دانشکده علوم تربیتی و روانشناسی
گروه آموزشی کتابداری و اطلاع رسانی

پایان نامه دکترای کتابداری و اطلاع رسانی

بررسی کارآمدی روش های موجود در بازیابی اطلاعات بین زبانی فارسی- انگلیسی با استفاده از واژه نامه دوزبانه ماشین خوان

استادان راهنما

دکتر رحمت الله فتاحی
دکتر محمد رضا داورپناه

استاد مشاور
دکتر محمد حسین دیانی

نگارش
حمید علیزاده

تابستان ۱۳۸۸

تقدیم به:

همسر مهربانم

به پاس همراهی همیشگی اش...

و

پرند عزیزم

که حضورش شادی بخش همه لحظه های زندگی است...

قدردانی و سپاس:

با تشکر و سپاس از همه آنها که در مراحل مختلف زندگی یاری ام کردند. بویژه پدر و مادرم که همیشه قدردان زحماتشان هستم.

از راهنمایی ها ی ارزشمند استاد گرامی جناب آقای دکتر فتاحی که همواره در طول مدت تحصیل و نگارش این پژوهش مرا از نظرات ارزشمندشان بهره مند ساختند تشکر می نمایم.

از راهنمایی های همیشگی و دقت نظر مدبرانه استاد گرامی جناب آقای دکتر داورپناه سپاسگزاری می کنم.

از مشاوره های ارزشمند استاد گرامی جناب آقای دکتر دیانی نیز صمیمانه تشکر می کنم.

از داوران پایان نامه استاد گرانپایه جناب آقای دکتر مهراد و سرکار خانم دکتر صنعت جو به خاطر تیزبینی و نکات ارزشمندی که موجب ارتقاء کیفی این پژوهش شد سپاسگزاری می کنم.

از همدوره ای ها و سایر دوستان در دانشگاه فردوسی و مرکز منطقه ای اطلاع رسانی علوم و فناوری نیز ممنون و سپاسگزارم.

چکیده

پژوهش حاضر به بررسی کارآمدی شیوه های موجود در بازیابی اطلاعات بین زبانی(بازبین)فارسی- انگلیسی با رویکرد واژه نامه دوزبانه ماشین خوان پرداخته است. این پژوهش کاربردی بوده و با استفاده از روش های پژوهش نیمه تجربی و تحلیل

محتوا انجام شده است. در این پژوهش ضمن بررسی میزان تاثیر انجام پردازش های زبان طبیعی بر روی ترجمه عبارت های جستجو، میزان و نحوه این تاثیر با آزمون فرضیه های پژوهش مشخص گردید. فنون پردازش زبان طبیعی که در این پژوهش بکار گرفته شد شامل قطعه بندی متن، شناخت گونه های زبانشناختی، حذف سیاهه بازدارنده، تحلیل مورفولوژیک و برچسب زنی انواع کلام بوده است. جامعه آماری این پژوهش پیشینه های موجود در موتور جستجوی گوگل بوده است که با استفاده از ۴۰ عبارت جستجو، بازیابی شده و آنگاه با استفاده از روش نمونه گیری مخزن سازی و انجام قضاوت های ربط، داده های مورد نیاز برای انجام این پژوهش مهیا شده است. آزمون فرضیه اول این پژوهش نشان داد که در هنگام ترجمه عبارت های جستجوی فارسی به انگلیسی با واژه نامه دوزبانه ماشین خوان، استفاده از روش ترجمه انتخاب اولین برابرنهاده منجر به دستیابی به میانگین متوسط دقت بازیافت(MAP) معادل ۲۲۳/. شد که در مقایسه با شیوه انتخاب همه برابرنهاده ها که میانگین متوسط دقت بازیافت برای آن ۱۶۸/. بود موجب کارآمدی بیشتر بازبین می گردد. آزمون فرضیه دوم نشان داد که اگرچه تحلیل مورفولوژیک واژه هایی که به وسیله واژه نامه ترجمه نشدند باعث افزایش ضریب دقت بازیافت می گردد اما تفاوت معناداری با عدم انجام این تحلیل ایجاد نمی نماید. بررسی فرضیه سوم این پژوهش نیز نشان داد که ترجمه عبارتی در مقایسه با ترجمه واژه به واژه باعث کارآمدی بیشتر بازبین فارسی- انگلیسی می گردد. نمرات میانگین متوسط دقت بازیافت برای این دو شیوه ترجمه به ترتیب ۳۱۹/. و ۲۲۳/. بود. نتایج دیگر این پژوهش نیز نشان داد که دگرنویسی واژه های فارسی ترجمه ناپذیر به حروف زبان انگلیسی و قرار دادن آنها در عبارت جستجوی نهایی در مقایسه با حذف آنها از عبارت های جستجو می تواند منجر به افزایش کارآمدی بازبین فارسی- انگلیسی به میزان ۷۴٪ گردد. در نهایت بررسی های انجام شده نشان داد که بازبین فارسی- انگلیسی در بهترین حالت بازیابی، به ۶۴٪ کارآمدی بازیابی یک زبانه عبارت های جستجوی انگلیسی دست می یابد.

