بسمعه تعالی

وزارت علـوم، تحقیقات و فناوری

مرکز منطقه‌ای اطلاع رسانی علوم و فناوری

گزارش عملکرد سالانه
گروه پژوهشی زبان شناسی رایانه ای
۱۳۸۸

گروه طرح و برنامه – خرداد ۱۳۸۹

گروه پژوهشی زبان شناسی رایانه ای

مقدمه:

رايانه اگرچه از دستاوردهاي غرب است اما در دو دهه اخير آنچنان در جوامع مختلف از جمله جامعه ما رخنه كرده كه امروزه در بعضي از مراكز حتي تصور اينكه روزي رايانه در آن مركز نباشد، بسيار مشكل است. با وجود اين هنوز تنها تعداد محدودي از زبانها با رايانه تطبيق داده شده و در آن قابل استفاده مي باشند و زبان فارسي از اين جهت در ابتداي راه است .

زبانشناسي رايانه اي به قسمتي از زبانشناسي گفته مي شود كه مسائل مربوط به زبان و رايانه را بطور عمومي مورد بحث و بررسي قرار مي دهد و نشان مي دهد كه زبانشناسي چگونه مي تواند به حل مسائل موجود در مورد تطبيق خط و زبان در رايانه، كمک نمايد.

اعضاء اين گروه سعي دارند كه با اجراي طرحها و نشر كتابها و مقالات و سخنرانيها و ... بتوانند در برطرف كردن مشكلات موجود در خصوص هم آهنگي خط و زبان فارسي با رايانه كمک كرده و در ايجاد يک نگرش علمي نسبت به آن موثر واقع شوند.

در زير به تعدادي از موارد كاربرد علم زبانشناسي رايانه اي اشاره مي شود:

۱- ساخت ريشه ساز زبان فارسي: ريشه سازي يكي از ابزارهايي است كه در بازيابي اطلاعات در برخورد با مسئله عدم انطباق واژگان مورد استفاده قرار مي‌گيرد. (يعني عدم تطبيق واژه هاي پرس و جو با واژه هاي مدرک) به فرايند حذف هر پسوند از واژه ها و تبديل اين واژه ها به ريشه هاي آنها ريشه سازي گفته مي شود. براي مثال ريشه سازي واژه "ورزش" را به شكل "ورز" (ريشه مضارع فعل ورزيدن) به وجود مي آورد. ريشه سازها عموماً متناسب با هر زبان خاص تهيه مي شوند. طراحي ريشه سازها مستلزم خبرگي زبانشناسي در زبان و درک نيازهاي مرتبط با بازيابي اطلاعات مي باشد.

۲- ساخت ماشين هاي ترجمه: ترجمه ماشيني يكي از نخستين استفاده هايي است كه از پردازش زبان طبيعي به عمل آمده است. براي ساخت اينگونه ماشينها استفاده از تئوري هاي نحوي و معنايي موجود در زبانشناسي ضروري است. در اين زمينه زبان شناسان مي توانند در حيطه هاي: **ايجاد *سيستم هاي ترجمة تماماً خودكار، ترجمة ماشيني نيازمند به انسان، ترجمه با كمک ماشين، ساخت ويرايش گرهاي املايي، ايجاد سيستم هاي ترجمة روي خط، ايجاد اصطلاحنامه هاي روي خط و سرانجام ماشين هاي همگرا و واگرا*** مفيد باشند.

نكتة حائز اهميت آنكه براي ساخت ماشين هاي ترجمه از نظرية اطلاعات و رمزنگاري نيز استفاده مي شود كه اين موضوع ارتباط حيطه هاي اطلاع رساني با زبانشناسي رايانه اي را نشان مي دهد. بديهي است مراكز اطلاع رساني برخوردار از اين مزيت قادر خواهند بود اطلاعات مورد نياز كاربران را به بيش از يک زبان ارايه نمايند.

۳- بازيابي اطلاعات: اساس كار مراكز و پايگاههاي اطلاع رساني بر جستجو و بازيابي اطلاعات بنا شده است. زير بخشي از زبانشناسي رايانه اي كه به ***معناشناسي رايانه اي*** موسوم است تاكنون توانسته است كيفيت اين گونه سيستم ها را به نحو مطلوبي ارتقاء دهد. از ***معناشناسي منطقي*** نيز استفاده هاي فراواني به عمل مي آيد.

به اختصار مي توان گفت كه اساس كار سيستم هاي بازيابي اطلاعات اسناد موجود نيست،بلكه بازنمودهاي معنايي است و ايجاد اين بازنمودها كاري است در حيطه

معناشناسي رايانه اي. دو مورد از مشكلات عمده موجود در سيستم هاي بازيابي اطلاعات به *ابهام و تشابه واژگاني* مربوط است كه رفع اين مشكلات نيز در حيطة زبانشناسي رايانه اي است.

