی مبتنی بر روابط مفهومی عرضه شد که پیاده‌سازی نوعی الگو از واژگان ذهنی محسوب می‌شود . این پایگاه داده‌ها که نخستین نسخه آن در سال ۱۹۹۰ عرضه‌شد، «وردنت» نام دارد. این سامانه رایانه‌ای که در آغاز صرفاً بازنمود عینی واژگان ذهنی انسان و موضوعی جذاب

در روان‌شناسی زبان محسوب می‌شد، رفته رفته در گروه‌ها و محافل زبان‌شناسی رایانه‌ای مطرح شد و به عنوان ابزاری کارآمد در طرح‌های مختلف پردازش زبان طبیعی مورد استفاده قرار گرفت [۸] .

۵-۱-۲ شبکه‌های واژگانی چندزبانه

وردنت پرینستون صرفاً زبان انگلیسی را پوشش می‌داد و نیاز مبرمی برای ساخت منبعی مشابه برای دیگر زبان‌ها وجود داشت. به ویژه، منبعی مورد نیاز بود که وردنت‌های زبان‌های مختلف را در قالب یک منبع واژگانی چندزبانه به هم پیوند دهد. ابتدا طرح شبکه واژگانی اروپایی در سال ۱۹۹۶ با هدف ساخت شبکه واژگانی زبان‌های هلندی، اسپانیایی و ایتالیایی و سپس پیوند آنها به شبکه واژگانی انگلیسی، آغاز شد. پس از آن با اضافه شدن زبان‌های آلمانی، فرانسوی، چکی و استونیایی، این پروژه در سال ۱۹۹۹ پایان پذیرفت. بعد از آن طرح شبکه واژگانی بالکان که هدف آن ایجاد شبکه واژگانی زبان‌های کمتر مطالعه شده ی اروپایی بوده، در سال‌های ۲۰۰۱ تا ۲۰۰۴ در دانشگاه پاتراس یونان اجرا شد. در این طرح که ادامه طرح شبکه واژگانی اروپایی محسوب می‌شود، زبان‌های آلمانی، ترکی، بلغاری، صربی، و رومانیایی به پایگاه داده‌های آن شبکه افزوده شده‌اند[۴۲].

در این کنسرسیوم،تیم های چک و فرانسه نیز به منظور سازگاری کامل با شبکه‌های واژگانی قبلی حضور داشتند، توسعه پایگاه داده یورو وردنت با ۵ زبان جدید تأثیر سیاسی و فرهنگی-اجتماعی مهمی در بر دارد.

بعدها در جهت تلاش برای طراحی شبکه‌ی واژگانی چند زبانه از زبانهای آسیایی نیز آسیانت طراحی شد که شامل زبانهای ژاپنی، تایلندی، مغولی، اندونزیایی می شد.در ادامه به بررسی سه شبکه واژگانی چند زبانه اروپایی(یوروردنت)، بالکانت و آسیانت می پردازیم.

۵-۱-۲-۱ یورو وردنت

مهمترین تفاوت شبکه اروپایی با وردنت پرینستون چندزبانه بودن آن است. چندزبانگی در اصل با اضافه کردن یک رابطه هم ارزی[۱] برای هر یک از دسته‌های هم‌معنای یک زبان به نزدیک‌ترین دسته هم‌معنا به آن در وردنت پرینستون حاصل می‌شود. دسته‌های هم‌معنایی که به یک دسته هم‌معنای یکسان در وردنت پرینستون پیوند داده شده باشند، هم‌ارز و یا با معنای نزدیک فرض می‌شوند و بنابراین می‌توانند با هم انطباق داده شوند.

---

[۱] Equivalence

به طور کلی، پایگاه داده‌های شبکه واژگانی اروپایی بر مبنای ساختار شبکه واژگانی پرینستون و به ویژه نسخه ۱/۵ آن طراحی شده است. مفهوم دسته هم‌معنا و روابط معنایی اصلی در شبکه واژگانی اروپایی حفظ شده‌اند. اما تغییرات خاصی در طراحی پایگاه داده‌های آن اعمال شد که مبتنی بر اصول زیر بودند: الف) ساخت یک پایگاه داده‌ه چند زبانه، ب) حفظ روابط زبان-ویژه در شبکه‌های واژگانی و ج) ساخت شبکه‌های واژگانی به صورت نسبتاً مستقل با استفاده از منابع موجود.

