بررسی کارآمدی روش‌های موجود در بازیابی اطلاعات بین زبانی فارسی- انگلیسی با استفاده از واژه‌نامهٔ دوزبانهٔ ماشین‌خوان

حمید علیزاده*
دکترای کتابداری و اطلاع‌رسانی
استادیار مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

رحمت‌الله فتاحی
دکترای کتابداری و اطلاع‌رسانی
استاد دانشگاه فردوسی مشهد

محمدرضا داورپناه
دکترای کتابداری و اطلاع‌رسانی
دانشیار دانشگاه فردوسی مشهد

دریافت: ۱۳۸۸/۰۶/۱۱ **پذیرش:** ۱۳۸۸/۰۸/۱۱ مقاله برای اصلاح به مدت یک ماه و شش روز نزد پدیدآوران بوده است.

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا (چاپی) ۵۲۰۶-۱۷۳۵
شاپا (الکترونیکی) ۵۵۸۳-۲۰۰۸
نمایه در LISA و SCOPUS
http://jist.irandoc.ac.ir
دورهٔ ۲۵ | شمارهٔ ۱ | ص‌ص ۵۳-۷۰
پاییز ۱۳۸۸

نوع مقاله: علمی پژوهشی

چکیده: در این پژوهش میزان تأثیر انجام پردازش‌های زبان طبیعی بر روی ترجمهٔ عبارت‌های جستجو با آزمون فرضیه‌های پژوهش مشخص گردید. فنون پردازش زبان طبیعی که برای پردازش عبارت‌های جستجو به کار گرفته شد شامل قطعه‌بندی متن، شناخت گونه‌های زبانشناختی، حذف سیاههٔ بازدارنده، تحلیل مورفولوژیک، و برچسب‌زنی انواع نقش دستوری بود. آزمون فرضیهٔ اول نشان داد که استفاده از روش ترجمهٔ اولین برابرنهاده در مقایسه با شیوهٔ انتخاب همهٔ برابرنهاده‌ها موجب کارآمدی بیشتر در بازیابی می‌گردد. آزمون فرضیهٔ دوم نشان داد که اگرچه تحلیل مورفولوژیک واژه‌هایی که به وسیلهٔ واژه‌نامه ترجمه نشدند باعث افزایش ضریب دقت بازیافت می‌گردد، اما تفاوت معناداری با عدم انجام این تحلیل ایجاد نمی‌نماید. بررسی فرضیهٔ سوم نیز نشان داد که ترجمهٔ عبارتی در مقایسه با ترجمهٔ واژه به واژه باعث کارآمدی بیش‌تر می‌گردد. یافته‌های دیگر این پژوهش نیز نشان داد که دگرنویسی واژه‌های فارسی ترجمه‌ناپذیر با حروف انگلیسی و قرار دادن آن‌ها در عبارت جستجوی نهایی در مقایسه با حذف آن‌ها از عبارت‌های جستجو، می‌تواند منجر به افزایش کارآمدی گردد.

کلیدواژه‌ها: بازیابی اطلاعات بین زبانی؛ پردازش زبان طبیعی؛ واژه‌نامهٔ دوزبانهٔ ماشین‌خوان؛ ارزیابی بازیابی اطلاعات

* پدیدآور رابط halizade@gmail.com

۱. مقدمه

با گسترش روزافزون استفاده از اینترنت و غلبه بر محدودیت‌های فنی و شبکه‌ای که به مدد توسعهٔ فناوری اطلاعات و ارتباطات حاصل شده است، کاربران و جستجوگران اطلاعات دیگر تنها به منابع اطلاعاتی که به زبان آن‌ها نوشته شده اکتفا نمی‌کنند. دسترسی به همهٔ اطلاعات مرتبط در دیگر زبان‌ها، اکنون نه آرزو، بلکه حق طبیعی کاربران شناخته می‌شود. این تنوع زبانی اگرچه در ابتدا مفید به نظر می‌رسد، اما می‌تواند مانعی برای دسترسی به اطلاعات تلقی شود (علیزاده ۱۳۸۳). بنابراین امروزه بازیابی اطلاعات به فرایندهای سنتی آن خلاصه نمی‌شود، بلکه هدف‌های بزرگ‌تر (یعنی غلبه بر موانع زبانی در هنگام جستجو و بازیابی اطلاعات) نیز در این حوزه مطرح شده است.