کلیدواژه ها: بازیابی اطلاعات بین زبانی، واژه نامه دو زبانه ماشین خوان، پردازش زبان طبیعی

فهرست مندرجات

فهرست پیوست ها

فهرست جدول ها

فصل اول

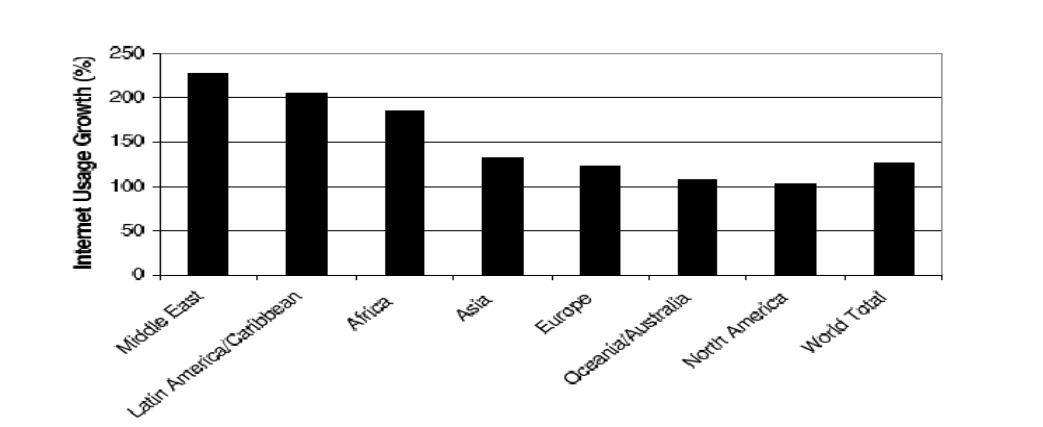
درآمدی بر پژوهش

۱-۱. مقدمه

با گسترش روزافزون استفاده از اینترنت و غلبه بر محدودیت های فنی و شبکه ای که به مدد توسعه فناوری اطلاعات و ارتباطات حاصل شده است، کاربران و جستجوگران اطلاعات تنها به منابع اطلاعاتی که به زبان آنها نوشته شده است اکتفا نمی کنند(Adriani, 2000) بلکه دسترسی به همه اطلاعات مرتبط در سایر زبانها را دیگر نه آرزو، بلکه حق طبیعی خود می دانند. امروزه وظیفه بازیابی اطلاعات به فرایندهای سنتی آن محدود نمی شود، بلکه هدفهای گسترده تر، یعنی غلبه بر موانع زبانی در هنگام جستجو و بازیابی اطلاعات نیز در این حوزه مطرح شده است.

امروزه تعداد زبان های زنده دنیا را چیزی حدود ۴۵۰۰ زبان تخمین می زنند، که از میان آن ها در حدود ۳۰ زبان وجود دارد که هرکدام توسط حداقل ۳۰ میلیون نفر استفاده می شود (Edwards, 1994). بدیهی است که برای تبادل اطلاعات در این جامعه اطلاعاتی چند زبانه، دیگر مطلوب نیست که به اطلاعات یک زبان خاص محدود شد.

