



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

پرنده 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 های 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 مهرداد 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 ۰/۲۲۳ شد که در مقایسه با شیوه انتخاب همه برابرنهاده ها که میانگین متوسط دقت بازیافت برای آن ۰/۱۶۸ 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 ۰/۳۱۹ و ۰/۲۲۳ 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 ۰/۷۴ گردد. در نهایت بررسی های انجام شده نشان داد که بازبین فارسی- انگلیسی در بهترین حالت بازیابی، به ۰/۶۴ 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

۱	فصل اول: درآمدی بر پژوهش
۲	۱-۱. مقدمه

۶	۱-۱-۱. بازبین فارسی - انگلیسی
۷	۲-۱-۱. رویکرد این پژوهش
۱۰	۲-۱. بیان مساله
۱۱	۳-۱. ضرورت پژوهش
۱۲	۴-۱. هدف های پژوهش
۱۳	۵-۱. فرضیه های پژوهش
۱۴	۶-۱. تعریف های عملیاتی و مفاهیم کلیدی
۱۸	فصل دوم: مبانی نظری و پیشینه پژوهش
۱۹	۱-۲. مقدمه
۱۹	۲-۲. زیربنای نظری بازبین
۱۹	۱-۲-۲. پردازش زبان طبیعی و بازبین
۲۱	۲-۲-۲. تحلیل ساخت واژه
۲۳	۳-۲-۲. تحلیل نحوی
۲۵	۴-۲-۲. حذف واژه های سیاهه بازدارنده
۲۵	۵-۲-۲. زبان فارسی و بازبین
۲۷	۶-۲-۲. ترجمه عبارت های جستجو
۲۸	۷-۲-۲. دشواری های ترجمه
۳۲	۳-۲. رویکردهای بازیابی اطلاعات بین زبانی

۳۲	۲-۳-۱ رویکردهای موجود
۳۴	۲-۳-۲. رویکرد مجموعه متن
۳۵	۲-۳-۳. رویکرد مبتنی بر دانش
۳۵	۲-۳-۴. ترجمه ماشینی
۳۷	۲-۳-۵. ساختارهای دانش
۳۸	۲-۴. پیشینه علمی و مرور نوشته های مربوط
۳۸	۲-۴-۱. پیشینه علمی و مرور نوشته های مربوط در داخل کشور
۳۹	۲-۴-۲. پردازش رایانه ای زبان فارسی
۴۱	۲-۴-۳. ترجمه ماشینی
۴۲	۲-۵. پیشینه علمی و مرور نوشته های مربوط در خارج
۴۴	۲-۵-۱. ترجمه عبارتی
۴۶	۲-۵-۲. ابهام در ترجمه
۴۹	۲-۵-۳. واژه های ترجمه ناپذیر یا خارج از واژه نامه
۵۰	۲-۶. استنتاج از مرور پیشینه پژوهش

۵۴	فصل سوم: روش شناسی پژوهش
۵۵	۳-۱. روش شناسی پژوهش
۵۷	۳-۱-۲. شیوه انجام پژوهش
۵۷	۳-۱-۳. مرحله اول
۶۰	۳-۱-۴. مرحله دوم
۶۲	۳-۱-۵. مرحله سوم
۶۳	۳-۲. جامعه آماری، حجم نمونه و روش نمونه گیری
۶۴	۳-۳. شیوه گردآوری و پردازش داده ها
۷۰	۳-۴. روایی و پایایی ابزار پژوهش
۷۳	۳-۵. شیوه تجزیه و تحلیل یافته ها
۷۵	۳-۶. محدودیت ها و مشکلات پژوهش
۷۸	فصل چهارم : تجزیه و تحلیل یافته ها
۷۹	۴-۱. مقدمه
۷۹	۴-۱-۱. ویژگی های جامعه آماری
۸۲	۴-۲. پردازش زبان طبیعی
۸۳	۴-۳. ارائه توصیفی یافته ها