«ادواردز» تعداد زبان‌های زندهٔ دنیا را چیزی حدود ۴۵۰۰ زبان تخمین می‌زند که از میان آن‌ها در حدود ۳۰ زبان هستند که هر یک توسط حداقل ۳۰ میلیون نفر استفاده می‌شود (Edwards 1994, 15). بدیهی است که برای تبادل اطلاعات در این جامعهٔ اطلاعاتی چندزبانه، دیگر نمی‌توان به یک زبان خاص محدود شد. اینترنت به عنوان محل تلاقی این زبان‌ها، بیش‌ترین نمود این گوناگونی را در خود جا داده است. آمارها نشان می‌دهد که استفاده از اینترنت در چند سال اخیر رشد قابل ملاحظه‌ای داشته است. این نرخ رشد بویژه در خاورمیانه، آمریکای جنوبی، و آفریقا بسیار چشمگیر است (Wang 2005). این تنوع جغرافیایی با تنوع زبانی نیز همراه است، به‌طوری که با رشد منابع اینترنتی، مشکلات زبانی در دسترسی و بهره‌گیری از این منابع نیز بیش‌تر شده است.

راه حل غلبه بر این مشکلات، بهره‌گیری از بازیابی اطلاعات بین‌زبانی[1] است (در این پژوهش از واژهٔ **بازبین** برای بیان اختصاری «بازیابی اطلاعات بین زبانی» استفاده شده است). بازیابی اطلاعات بین زبانی نوعی از بازیابی اطلاعات است که در آن حداقل دو زبان وجود دارد: زبان عبارت جستجو، و زبان مجموعهٔ مدارک. زبان عبارت جستجو را زبان اصلی[2]، و زبان مجموعهٔ مدارک را زبان هدف یا مقصد[3] می‌نامند. یک نظام بازیابی اطلاعات بین‌زبانی (بازبین)، مدرک را در زبانی که با زبان عبارت جستجو متفاوت است

[1] Cross Language Information Rretrieval (clir)
[2] Source Language
[3] Target Language

بازیابی می‌کند. البته کاربر نظام بازبین، عبارت جستجو را به زبان بومی خویش ارائه می‌کند، اما مدارک دریافتی بر اساس زبان مجموعهٔ مدارک 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 بازیابی‌شده تمایز قائل شوند. مدارکی که مربوط تشخیص داده شده‌اند، سپس با استفاده از عامل انسانی یا ماشینی، ترجمه و استفاده می‌شوند (Ballesteros and Croft 1998).

۲. بیان مسئله

در حال حاضر، نظام بازیابی اطلاعات بین‌زبانی فارسی- انگلیسی وجود ندارد که کاربران گوناگون از آن در جهت برآورده ساختن نیاز خود در هنگام جستجو در محیط‌های جدید استفاده کنند. ازآنجاکه تاکنون تحقیقی نیز در زمینهٔ کاربرد زبان فارسی در این حوزه به عمل نیامده، تصویر روشنی از ظرفیت‌ها و امکانات بالقوهٔ این زبان در فرایند بازبین وجود ندارد. پس از گذشت دو دهه از به‌کارگیری نظام‌های بازبین در دنیا، هنوز در حوزهٔ زبان فارسی حرکتی برای شناخت، طراحی و بهره‌گیری از مزایای این نظام انجام نشده و به همین علت پیش‌زمینه‌ها و ملزومات آن نیز شناخته نشده است. امروزه اگر از فعالیت‌های انبوه انجام‌شده در زبان‌هایی چون انگلیسی و فرانسه بگذریم، حتی از زبان‌هایی چون عربی، تایلندی و اندونزیایی هم در این حوزه، سال‌ها عقب هستیم.

بر اساس آنچه که در ارتباط با اهمیت چنین نظامی مطرح شد، این تحقیق می‌کوشد راه‌حل‌هایی برای مسئله‌های زیر بیابد:

مشخص نیست که استفاده از فنون پردازش زبان طبیعی از قبیل تحلیل مورفولوژیک، برچسب‌زنی انواع نقش‌های دستوری، و دگرنویسی به چه میزان در ارتقاء کارآمدی نظام بازبین تأثیر دارد. همچنین مشخص نیست که در هنگام ترجمهٔ عبارت‌های جستجو، اتخاذ چه رویکردی در انتخاب برابرنهاده‌های ارائه‌شده در واژه‌نامه، منجر به دستیابی به نتیجهٔ مناسب‌تری برای عملکرد نظام بازبین می‌گردد. بدیهی است که بسیاری از نام‌های خاص، در واژه‌نامه موجود نیست؛ نحوهٔ رفتار با کلمات ترجمه‌ناپذیر و خارج از واژه‌نامه نیز مسئله‌ای است که در مقابل یک نظام بازبین قرار دارد. انسان در ارتباط‌های کلامی از

الگوی عبارتی برای انتقال بهتر مفاهیم بهره می‌گیرد؛ اما مشخص نیست که در ترجمهٔ عبارت‌های جستجو، شناسایی عبارت‌ها و ترجمهٔ آن‌ها به‌صورت عبارتی، تا چه میزان بر کارآمدی نظام بازیابی اطلاعات بین زبانی فارسی- انگلیسی تأثیر دارد. در نهایت، فراهم‌آوردن داده‌های تجربی برای بررسی این نادانسته‌ها می‌تواند زوایای ناشناختهٔ بازیابی اطلاعات بین‌زبانی فارسی- انگلیسی را روشن نماید.