منطق توصيفي، *شبكه هاي معنايي و صورت گرايي قالب-* محور مبناي كار سيستمهاي اطلاع رساني در حيطة معناست.

۴- بازشناسي گفتار و سيستم هاي گفتار به متن: بازشناسي گفتار بر اساس مدلهاي زبانشناختي بنا شده كه از آن جمله مي توان به مدل ماركو اشاره نمود. *دستور حالتهاي* محدود *چامسكي* نيز كاربرد فراوان دارد. در سيستم هاي بازشناسي گفتار شناخت ***امواج صوتي، چگونگي تعبير و تفسير آنها، رمزگذاري و رمز گشايي و چگونگي تغيير مشخصه هاي آوايي*** مسايلي عمده و اساسي مي باشند كه يک آواشناس يا واج شناس از عهده آن بر مي آيد. (آواشناسي و واج شناسي يكي از زير شاخه هاي علم زبانشناسي است). تاكنون در زمينه تلفظ و املا الگوهاي مختلفي ارائه گرديده است.

۵- برچسب زني كلمات: اين حيطه نيز از حيطه هاي مطرح در علم اطلاع رساني است و به يكي ديگر از زير شاخه هاي علم زبانشناسي كه نحو نام دارد مربوط است. *شناخت مقوله هاي مختلف دستوري و چگونگي ايجاد ارتباط و تمايز* بين دسته هاي مختلفي از كلمات به اطلاعات تخصصي در حيطه نحو نيازمند است.

۶- سيستم هاي متن به گفتار: حيطه رايج ديگري در علم اطلاع رساني است كه از واج شناسي و آواشناسي رايانه اي استفاده هاي زيادي به عمل مي آورد. ***شناخت اندام هاي صوتي، آوانويسي آواهاي زباني، استفاده از الگوها و قواعد واجي و واج شناسي*** در سيستم هاي متن به گفتار همگي نيازمند به فردي است كه در حيطه واج شناسي به مطالعه و تحقيق پرداخته باشد.

در ادامه براي رعايت اختصار كاربردهاي عمده ديگر اين حيطه فهرست وار ارايه مي گردد.

۷- واژه سازي: جستجو و پيشنهاد معادل هاي مناسب براي واژه هاي فاقد معادل فارسي (البته براي اين كار لازم است با هماهنگي و يا همكاري فرهنگستان زبان و ادب فارسي اقدام گردد).

۸- نمودارهاي N: از اين نمودارها براي شمارش اعداد در متن و ... استفاده مي شود و مطالب موجود تا حد زيادي به رياضيات و علم رايانه نيز مربوط است.

۹- دستورهاي بافت وابسته و مستقل از بافت: سيستم هاي بازيابي با استفاده از اين گونه الگوها مي توانند به جستجو يا تجزيه متن دست بزنند كه از آن جمله مي توان به *تجزيه هاي از كل به جز و از جز به كل* اشاره نمود.

۱۰- تجزيه و تحليل معنايي

۱۱- گفتمان: كه در آن در خصوص چگونگي تعبير و تفسير ضمايز و يافتن مرجع آنها و نكات مرتبط ديگر بحث مي شود كه در ماشينهاي ترجمه و ... كاربرد فراوان دارد.

۱۲- تبديل متن به گفتار

۱۳- تبديل گفتار به متن

۱۴- خلاصه سازي خودكار

۱۵- موتورهاي كاوش هوشمند

اهداف مهم گروه پژوهشي زبانشناسي رايانه اي بشرح زير مي باشد:

۱- مطالعه ساختار و واژگان هاي علمي و پژوهشي جهت دريافت توصيفگر هاي مناسب

۲- مطالعه و بررسي علوم اطلاع رساني و كتابداري و زبانشناسي و افزايش بهره وري آنها در سطح كشور و منطقه

۳- مكانيزه كردن نمايه سازي و چكيده نويسي مدارک علمي كشور و منطقه

۴- مطالعه در ساختار زبان فارسي و ارتباط منطقي واژه ها و پراكنش آماري آنها

۵- مطالعه در ساختار ترجمه در زبانهاي مختلف با استفاده از رايانه و منطق زبانشناسي

۶- افزايش بهره وري اطلاع رساني با استفاده از علم زبانشناسي در سطح كشور و منطقه

۷- تعيين راهبردهای جستجو با استفاده از اصلاحنامه ها و تزاروسها و واژه نامه هاي كامپيوتري