	و پاسخگویی به فرضیه های پژوهش
۸۳	۴-۳-۱. بررسی فرضیه اول
۹۶	۴-۳-۲. بررسی فرضیه دوم
۱۰۵	۴-۳-۳. بررسی فرضیه سوم
۱۱۲	۴-۳-۴. بررسی فرضیه چهارم
۱۲۰	۴-۳-۵. بررسی فرضیه پنجم
۱۲۸	۴-۴. جمع بندی کلی
	فصل پنجم: بحث و نتیجه گیری
۱۲۹	
۱۳۰	۵-۱. مقدمه
۱۳۰	۵-۱-۱. خلاصه یافته ها
۱۳۵	۵-۱-۲. بحث و بررسی
۱۴۳	۵-۱-۳. مدل پیشنهادی نظام بازیبن فارسی - انگلیسی
۱۴۶	۵-۲. پیشنهاد برای پژوهش های آتی
۱۴۸	۵-۳. کلام پایانی
۱۴۹	فهرست منابع و مآخذ

فهرست پیوست ها	
۱۵۹	پیوست ها
۱۶۰	پیوست ۱. سیاهه عبارت های جستجوی اصلی به زبان انگلیسی
	پیوست ۲. نتایج ترجمه واژه به واژه عبارت های جستجو
۱۷۰	پیوست ۳. نتایج تحلیل ساخت واژه (مورفولوژیک) اصطلاح های ترجمه
۱۷۸	نشده عبارت های جستجو
	پیوست ۴. نتایج برچسب زنی انواع کلام و استخراج عبارت های احتمالی از
۱۸۵	عبارت های جستجو
	پیوست ۵. سیاهه واژه های دگرنویسی شده
۱۹۱	پیوست ۶. سیاهه واژه های تهی یا سیاهه بازدارنده
۱۹۳	

فهرست جدول ها

۸۱	جدول ۱-۴. مشخصات واژگانی عبارت های جستجو
۸۲	جدول ۲-۴. سیاهه مراحل مختلف پردازش زبان طبیعی که در این پژوهش انجام شد

۸۵	جدول ۴-۳. نتایج انتخاب اولین برابرنهاده در هنگام ترجمه عبارت جستجو
۸۶	جدول ۴-۴. انتخاب دومین برابرنهاده در هنگام ترجمه عبارت جستجو
۸۸	جدول ۴-۵. انتخاب سومین برابرنهاده در هنگام ترجمه عبارت جستجو
۸۹	جدول ۴-۶. انتخاب چهارمین برابرنهاده در هنگام ترجمه عبارت جستجو
۹۰	جدول ۴-۷. مقایسه میانگین متوسط دقت بازیافت رویکردهای انتخاب اولین برابرنهاده و همه برابرنهاده ها
۹۲	جدول ۴-۸. محاسبه میانگین دقت بازیافت در سطوح مختلف بازیابی بر اساس انتخاب برابرنهاده های مختلف
۹۳	جدول ۴-۹. نتایج آزمون بررسی نرمال بودن

	داده های مربوط به متغیرهای انتخاب اولین برابرنهاده و همه برابرنهاده ها
۹۴	جدول ۴-۱۰. نتایج آزمون لون برای همسانی واریانس ها
۹۵	جدول ۴-۱۱. داده های توصیفی کلی مربوط به متغیرهای اولین برابرنهاده و همه برابرنهاده ها
۹۵	جدول ۴-۱۲. نتایج آزمون T برای تفاوت میانگین مربوط به متغیرهای اولین برابرنهاده و همه برابرنهاده ها
۹۷	جدول ۴-۱۳. متوسط دقت بازیافت عبارت های جستجو بدون پردازش مورفولوژیک
۹۹	جدول ۴-۱۴. میانگین دقت بازیافت عبارت های جستجو با استفاده از پردازش مورفولوژیک
۱۰۰	جدول ۴-۱۵. مقایسه بازیبن با استفاده از پردازش مورفولوژیک واژه های ترجمه نشده و عدم انجام این پردازش
۱۰۱	جدول ۴-۱۶. محاسبه میانگین دقت بازیافت