هدف کلی این پژوهش، توسعهٔ نظری و عملی دانش بازبین فارسی با تمرکز خاص بر ابزار ترجمهٔ عبارت‌های جستجو است. اهداف دیگر این پژوهش شامل موارد زیر است:

- ◇ انتخاب مناسب‌ترین شیوه از بین رویکردهای مختلف در ترجمهٔ عبارت جستجو با واژه‌نامهٔ دوزبانهٔ ماشین‌خوان. وجود رویکردهای مختلف، تصمیم‌گیری برای انتخاب مناسب‌ترین شیوه را منوط به انجام پژوهشی نظام‌مند می‌سازد. در این پژوهش سعی می‌شود با بررسی رویکردهای مختلف به ترجمهٔ عبارت‌های جستجو در یک محیط اطلاعاتی و با استفاده از مجموعهٔ عبارت‌های جستجوی آزمون‌شده، مناسب‌ترین شیوهٔ ترجمه انتخاب، و با استفاده از معیارهایی چون قضاوت ربط، کارآمدی آن ارزیابی گردد.
- ◇ یک نظام بازیابی اطلاعات بین زبانی دارای اجزایی است که یکی از مهم‌ترین آن‌ها، قسمتی است که برای ترجمهٔ عبارت‌های جستجو استفاده می‌شود. این تحقیق در پی بررسی لزوم بهره‌گیری از ابزار پردازش زبان طبیعی (از قبیل ابزار حذف واژه‌های سیاههٔ بازدارنده، و تحلیل مورفولوژیک) در کنار نظام ترجمه است.
- ◇ اگرچه این تحقیق به زبان‌های فارسی و انگلیسی محدود شده، اما نتایج آن و رویکردهای به‌کارگرفته‌شده می‌تواند نقشهٔ راهی برای انجام پژوهش‌هایی مشابه در دیگر زبان‌ها ایجاد کند و در ایجاد نظام‌های بازبین فارسی به دیگر زبان‌ها نیز راهگشا باشد.

فرضیه‌های این پژوهش به شرح زیر هستند:

۱. در هنگام استفاده از واژه‌نامهٔ دوزبانهٔ ماشین‌خوان، رویکرد ترجمهٔ اولین برابرنهاده در مقایسه با رویکرد همهٔ برابرنهاده‌ها، باعث کارآمدی بیش‌تر (افزایش ضریب دقت) در نتایج بازیابی اطلاعات بین‌زبانی فارسی- انگلیسی می‌گردد.

۲. پردازش مورفولوژیک واژه‌های عبارت جستجوی فارسی (که در واژه‌نامهٔ مورد بررسی وجود ندارد) پیش از ترجمهٔ آن‌ها باعث کارآمدی بیش‌تر (افزایش ضریب دقت) در نتایج بازیابی اطلاعات بین زبانی فارسی- انگلیسی می‌گردد.
۳. در هنگام ترجمهٔ عبارت‌های جستجوی فارسی، شیوهٔ ترجمهٔ عبارتی در مقایسه با ترجمهٔ کلمه به کلمه باعث کارآمدی بیش‌تر (افزایش ضریب دقت) در نتایج بازیابی اطلاعات بین زبانی فارسی- انگلیسی می‌گردد.
۴. استفاده از روش دگرنویسی اصطلاح‌های ترجمه‌ناپذیر یا خارج از واژه‌نامهٔ فارسی به حروف انگلیسی و بازیابی بر اساس آن (در مقایسه با حذف این‌گونه واژه‌ها از عبارت جستجو) باعث کارآمدی بیش‌تر در بازیابی اطلاعات بین زبانی فارسی- انگلیسی می‌گردد.

۳. پیشینهٔ پژوهش

همان‌گونه که پیش از این ذکر شد، یکی از انگیزه‌های اصلی پژوهش در حوزهٔ بازیابی اطلاعات بین زبانی فارسی- انگلیسی، کمبود منابع فارسی در اینترنت در مقایسه با دیگر زبان‌های بزرگ است. کاستی‌های نرم‌افزاری را می‌توان یکی از بزرگ‌ترین دلایل غیبت زبان فارسی در اینترنت دانست. راه غلبه بر چنین مشکلی، توجه به طراحی و راه‌اندازی نرم‌افزارها و نظام‌های اطلاعاتی خاص زبان فارسی است. اکنون که توجه زیادی به تولید محتوای فارسی در وب می‌شود، سرمایه‌گذاری در پشتیبانی فنی این محتوا و ایجاد نظام‌هایی همچون نظام بازبین نیز ضروری به نظر می‌رسد. با جستجو در نوشته‌های داخلی، تنها مقاله‌ای که در آن به بازیابی اطلاعات بین زبانی فارسی اشاره شده، نوشتهٔ «آزادنیا» است که در طرح پیشنهادی خود، برای سند توسعهٔ بومی‌سازی در ایران و به عنوان یکی از محورهای اصلی این سند، به بازیابی چندزبانی اشاره نموده و اینترنت را محملی دانسته که برای بهره‌گیری بهینه از آن، چاره‌ای جز توسعهٔ بازیابی بین زبانی نیست (آزادنیا ۱۳۸۳). تا آنجا که پژوهشگر بررسی نموده، این گفته در حد پیشنهاد باقی مانده و هنوز اقدامی در عملی‌سازی این طرح انجام نشده است.