به منظور حفظ ساختارهای زبان-ویژه و امکان ایجاد منابع مستقل دیگر، میان واحدهای[۱] زبان-ویژه و واحدهای جداگانه مستقل از زبان تمایز ایجاد شده است. هر واحد زبانی نمایانگر نظامی یکتا و خودسامان از روابط درون‌زبانی میان دسته‌های هم‌معنا است. روابط میان دسته های هم معنای زبان‌های مختلف به کمک نمایه میان زبانی و با بهره گیری از یک هستان شناسی سطح بالا تامین می شود.

۵-۱-۲-۲ بالکانت

در مورد شبکه واژگانی بالکانت، اصول کلی و چارچوب ساخت شبکه واژگانی، همان اصول و چارچوب‌های به‌کار رفته در شبکه واژگانی اروپایی هستند اما ابزارهای ساخت بهبود یافته‌اند و همچنین مجموعه مفاهیم بنیادی برای ایجاد شبکه‌های واژگانی گسترده‌تر و پیراسته‌تر شده‌اند. مفاهیم بنیادی و ابزارهای ساخت شبکه واژگانی بالکان، اکنون به‌عنوان ابزارها و مواد استاندارد ساخت شبکه‌های واژگانی جدید شناخته می‌شوند.

اهداف اصلی در پروژه بالکانت نیز عبارتند از:

۱) توسعه حداقل ۸۰۰۰ دسته هم معنا برای هر وردنت زبان-ویژه جدید بطوریکه با این اندازه کوچک، وردنت ها باید در کاربردهای واقعی مفید باشند.

۲) اطمینان از همپوشی حداکثر بین زبانی میان وردنت های بالکانت و سازگاری با وردنت های پروژه یورو وردنت.

۳) ساخت ابزارهای نرم افزاری رایگان برای مدیریت کارآمد و استخراج واژگان معنایی چند زبانه

۴) توسعه نشانگرهای کاربردی از قبیل رفع ابهام معنایی کلمه(**WSD**)، شاخص بندی هوشمند، بازیابی اطلاعات میان زبانی

برای ضمانت سازگاری وردنت های تک زبانه، کنسرسیوم شبکه‌واژگانی اروپائی، رویکرد هایی را اتخاذ کرد که مهمترین آنها عبارتند از:

---

[۱] Modules

۱.نمایه میان زبانی یورو وردنت

۲.روابط واژی-معنایی یورو وردنت

۳. هستان شناسی مرتبه بالای یوروواردنت

۴.مجموعه مفاهیم بنیادی(BCS) .

هدف از ساخت بالکانت، یکپارچه سازی وردنت‌ها برای زبان‌های شبه جزیره بالکان، از قبیل بلغاری، یونانی، رومانی، صربستانی و ترکی می‌باشد. در واقع ساخت یک پایگاه داده‌ی لغوی چند جانبه از وردنت در چندین زبان اروپای شرقی و مرکزی است. بالکانت همانند یوروواردنت ساخته شده و علاوه بر آن ویژگی‌های جدیدی هم به آن اضافه شده است، که در نتیجه آن از ساختار ILI جهت اطمینان ارتباط هم ارزی مفهومی بین وردنت‌ها استفاده می‌شود که مدیریت وردنت داخل شبکه‌ای را گسترش می‌دهد. بنابراین هر شرکت کننده علاوه بر اینکه به طور مستقل از وردنت محلی خودش اطلاع دارد به طور همزمان می‌تواند وردنت‌های دیگر را ببیند و توانایی‌های آنها را هم بررسی کند.

۵-۱-۲-۳ آسیانت

وردنت آسیا (AWN) حاصل تلاش جهت طراحی وردنتی به هم بافته از زبان‌های آسیایی مانند ژاپنی، تایلندی، مغولی، اندونزیایی، کره ای، میانماری، ویتنامی و بنگالی است. نقطه آغازین این طرح فرهنگ‌های واژگانی دوزبانه موجود انگلیسی و زبان های مورد نظر انجام شده است. واژگان این وردنت را می‌توان به کمک معادل‌های انگلیسی به واژگان وردنت پرینستون اتصال داد. برخی مراکز و موسسات اصلی دست‌اندرکار این وردنت بزرگ به شرح زیرند:

- آزمایشگاه زبان شناسی رایانه ای تایلندی (TCL)
- موسسه ملی اطلاعات و فناوری ارتباطات ژاپن (NICT)
- مرکز ملی الکترونیک و تکنولوی کامپیوتر تایلند (NECTEC)
- موسسه ارزیابی و کاربرد فناوری اندونزی (BPPT)
- دانشگاه ملی مغولستان (NMU)
- فدراسیون کامپیوتر میانمار (MCF)

آسیانت که بر اساس وردنت انگلیسی طراحی شده است چارچوبی برای تکمیل وردنت‌های زبان‌های آسیایی به شمار می آید. از آنجا که ساختار آسیانت از ساختار وردنت انگلیسی الگوبرداری شده است، از طریق آن می‌توان به سایر وردنت‌های جهان نیز دسترسی داشت.

۵-۱-۳ وردنت عربی

شروع شکل گیری و ساخت وردنت عربی را می توان ساخت یک فرهنگ الکترونیکی دو زبانهٔ عربی-انگلیسی دانست که بر اساس وردنت انگلیسی ساخته شده است[۴۷]. ویژگی های این فرهنگ که در قالب مقاله ای در مجموعه مقالات کنفرانس جهانی وردنت ۲۰۰۴ آورده شده است نشان می دهد که هدف از تهیهٔ آن فراهم آوردن مقدمات ساخت وردنت عربی بوده است.

همانگونه که می دانیم یکی از ویژگی های زبان عربی که آن را از زبان هایی مانند زبان اروپایی متمایز می سازد، وجود ساخت واژهای تصریفی و اشتقاقی فراوان در این زبان است. به گونه ای که بر اساس مجموعه ای از ریشه های کلمات و براساس مجموعه ای از وزن های مشخص که دربر دارندهٔ واقعیت های تصریفی و اشتقاقی هستند، کلمات مختلف زبان عربی که که ریشهٔ یکسان دارند ساخته می شوند. نشان دادن این ویژگی خاص عربی جمله مواردی است که در ساخت این فرهنگ دو زبانه مورد توجه قرار گرفته است. به عنوان مثال در مورد کلمه ای مثل "مولود" ریشهٔ این کلمه یعنی "و ل د" و نیز وزن آن یعنی "مفعول" و رابطهٔ ان با کلماتی مانند "تولید" یا "والد" که ریشهٔ یکسان دارند نشان داده شده است. از دیگر ویژگی های این فرهنگ الکترونیکی آن است که کلمات عربی در آن با خط عربی نوشته شده اند و با الفبای آوانگاری نوشته نشده اند. یعنی اگرچه این فرهنگ یک دادگان دو زبانهٔ الکترونیکی است، اما در طراحی آن امکان کار با خط عربی پیش بینی شده است. نکته ای که باید به آن اشاره کرد این است که در پروژه مذکور در واقع حاصل کار به دست دادن یک فرهنگ دو زبانهٔ مبتنی بر وردنت انگلیسی نبوده است. بلکه فراهم کردن مقدمات کار و انجام پیش بینی های لازم برای این کار با توجه به ویژگی های خاص زبان عربی مورد توجه بوده است.

۵-۲ وردنت‌های فارسی

در چند سال اخیر در داخل و خارج از کشور فعالیت‌هایی برای تهیه‌ی وردنت فارسی آغاز شده است. با توجه به کاربردهایی که امروزه برای وردنت در زمینه‌ی بازیابی اطلاعات و حوزه‌های مختلف پردازش معنایی تعریف شده است، علاوه بر زبان‌شناسان، توجه بسیاری از متخصصان علوم کامپیوتر نیز به تهیه‌ی وردنت فارسی معطوف شده است.

۵-۲-۱ فارس‌نت

شرح:

هدف از توسعه این پایگاه دانش، ایجاد یک شاخه فارسی برای وردنت جهانی است که در تحقیقات و پژوهش‌های زبان فارسی قابل استفاده بوده وامکانات تبدیل چندزبانه را نیز فراهم کند. فارس‌نت در واقع وردنت

فارسی حوزه عمومی می‌باشد در حال حاضر این پایگاه دانش حاوی بیش از ۳۰۰۰۰ مدخل شامل پرکاربردترین واژه ها در حوزه عمومی است.