	<p>در سطوح مختلف بازیابی بر اساس مقایسه انجام پردازش مورفولوژیک واژه های ترجمه نشده و عدم انجام این پردازش</p>
۱۰۲	<p>جدول ۴-۱۷. بررسی نرمال بودن داده های مربوط به متغیرهای ترجمه با پردازش مورفولوژیک و ترجمه بدون پردازش مورفولوژیک</p>
۱۰۳	<p>جدول ۴-۱۸. آزمون همسانی واریانسهای متغیرهای ترجمه با پردازش مورفولوژیک و ترجمه بدون پردازش مورفولوژیک</p>
۱۰۳	<p>جدول ۴-۱۹. داده های توصیفی متغیرهای ترجمه با پردازش مورفولوژیک و ترجمه بدون پردازش مورفولوژیک</p>
۱۰۴	<p>جدول ۴-۲۰. نتایج آزمون T برای تفاوت میانگین مربوط به متغیرهای ترجمه با پردازش مورفولوژیک و ترجمه بدون پردازش مورفولوژیک</p>
۱۰۶	<p>جدول ۴-۲۱. نتایج بازیابی با استفاده از ترجمه عبارتی</p>

۱۰۷	جدول ۴-۲۲. نتایج بازبین با استفاده از ترجمه واژه به واژه
۱۰۸	جدول ۴-۲۳. مقایسه کارآمدی شیوه ترجمه عبارتی با شیوه ترجمه واژه به واژه
۱۰۹	جدول ۴-۲۴. نتایج بررسی میانگین دقت باز یافت در سطوح مختلف بازیابی بر اساس ترجمه عبارتی و ترجمه واژه به واژه
۱۱۰	جدول ۴-۲۵. بررسی نرمال بودن داده های مربوط به متغیرهای ترجمه عبارتی و ترجمه واژه به واژه
۱۱۱	جدول ۴-۲۶. آزمون همسانی واریانس های متغیرهای ترجمه عبارتی و ترجمه واژه به واژه
۱۱۱	جدول ۴-۲۷. داده های توصیفی مربوط به متغیرهای ترجمه عبارتی و ترجمه واژه به واژه
۱۱۲	جدول ۴-۲۸. نتایج آزمون T دو نمونه مستقل برای متغیرهای ترجمه عبارتی و ترجمه واژه به واژه

۱۱۳	<p>جدول ۴-۲۹. نتایج متوسط دقت بازیافت عبارت های جستجوی انگلیسی در بازیابی یک زبانه</p>
۱۱۵	<p>جدول ۴-۳۰. نتایج متوسط دقت بازیافت عبارت های جستجو در بازیابی اطلاعات بین زبانی فارسی - انگلیسی</p>
۱۱۶	<p>جدول ۴-۳۱. مقایسه میانگین متوسط دقت بازیافت بازیابی اطلاعات یک زبانه انگلیسی و بازیابی اطلاعات بین زبانی فارسی - انگلیسی</p>
۱۱۷	<p>جدول ۴-۳۲. نتایج مقایسه میانگین دقت بازیافت بازیابی یک زبانه و بازیابی فارسی - انگلیسی در سطوح مختلف بازیابی</p>
۱۱۷	<p>جدول ۴-۳۳. نتایج آزمون نرمال بودن داده های مربوط به متغیرهای بازیابی یک زبانه و بین زبانی</p>
۱۱۸	<p>جدول ۴-۳۴. آزمون همسانی واریانس های متغیرهای</p>