پژوهش‌های غیرفارسی در حوزهٔ بازبین بسیار زیاد است و در اینجا برای نمونه به مواردی اشاره می‌شود. اولین تجربهٔ بازبین با رویکرد واژه‌نامه نشان داد که ترجمهٔ واژه به

واژهٔ عبارت‌های جستجو می‌تواند منجر به کاهش بین ۴۰٪ تا ۶۰٪ کارآمدی بازیابی در مقایسه با بازیابی اطلاعات یک‌زبانه بر اساس همان عبارت‌های جستجو گردد (Hull and Grefenstette 1996). در این پژوهش که دربارهٔ بازبین فرانسوی- انگلیسی انجام شد، با استفاده از محاسبهٔ متوسط دقت بازیافت، نتایج زیر به دست آمد: ترجمهٔ عبارت‌های جستجو به روش واژه به واژه منجر به دستیابی به متوسط دقت بازیافت برابر با ۰/۲۳۵ می‌گردد که این مقدار، معادل ۶۰٪ میزان کارآمدی در بازیابی یک‌زبانه در این پژوهش بود. این میزان برای روش ترجمهٔ عبارتی عبارت‌های جستجو افزایش پیدا کرد و به ۰/۳۵۷ (یعنی ۹۱٪ عملکرد بازیابی یک‌زبانه که برابر با ۰/۳۹۳ بود) افزایش یافت. آن‌ها نتیجه گرفتند که ترجمهٔ عبارتی در مقایسه با ترجمهٔ واژه به واژهٔ عبارت‌های جستجو، منجر به نتیجهٔ بهتری می‌شود.

«چن» نیز با انجام پژوهشی به بررسی کارآمدی ترجمهٔ عبارتی در بازبین چینی- انگلیسی پرداخت. وی که با استفاده از روش تحقیق ارزیابی برنامه، این پژوهش را انجام داد، ترجمهٔ عبارتی را موفق‌تر از ترجمهٔ واژه به واژه دانست. یافته‌های او نشان داد که در مقایسه با میانگین دقت بازیافت در بازیابی یک‌زبانه، ترجمهٔ عبارتی به ۵۳٪ کارآمدی دست می‌یابد. این در حالی است که این میزان برای ترجمهٔ واژه به واژه به ۴۲٪ رسید. وی اشاره می‌کند که با بهره‌گیری از منابع اضافی و کامل‌تر برای ترجمهٔ عبارت‌ها، می‌توان این میزان از کارآمدی را افزایش داد و به ۸۳٪ کارآمدی به‌دست‌آمده برای بازیابی یک‌زبانه رسید (Chen 2002).

تلاش برای رفع ابهام در ترجمه نیز از جمله مواردی است که در پژوهش‌ها به آن توجه شده است. «بالستروس» و «کرافت» (Ballesteros & Croft 1998) برای رفع ابهام در ترجمه، روش آمار هم‌رخدادی[1] را مطرح کرده‌اند. رویکرد هم‌رخدادی بر این اندیشه استوار است که ترجمه‌های صحیح واژه‌های عبارت جستجو، گرایش به رخداد همزمان در متن‌های زبان هدف دارند، ولی ترجمه‌های نادرست این گرایش را ندارند. آن‌ها در پژوهش خود از یک ابزار برچسب‌زنی نوع نقش دستوری برای انتخاب برابرنهاده‌های زبان هدف، که دارای برچسب‌های نوع نقش دستوری مشابه با اصطلاح‌های عبارت جستجو در

[1] co-occurrence

زبان اصلی باشد، استفاده کردند. در نهایت با انجام محاسبه‌های مختلف، آمار هم‌رخدادی مشخص، و برابرنهاده‌هایی که در سیاههٔ رتبه‌بندی بر اساس آن، بالاتر قرار گرفته بودند، به عنوان ترجمهٔ مناسب انتخاب شدند (Ballesteros and Croft 1998).