اینترنت به عنوان محل تردد این زبان ها بیشترین نمود این گوناگونی را به خود اختصاص داده است. آمارها نشان می دهد که استفاده از اینترنت در چند سال اخیر رشد قابل ملاحظه ای داشته است. این نرخ رشد به ویژه در خاورمیانه، آمریکای جنوبی و آفریقا بسیار چشمگیر است(این نکته در شکل ۱,۱ نشان داده شده است). این تنوع جغرافیایی با تنوع زبانی نیز همراه است. بطوری که با رشد منابع اینترنتی، مشکلات و هم سودمندی دسترسی و بهره گیری از منابع به زبانهای دیگر نیز بیشتر شده است.



شکل ۱-۱ رشد بهره گیری از اینترنت در سال ۲۰۰۵ در مقایسه با سال ۲۰۰۰ ماخذ: (Wang, 2005)

یکی از راه حل های غلبه بر این مشکلات، بهره گیری از بازیابی اطلاعات بین زبانی[۱] (بازبین) است. بازیابی اطلاعات بین زبانی نوعی از بازیابی اطلاعات است که در آن حداقل دو زبان حضور دارد، زبان عبارت جستجو[۲] و زبان مجموعه مدرک. زبان عبارت جستجو را زبان اصلی[۳] و زبان مجموعه مدرک را زبان هدف یا مقصد[۴]می نامند. یک نظام بازیابی اطلاعات بین زبانی(بازبین) مدرک را در زبانی که با زبان عبارت جستجو متفاوت است بازیابی می کند. در این شیوه، کاربر نظام بازبین عبارت جستجو را به زبان بومی خویش ارائه می کند، اما مدارک دریافتی به زبان مجموعه مدرک خواهد بود. نظام بازبین کار جستجوگرانی که به چند زبان تسلط دارند را ساده می کند و در عین حال جستجوگرانی را که تنها به یک زبان تسلط دارند، قادر می سازد عبارت جستجو را به زبان خود ارائه کنند و آنگاه با استفاده از دانش خود یا با بهره گیری ازکمک دیگران، بین مدارک بازیابی شده تمایز قائل شوند و سپس مدارکی را که مربوط تشخیص داده می شود، با

[1]- Cross Language Information Rretrieval(CLIR)
[2]- Query
[3] - Source Language
[4] - Target Language

استفاده از عامل انسانی یا ماشینی ترجمه نموده و مورد استفاده قرار دهند (Ballesteros& Croft, 1998).

به طور کلی کاربران این نظام را می توان به چند دسته تقسیم کرد: دسته اول کاربرانی هستند که تا حدودی توانایی خواندن مدرک به یک زبان دیگر، غیر از زبان اصلی خود را دارند، ولی در ابراز دقیق نیاز اطلاعاتی خویش به آن زبان ناتوانند. در این صورت، چنین کاربری می تواند عبارت جستجو را به زبان خود بنویسد تا نظام اطلاعاتی مدارک را در زبان دیگر برای وی جستجو و بازیابی کند. دسته بعدی کاربران نظام بازبین، کاربرانی هستند که به چند زبان تسلط دارند و در یک مجموعه مدرک چند زبانه(مثل اینترنت) جستجو می کنند. چنین کاربری می تواند عبارت جستجو را به یک زبان ارائه کرده و مدارک را به زبانهای مختلف بازیابی کند. این عمل به صرفه جویی در وقت و انرژی کاربر می انجامد، زیرا دیگر مجبور نیست سؤال را به زبان های مختلف ترجمه و چند بار جستجو کند. دسته سوم کاربرانی هستند که هیچگونه آشنایی با زبان خارجی ندارند. این کاربران، در صورت دسترسی به یک نظام یا منبع ترجمه (انسانی یا ماشینی)، می توانند با بهره گیری از نظام بازبین به تعدادی مدارک مرتبط با جستجوی اطلاعاتی خویش دست یابند و آنگاه با ترجمه مدارک به زبان خود، از آنها استفاده کنند. با این عمل از تعداد مدارکی که کاربر مجبور به ترجمه آن است کاسته می شود.

در بازیابی اطلاعات بین زبانی، هم مدرک و هم عبارت جستجو می تواند ترجمه شود. چون ترجمه عبارت جستجو در مقایسه با ترجمه مدرک هم ارزان تر است و هم به صرف وقت و کار علمی کم تری نیاز دارد، در پژوهش های انجام شده در این حوزه بیشتر به ترجمه عبارت جستجو توجه شده است(Oard, 1997; Ballesteros& Croft, 1996). در ترجمه عبارت