ویژگی‌های فنی:

- پوشش کلمات پرکاربرد زبان فارسی حوزه عمومی در مقوله نحوی اسم، فعل، صفت و قید
- وجود روابط میان مقوله ای و درون مقوله ای
- افزودن ساختار آرگومانی (قاب) افعال (ساختاری شبیه ورب نت)
- نگاشت مفاهیم آن به وردنت ۳ انگلیسی

متولی:

آزمایشگاه پردازش زبان دانشگاه شهید بهشتی

۵-۲-۲ فاوانت

شرح:

هدف از توسعه این پایگاه دانش ایجاد هستان‌شناسی واژگان حوزه فاوا بوده است. در حال حاضر این پایگاه دانش شامل ۳۰۰۰۰ مدخل از پرکاربردترین واژه‌های حوزه فاوا است.

ویژگی فنی:

- پوشش کلمات پرکاربرد فارسی حوزه فاوا در مقوله نحوی اسم، فعل، صفت و قید
- وجود روابط میان مقوله ای و درون مقوله ای
- افزودن ساختار آرگومانی (قاب) افعال (ساختاری شبیه ورب نت)
- امکان جستجوی پیشرفته و دوزبانه
- نگاشت مفاهیم آن به وردنت ۳ انگلیسی

متولی:

دانشگاه بوعلی‌سینا همدان

۵-۳ شاخص‌های ارزیابی وردنت

- پوشش

- تعداد مفاهیم
- تعداد کلمات
- تعداد روابط

- دقت

۶ نتیجه گیری:

در این مستند اصلی‌ترین ابزارها و سیستم‌های موجود در حوزه پردازش زبان طبیعی (زبان فارسی) بررسی گردیدند. در فصول این مستند مشخصات و ویژگی‌های هریک از این ابزارها استخراج شده و در صورت امکان مقایسه‌های بین ابزارها و سرویس‌های موجود در هر بخش انجام گردیده است.

از بررسی سامانه‌های ترجمه موجود به این نتیجه می‌رسیم که این سامانه‌ها از جنبه کیفی و کمی نیازمند توسعه و ارتقاء می‌باشند. همچنین برخی از این سامانه‌ها در برخی موارد دارای نقطه ضعف و قوتی بوده به باید برنامه‌ریزی دقیقی با هدف ارتقاء این سامانه‌ها و رسیدن به کیفیت نمونه‌های خارجی انجام پذیرد.

یکی از منابع اولیه و مورد نیاز در حوزه پردازش زبان طبیعی، پیکره‌های تک زبانه و دوزبانه می‌باشد. با بررسی انجام گرفته این نتیجه بدست آمد که وضعیت این منابع برای زبان فارسی در سطح مطلوبی نبوده و نیازمند تکمیل این منابع می‌باشیم. بدین منظور پیشنهاد می‌گردد با تعریف پروژه‌هایی در این حوزه اقدام به توسعه این پیکره‌ها در حوزه‌های مختلف نمائیم.

سایر ابزارهای پردازش زبان فارسی تشریح گردیده در این مستند نیز در وضعیتی بوده که نیازمند برنامه‌ریزی جهت توسعه این ابزارها می باشند. نکته قابل توجه در توسعه این ابزارها این است که هریک از ابزارها در حوزه‌ای جداگانه و یا در بخشی دارای قوت بوده و نمی‌توان مقایسه‌ای بین آنها انجام داد.

مراجع

[۸] فامیان, ع. و د. آقاجانی, طراحی شبکه واژگانی صفات زبان فارسی. مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه, ۱۳۸۵: ۴۲-۵۴ .p.

[*] مهرنوش شمس فرد، سمیه باقر بیگی، مصطفی عاصی، نیلوفر منصوری، علی فامیان، یاسمن معتضدی و مسعود روحی زاده، مطالعه و بررسی وردنت های مطرح جهان، از نظر حوزه پوشش، ساختار و ویژگی ها، ۱۳۸۸.

[*] محرم منصوری زاده، محمد نصیری، مطالعه تطبیقی وردنت‌های فارسی و غیر فارسی، ۱۳۹۱.

[۴۲] Tufis, D., ed, *Special Issue on the BalkaNet Project.* Romanian Journal of Information, Vol. ۷, ۲۰۰۴: p. Nos ۱–۲.

[۴۷] Black, W.a.E., S, *A prototype English –Arabic dictionary based on WordNet.* GWC, ۲۰۰۴: p. ۶۷-۷۴.