٤. روش پژوهش

این پژوهش کاربردی است و برای انجام آن از روش پژوهش نیمه‌تجربی استفاده شده. جامعهٔ آماری این پژوهش، رکوردهای بازیابی‌شده از موتور جستجوی گوگل بر اساس جستجوی عبارت‌های جستجوی این پژوهش است. روش نمونه‌گیری این پژوهش، روش نمونه‌گیری «کنفرانس بازیابی متن (ترک)[1] برای ارزیابی نظام‌های بازیابی اطلاعات» است که امروزه در اغلب پژوهش‌های مربوط به ارزیابی نظام‌های بازیابی اطلاعات از آن استفاده می‌شود (Voorhees 1998). این شیوه، مخزن‌سازی[2] نام دارد. شیوهٔ مخزن‌سازی که در این پژوهش انجام شد به این شرح است: ابتدا عبارت‌های جستجو در اختیار تعدادی جستجوگر قرار گرفت. هر جستجوگر پس از انجام جستجو در سامانهٔ بازیابی اطلاعات (که در این تحقیق، موتور جستجوی گوگل بود) سیاههٔ نتایج بازیابی شامل ۱۰۰ رکورد اول بازیابی‌شده برای هر عبارت جستجو را ارائه نمود. سپس با استفاده از این سیاهه‌ها برای هر عبارت جستجو، مخزنی تشکیل شد. رکوردهای خارج از مخزن نیز نامربوط تلقی شدند (و به دلیل بزرگی مجموعه‌های مدارک، چاره‌ای جز این نیست).

با توجه به حجم بالای پردازش‌های زبان طبیعی و مراحل مختلف بازیابی، تعداد ۴۰ عبارت جستجو از مجموعه عبارت‌های جستجوی «ترِک» انتخاب شد. هر عبارت جستجو از سه بخش تشکیل شده: بخش اول شامل یک عنوان کوتاه است. بخش دوم توصیف یک‌جمله‌ای از موضوع (جملهٔ نیاز اطلاعاتی) است. بخش سوم از چند جمله که معیارهای ربط را مشخص می‌کنند، تشکیل شده. در زیر متن انگلیسی یکی از عبارت‌های جستجوی «ترِک» مشاهده می‌شود:

Title: War and Radio.
Description: What role do radios play during war or armed conflict?
Narrative: The radio plays an important role by broadcasting propaganda and messages of hope or warning. Relevant documents talk about this role.

---

[1] Text Retrieval Conference (TREC)
[2] Pooling

در ساخت عبارت جستجوی نهایی برای ارائه به سامانهٔ اطلاعاتی، از فیلد «عنوان» استفاده شد و از اطلاعات موجود در فیلد «توصیف» نیز در صورت نیاز برای بسط جستجو استفاده گردید.

۵. یافته‌های پژوهش

۱. در هنگام استفاده از واژه‌نامهٔ دوزبانهٔ ماشین‌خوان، رویکرد «ترجمهٔ اولین برابرنهاده» در مقایسه با رویکرد «همهٔ برابرنهاده‌ها» باعث کارآمدی بیش‌تر (افزایش ضریب دقت) در نتایج بازیابی اطلاعات بین‌زبانی فارسی- انگلیسی می‌گردد.

ابهام در ترجمه هنگامی رخ می‌دهد که برابرنهاده‌های انتخاب‌شده، همخوانی مناسبی با واژهٔ موجود در عبارت جستجو نداشته باشند. موضوع ابهام‌زدایی در ترجمه، از سوی پژوهشگرانی چون «دیویس» و «دانینگ»، و «پیرکولا» و همکارانش مطرح، و راهکارهایی در مورد آن ارائه شده است (Davis and Dunning 1995; Pirkola et al 2001). این روش به این دلیل که اصطلاح‌های بعضاً نامربوط را به عبارت جستجوی ترجمه‌شده به زبان هدف اضافه می‌کند، می‌تواند منجر به بروز ابهام در ترجمه گردد که این، خود منجر به کارآیی ضعیف در بازیابی می‌گردد. «مهراد» و «ناصری» معتقدند که زبان‌های طبیعی، خواص ویژه‌ای دارند که از سودمندی نظام‌های بازیابی اطلاعات متنی می‌کاهد. این خواص عبارت‌اند از اختلاف زبانی، و ابهام. منظور از اختلاف زبانی، استفاده از واژه‌ها یا عبارت‌های مختلف برای انتقال یک ایدهٔ واحد می‌باشد. ابهام زبانی نیز زمانی رخ می‌دهد که واژه یا عبارت دارای بیش از یک تفسیر است (مهراد و ناصری ۱۳۸۷).

در پژوهش حاضر برای بررسی شیوهٔ غلبه بر این مشکل، انتخاب اولین برابرنهاده مطرح گردید و این شیوهٔ ترجمه با شیوهٔ دیگر (یعنی استفاده از همهٔ برابرنهاده‌ها) مقایسه شد. ازآنجاکه در هنگام استفاده از رویکرد «همهٔ برابرنهاده‌ها» کلماتی به عبارت جستجوی نهایی وارد می‌شوند که ممکن است در معنای مورد نظر نویسندهٔ مدرک کاربرد نداشته باشند (که این خود منجر به بازیابی ناخواستهٔ تعدادی مدرک نامرتبط می‌گردد) و این امر در مقایسه با رویکرد «اولین برابرنهاده» (که معمولاً رایج‌ترین معادل هر کلمه را برای ترجمه انتخاب می‌کند) می‌تواند باعث ناکارآمدی در بازیابی گردد، این فرضیه مطرح شده است.