۸- بررسي چگونگي تثبيت اصطلاحات فارسي و كاربرد صحيح آن زبان بعنوان يک زبان علمي

۹- بررسي و رفع مشكلات جستجو و بازيابي اطلاعات رايانه اي با استفاده از منطق و علم زبانشناسي

۱۰- بررسي ارتباط زبانشناسي و بازيابي اطلاعات با الگوهاي منطقي زبانشناسي

۱۱- بررسي املاي هوشمند از طريق ارتباط علم زبانشناسي و رايانه

۱۲- مطالعه ساختار و صرف و نحو زبانهاي طبيعي با استفاده از علم منطق زبانشناسي

۱۳- بررسي چگونگي ساختار (ميانجي) واسط هاي زبانهاي طبيعي براي سيستمهاي رايانه اى

۱۴- بررسي راهبردهاى تبديل گفتار به نوشتار و بالعكس با استفاده از علم و منطق زبانشناسي

۱۵- مطالعه خلاصه سازي خودكار در رايآنه با استفاده از علم و منطق زبانشناسي

۱۶- بررسي چگونگي رو در رويي (ارتباط) زبان طبيعي و سيستمهاي رايانه اى

۱۷- برچسب زني كلمات جهت بازشناسي رايانه اي، با استفاده از منطق زبانشناسي

۱۸- بررسي ساختار دستورهاي بافت وابسته و مستقل از بافت جهت تعيين الگوهايي براي بازيابي اطلاعات.

طرحهاي تحقيقاتي

پايان يافته در سال ۱۳۸۸

مفردسازي ماشيني صيغه هاي جمع اسامي در زبان فارسي

مجري: شاپوررضا برنجيان

تاريخ شروع: ۸۶/۶/۱۴

تاريخ پايان: ۸۸/۸/۳۰

چكيده:

اهميت بررسي و تحقيقات در حيطه زبان فارسي بويژه در برخورد اين زبان و خط آن با علم رايانه در ماشينهاي ترجمه، ريشه ياب و ... بر كسي پوشيده نيست، اگر چه تاكنون پژوهشهايي نيز در اين زمينه صورت پذيرفته، امّا هنوز موارد بسياري يافت مي شود كه هيچ گونه بررسي درباره آنها انجام نپذيرفته و تحقيقات در اين زمينه ضروري بنظر مي رسد. مفردسازي صيغه هاي جمع اسامي نيز يكي از اين موارد مي باشد.

مسئلة اسم فارسي و ويژگي هاي آن در برابر موتورهاي جستجو در شبكه هاي كامپيوتري نيز سالهاست كه مورد توجه اطلاع رسانها قرار گرفته و اين امر كم و بيش در كتابخانه ها و مراكز اطلاع رساني مورد بررسي قرار گرفته است، طرح حاضر تلاشي است در جهت هموارتر كردن اين راه ناهموار.

طرحهاي در دست اجرا

در سال ۱۳۸۸

ريشه ياب ماضي و مضارع از مصدر
افعال ناگذر در زبان فارسي

مجري: شاپوررضا برنجيان

تاريخ شروع: ۸۷/۱۰/۱

چكيده:

تقريباً تمامي زبانهاي موجود در دنيا در هر برهه از زمان بدليل ماهيت خود زبان اقدام به واژه سازي مي كنند، اين عمل از دو طريق (اشتقاق و تصريف) صورت مي گيرد، كه اكثر آنها از بن هاي مضارع يا ماضي افعال ساخته مي شوند. مانند: خواستار، گفتار، كردار، دانش ها و ... و همچنين صرف افعال فارسي (در تمام زمانها) نياز به مشخص نمودن بن ماضي و مضارع افعال دارند، از اين رو ساخت ريشه يا بن هاي ماضي و مضارع در افعال فارسي در ذخيره سازي و بازيابي اطلاعات، سيستمهاي ماشينهاي ترجمه، تبديل متن به گفتار، گفتار با متن و ... از اهميت خاصي برخوردار است.

چالش هاي پردازش زبان طبيعي فارسي

مجري طرح: دكتر حميد عليزاده

زمان شروع: ۸۸/۴/۴

چكيده:

امروزه اهميت كاربرد اصول و تكنيک هاي پردازش زبان طبيعي در بازيابي اطلاعات به وضوح اثبات شده است. اما عليرغم آنكه در زبان هاي پركاربردي همچون انگليسي، فرانسه، چيني و ... پژوهش هاي بسياري در زمينه چالش هاي موجود براي بكارگيري اصول پردازش زبان طبيعي انجام شده است. در زبان فارسي جدا از چند كار پراكنده كه هركدام وجهي از كاربردهاي اين حوزه را درنظر گرفته اند. پژوهشي جامع كه ظرفيت ها، امكانات و كاستي هاي زبان فارسي را براي كاربرد اصول پردازش زبان طبيعي بررسي نمايد انجام نشده است. در جهت طراحي و ساخت ابزار پردازش زبان طبيعي نظير ريشه سازها و ابزار تحليل انواع كلام، ابتدا بايد ساختارهاي مربوط به زبان فارسي تحليل گردد تا چنين ابزارهايي با درک صحيح از نيازمندي ها و امكانات موجود طراحي گردد. طرح پژوهشي حاضر، با بررسي چالش هاي موجود در پردازش زبان طبيعي فارسي، به روشن سازي مسائل اصلي از قبيل امكانات و ساختار خط فارسي، الگوهاي زباني، وضعيت زبان فارسي در پشتيباني از نظام هايي مثل مترجم هاي ماشيني و استخراج اطلاعات پرداخته و گزارش نهايي آن كه به شكل يک كتاب ارائه خواهد شد، به عنوان مبناي دانش براي پژوهش هاي مربوط به اين حوزه در گروه زبانشناسي رايانه اي مركز منطقه اي اطلاع رساني علوم و فناوري و ساير پژوهشگران استفاده خواهد شد.

تکواژگونه و واجگونه

نویسنده: شاپوررضا برنجیان

ارایه شده در فصلنامه علمي – پژوهشي اباختر پائیز ۱۳۸۸ سال پنجم شماره پیاپي ۱۷ و ۱۸

چکیده:

زبان دستگاهي است که از اصوات تشکیل شده است و کارکرد اجتماعي دارد و زبان شناس ابتدا آنرا به صورت صوري مورد تجزیه و تحلیل قرار داده و تقطیع مي کند تا بتواند آن را به عنوان نظامي از نشانه ها تعریف کند.

در این مقاله بر آن هستیم تا ضمن اشاره به واحدهاي گفتار در زبان فارسي به بررسي تکواژ و واج پرداخته و سپس تعریفي از تکواژگونه و واجگونه ارائه نمائیم.

بررسي کارآمدي روش هاي موجود در بازیابي اطلاعات بین زباني فارسي - انگلیسي

نویسنده: دکتر حمید علیزاده

ارایه شده در فصلنامه علوم و فناوري اطلاعات دوره ۲۵، شماره ۱، ۱۳۸۸

مروري بر مدل هاي نوین دسترسي آزاد به نتایج تحقیقات علمي

نویسنده: دکتر حمید علیزاده

ارایه شده در فصلنامه کتاب ماه کلیات، سال دوازدهم، شماره دوازدهم، ۱۳۸۸

محورهاي زبان

سخنران: شاپوررضا برنجيان

تاريخ ارائه: ۸۸/۷/۱۵

مكان ارائه: مركز منطقه اي

چكيده:

موضوع اصلي چگونگي بازيابي اطلاعات از كامپيوتر است. در بازيابي متني ابتدا كاربر متني را درنظر دارد. مانند: "فرزندان ايران در نيروهاي مسلح خدمت مي‌كنند" كه در زمان بازيابي مي بايست به جاي هر يک از واژه ها در زنجيره ي افقي بتوان واژه هاي هم معني را نيز جايگزين كرد، مانند: "جوان" يا "كودک" بجاي "فرزند" تا بتوان اسناد و مدارک مرتبط با آن موضوع بازيابي گردد. صرف نظر از اصطلاحات، هر يک از اجزاي آن عبارت در زنجيره به نوبت در معرض بازيابي قرار مي گيرند.

بررسي كارآمدي روشهاي موجود در بازيابي اطلاعات بين زباني

سخنران: دكتر حميد عليزاده

زمان سخنراني: ۸۸/۴/۲۳

مكان: مركز منطقه اي

كاربرد پردازش زبان طبيعي در بازيابي اطلاعات بين زباني

سخنران: دكتر حميد عليزاده

زمان سخنراني: ۸۸/۸/۱۸

مكان: تهران، انجمن كتابداري و اطلاع رساني

مروري بر اصول و مفاهيم بازيابي بين زباني و كاربرد آن

سخنران: دكتر حميد عليزاده

زمان: بهمن ماه ۱۳۸۸

مكان: دانشگاه شيراز

كارگاههاي آموزشي

۱- كارگاه آموزشي آشنايي با ISC در دانشگاه پيام نور شيراز

مجري: دكتر حميد عليزاده

۲- كارگاه آموزشي طلايه داران علم ايران در دانشگاه اصفهان

مجري: دكتر حميد عليزاده

۳- كارگاه آموزشي طلايه داران علم ايران در دانشگاه آزاد مرودشت

مجري: دكتر حميد عليزاده