	بازیابی یک زبانه و بازیابی بین زبانی
۱۱۹	جدول ۴-۳۵. داده های توصیفی مربوط به متغیرهای بازیابی یک زبانه و بازیابی بین زبانی
۱۱۹	جدول ۴-۳۶. نتایج آزمون T برای تفاوت میانگین مربوط به متغیرهای بازیابی یک زبانه و بازیابی بین زبانی
۱۲۰	جدول ۴-۳۷. نتایج متوسط دقت بازیافت بازیبن با حذف واژه های ترجمه ناپذیر و خارج از واژه نامه
۱۲۲	جدول ۴-۳۸. نتایج متوسط دقت بازیافت بازیبن با استفاده از دگرنویسی واژه های ترجمه ناپذیر و خارج از واژه نامه
۱۲۳	جدول ۴-۳۹. مقایسه میانگین متوسط دقت بازیافت بازیبن با استفاده از دگرنویسی و بازیبن بدون استفاده از دگرنویسی
۱۲۴	جدول ۴-۴۰. مقایسه میانگین دقت بازیافت

۱۲۵	<p>بازبین با استفاده از دگرنویسی و بدون استفاده از دگرنویسی در سطوح مختلف بازیابی</p> <p>جدول ۴-۴۱. نتایج بررسی نرمال بودن داده های مربوط به متغیرهای بازبین با دگرنویسی و بازبین بدون دگرنویسی</p>
۱۲۶	<p>جدول ۴-۴۲. نتایج آزمون همسانی واریانس های متغیرهای بازبین با دگرنویسی و بازبین بدون دگرنویسی</p>
۱۲۷	<p>جدول ۴-۴۳. داده های توصیفی مربوط به متغیرهای بازبین با دگرنویسی و بازبین بدون دگرنویسی</p>
۱۲۷	<p>جدول ۴-۴۴. نتایج آزمون T برای تفاوت میانگین مربوط به بازبین با دگرنویسی و بازبین بدون دگرنویسی</p>

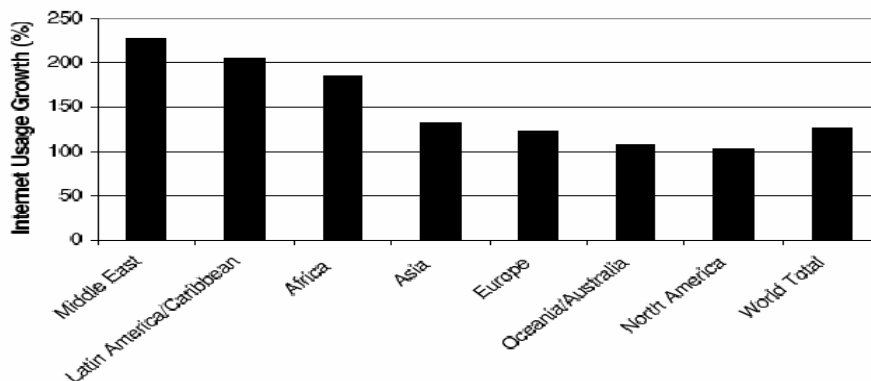
فصل اول

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 بهره‌گیری از اینترنت در سال ۲۰۰۵ در مقایسه با سال ۲۰۰۰ ماخذ: (Wang, 2005)

یکی از راه‌حل‌های غلبه بر این مشکلات، بهره‌گیری از بازیابی اطلاعات بین‌زبانی^۱ (بازبین) است. بازیابی اطلاعات بین‌زبانی 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 می‌نامند. یک نظام بازیابی اطلاعات بین‌زبانی (بازبین) مدرک را در زبانی که با زبان عبارت جستجو متفاوت است بازیابی می‌کند. در این شیوه، کاربر نظام بازبین عبارت جستجو را به زبان بومی خویش ارائه می‌کند، اما مدارک دریافتی به زبان مجموعه مدرک خواهد بود. نظام بازبین کار جستجوگرانی که به چند زبان تسلط دارند را ساده می‌کند و در عین حال جستجوگرانی را که تنها به یک زبان تسلط دارند، قادر می‌سازد عبارت جستجو را به زبان خود ارائه کنند و آنگاه با استفاده از دانش خود یا با بهره‌گیری از کمک دیگران، بین مدارک بازیابی شده تمایز قائل شوند و سپس مدارکی را که مربوط تشخیص داده می‌شود، با

^۱ - Cross Language Information Retrieval (CLIR)

^۲ - Query

^۳ - Source Language

^۴ - 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