در این پژوهش، کارآمدی عملکرد بازیابی اطلاعات بین زبانی، با اندازه‌گیری میانگین متوسط دقت بازیافت[1] مشخص شده است. برای محاسبهٔ این آماره، ابتدا باید متوسط دقت بازیافت برای همهٔ مخزن‌های مدارک، محاسبه شود. نحوهٔ محاسبهٔ این مقیاس به این صورت است که بررسی ربط مدارک موجود در یک مخزن، از بالای سیاهه شروع می‌شود و بلافاصله که یک مدرک مربوط شناخته شد، مقدار دقت بازیافت محاسبه می‌شود. وقتی به پایان سیاههٔ مدارک موجود در مخزن رسیدیم، همهٔ مقدارهای دقت بازیافت را جمع می‌زنیم و تقسیم بر تعداد کل مدرک‌های مرتبط موجود در مخزن می‌کنیم. بدین ترتیب متوسط دقت بازیافت، برای آن عبارت جستجو محاسبه می‌شود. در صورت وجود مجموعه‌ای از عبارت‌های جستجو، میانگین متوسط دقت بازیافت برای کل عبارت‌های جستجوی موجود، از طریق جمع‌کردن متوسط دقت بازیافت همهٔ عبارت‌های جستجو و تقسیم آن بر تعداد کل عبارت‌های جستجو، محاسبه می‌شود (Wang and Oard 2005).

جدول ۱. مقایسهٔ میانگین متوسط دقت بازیافت رویکردهای انتخاب اولین برابرنهاده، و همهٔ برابرنهاده‌ها

روش ترجمه	میانگین متوسط دقت بازیافت	درصد تغییر نسبت به رویکرد اولین برابرنهاده
استفاده از اولین برابرنهاده	۰/۲۲۳	-
استفاده از همهٔ برابرنهاده‌ها	۰/۱۶۸	٪۷۵

همان‌گونه که در جدول ۱ مشاهده می‌شود، از میان دو روش ترجمهٔ مختلف، استفاده از اولین برابرنهاده با دستیابی به عدد ۰/۲۲۳، میانگین متوسط دقت بازیافت بیش‌تری داشته است. رویکرد انتخاب همهٔ برابرنهاده‌ها با دستیابی به ٪۷۵ کارآمدی رویکرد اولین برابرنهاده، باعث کاهش میزان کارآمدی شده است. مقایسهٔ رویکردهای مختلف ترجمه، کارآمدی بیش‌تر رویکرد انتخاب اولین برابرنهاده در فرایند ترجمهٔ عبارت‌های جستجوی نظام بازبین را نشان می‌دهد. نتایج فوق با یافته‌های «الجلایل» و

[1] mean average precision

«فرایدر» که به بررسی رویکردهای مختلف ترجمه در بازبین عربی- انگلیسی پرداخته بودند مشابهت دارد. آنان نیز به این نتیجه رسیده بودند که انتخاب اولین برابرنهاده در مقایسه با انتخاب همهٔ برابرنهاده‌ها، موجب کارآمدی بیش‌تر نظام بازبین می‌گردد (Aljalayil and Frieder 2001).

مقیاس دیگری که در این پژوهش برای ارزیابی کارآمدی نظام بازبین استفاده شد، «میانگین دقت بازیافت در سطوح مختلف بازیابی»[1] بود. براساس این مقیاس، دقت بازیافت در سطوح مختلف بازیابی یک سیاههٔ رتبه‌بندی‌شده مشخص می‌شود. در این پژوهش، دقت بازیافت در سطوح مختلف بازیابی (یعنی در سطح ۵، ۱۰، ۲۰، و ۳۰ مدرک بازیابی‌شده) محاسبه شد. این مقیاس ابتدا برای هر عبارت جستجو، جداگانه محاسبه، و آنگاه با محاسبهٔ میانگین آن‌ها در کل مجموعهٔ عبارت‌های جستجو، شمای کلی متوسط عملکرد نظام مشخص گردید. محاسبهٔ دقت بازیافت در سطوح مختلف بازیابی می‌تواند تصویر کامل‌تری از نحوهٔ عملکرد یک نظام بازیابی اطلاعات را مشخص کند. این مقیاس، موقعیت‌هایی را مدل‌سازی می‌کند که در آن، کاربران ترجیح می‌دهند تنها مدارکی را بررسی کنند که در سیاهه‌های نتایج بازیابی، رتبهٔ بالاتری دارند. این امر بویژه در جستجوهای وبی که در آن، موتورهای جستجو مدارک بسیاری را بازیابی می‌کنند و کاربران، اغلب صفحه‌های اولیهٔ این نتایج را مرور می‌کنند، نمود پیدا می‌کند.

جدول ۲. محاسبهٔ میانگین دقت بازیافت در سطوح مختلف بازیابی بر اساس انتخاب برابرنهاده‌های مختلف

دقت بازیافت	اولین برابرنهاده	همهٔ برابرنهاده‌ها
در ۵ مدرک	۰/۴۸۵	۰/۳۸۴
در ۱۰ مدرک	۰/۴۴۱	۰/۳۶۱
در ۲۰ مدرک	۰/۳۹۹	۰/۳۱۶
در ۳۰ مدرک	۰/۳۱۸	۰/۲۸۷

[1] mean precision at various cut- off levels

اطلاعات موجود در جدول ۲ نیز مؤید این نکته است که رویکرد انتخاب اولین برابرنهاده در سطوح مختلف بازیابی، برتری محسوسی نسبت به رویکرد انتخاب همهٔ برابرنهاده‌ها دارد و باعث کارآمدی بیش‌تر بازیابی اطلاعات بین‌زبانی می‌گردد. میزان دقت بازیافت چنانکه مشاهده می‌شود روندی نزولی دارد و با حرکت به سمت انتهای سیاههٔ مدارک و افزایش مدارکی که مورد قضاوت ربط قرار می‌گیرد، کاهش می‌یابد.

۲. پردازش مورفولوژیک واژه‌های عبارت جستجوی فارسی (که در واژه‌نامهٔ مورد بررسی وجود ندارد) پیش از ترجمهٔ آن‌ها در مقایسه با عدم انجام این پردازش، باعث کارآمدی بیش‌تر (افزایش ضریب دقت) در نتایج بازیابی اطلاعات بین زبانی فارسی- انگلیسی می‌گردد.

در این قسمت به بررسی میزان اهمیت به‌کارگیری تحلیل مورفولوژیک اصطلاح‌های مورد ترجمه در نظام بازیابی اطلاعات بین زبانی پرداخته شده است. پژوهشگران در مورد نحوه و سطح انجام چنین تحلیل‌هایی و میزان سودمندی آن‌ها در ارتقاء عملکرد بازیابی اطلاعات، نظرات مختلفی داشته‌اند. یکی از مهم‌ترین کارها در این حوزه، توسط «پرتر» ارائه شده است (Porter 1980).

در تشریح تحلیل مورفولوژیک (ساختواژه) باید گفت که بسیاری از اصطلاح‌های موجود در واژه‌نامه‌ها شکل مفرد کلمات، بدون پیشوند و پسوند و ادات جمع است. ازآنجاکه بسیاری از شکل‌های مختلف صرفی کلمات در واژه‌نامه وجود ندارند، در هنگام ترجمهٔ عبارت‌های جستجو، با انبوهی از کلمات ترجمه‌نشده روبرو می‌شویم. تحلیل مورفولوژیک در حل این مشکل، ضروری به نظر می‌رسد. با این تحلیل، اشکال مختلف کلمات پردازش می‌شوند، پیشوند و پسوندها حذف، و کلمات جمع به مفرد تبدیل می‌شوند و سپس ترجمه انجام می‌گیرد. طرح این فرضیه، برای آزمون سودمند بودن و ضرورت انجام تحلیل مورفولوژیک در هنگام ترجمه صورت گرفته است.

در این پژوهش تحلیل مورفولوژیک بر روی واژه‌هایی انجام شد که به شکل اولیهٔ موجود در عبارت جستجو، در واژه‌نامه یافت نشد. با مقایسهٔ بازیابی بر اساس به‌کارگیری این تحلیل و بازیابی عبارت‌های جستجویی که واژه‌های ترجمه‌نشده از آن حذف شده است، میزان کارآمدی این روش مشخص گردید.

جدول ۳. مقایسهٔ بازبین با استفاده از پردازش مورفولوژیک واژه‌های ترجمه‌نشده و عدم انجام این پردازش

نوع اجرا	میانگین متوسط دقت بازیافت	درصد تغییر
بازبین بدون تحلیل مورفولوژیک	۰/۱۸۰	-
بازبین با تحلیل مورفولوژیک	۰/۲۲۳	٪۲۳

در جدول ۳ مشاهده می‌شود که میانگین متوسط دقت بازیافت در شیوهٔ اول که بدون پردازش مورفولوژیک بوده برابر با ۰/۱۸۰ است، در حالی که پردازش مورفولوژیک واژه‌هایی که ترجمه نشده و در مرحله قبلی حذف شدند (و آنگاه ترجمهٔ آن‌ها با واژه‌نامه) منجر به افزایش این رقم به میزان ٪۲۳ گردید.

همین مسئله این بار در سطوح مختلف بازیابی آزمایش شد که نتایج به‌دست‌آمده، در جدول ۴ به نمایش گذاشته شده است.

جدول ٤. محاسبهٔ میانگین دقت بازیافت در سطوح مختلف بازیابی بر اساس مقایسهٔ انجام پردازش مورفولوژیک واژه‌های ترجمه‌نشده و عدم انجام این پردازش

دقت بازیافت	بدون پردازش مورفولوژیک	با پردازش مورفولوژیک
در ۵ مدرک	۰/۳۹۶	۰/۴۸۵
در ۱۰ مدرک	۰/۳۷۲	۰/۴۴۱
در ۲۰ مدرک	۰/۳۱۰	۰/۳۹۹
در ۳۰ مدرک	۰/۲۶۳	۰/۳۱۸

نتایج مندرج در جدول ۴ نشان می‌دهد که انجام پردازش مورفولوژیک در سطوح مختلف بازیابی باعث افزایش میانگین دقت بازیافت و در نتیجه کارآمدی بیش‌تر این روش می‌گردد.

۳. در هنگام ترجمهٔ عبارت جستجوهای فارسی، شیوهٔ ترجمهٔ عبارتی در مقایسه با ترجمهٔ واژه به واژه، باعث کارآمدی بیش‌تر (افزایش ضریب دقت در بازیافت) در نتایج بازیابی اطلاعات بین زبانی فارسی- انگلیسی می‌گردد.

وجود «عبارت» در بعضی از جملات جستجو این سؤال را در ذهن پدید می‌آورد که یک عبارت را باید یک ترکیب واژگانی در نظر گرفت و کلمات آن را با هم ترجمه کرد، یا این‌که باید واژه‌های عبارت به‌صورت جدا از هم ترجمه گردد. پژوهشگران رویکردهای مختلفی را در مواجهه با این مورد به کار گرفته‌اند. برای مثال «بالستروس» و «کرافت» در پژوهشی به بررسی تأثیر رویکردهای «آماری» و «پردازش زبان طبیعی» در ترجمهٔ عبارتی پرداخته‌اند (Ballesteros and Croft 1998). این فرضیه به دلیل استفادهٔ زیاد از عبارت‌ها در ساخت جملات جستجوی فارسی، و در جهت بررسی کارآمدی روش ترجمهٔ عبارتی در بازبین، مطرح گردید.

شناسایی عبارت‌های موجود در جملات جستجو و ترجمهٔ آن‌ها به‌صورت عبارت (و نه به‌صورت واژه‌های مجزا) از چالش‌های پیش روی بازبین است. در این پژوهش با استفاده از برچسب‌زنی انواع نقش دستوری (یعنی مشخص‌کردن نوع دستوری واژه‌ها از قبیل اسم، صفت، قید، و ...) تلاش شد عبارت‌های موجود، مشخص و در صورت موجودبودن در واژه‌نامه، به‌صورت عبارتی ترجمه گردند. این شیوهٔ ترجمه سپس با شیوهٔ ترجمهٔ واژه به واژه مقایسه گردید.

جدول ۵. مقایسهٔ کارآمدی شیوهٔ ترجمهٔ عبارتی با شیوهٔ ترجمهٔ واژه به واژه

نوع اجرا	میانگین متوسط دقت بازیافت	درصد تغییر
بازبین با استفاده از ترجمهٔ واژه به واژه	۰/۲۲۳	-
بازبین با استفاده از ترجمهٔ عبارتی	۰/۳۱۹	٪۴۳

همان‌گونه که در جدول ۵ مشاهده می‌شود، شیوهٔ ترجمهٔ عبارتی با دستیابی به میانگین متوسط دقت بازیافتی برابر با ۰/۳۱۹ در مقایسه با ترجمهٔ واژه به واژه کارآمدتر بوده و باعث ارتقاء آن به میزان ۴۳٪ می‌گردد. این یافته‌ها با نتایج پژوهش‌های انجام‌شده

توسط «هال» و «گرفنستت» که در مورد بازبین فرانسوی- انگلیسی انجام شد و «چن» که به بررسی بازبین چینی- انگلیسی پرداخت همخوانی دارد (Hull and Grefenstette 1996; Chen 2002). در آن پژوهش‌ها نیز کارآمدی روش ترجمهٔ عبارتی در مقایسه با ترجمهٔ واژه به واژه مشخص گردیده بود.

کارآمدی ترجمهٔ عبارتی در مقایسه با ترجمهٔ واژه به واژه در سطوح مختلف بازیابی نیز بررسی شد که نتایج این بررسی در جدول ۶ نشان داده شده است.

جدول ٦. نتایج بررسی میانگین دقت بازیافت در سطوح مختلف بازیابی بر اساس ترجمهٔ عبارتی و ترجمهٔ واژه به واژه

دقت بازیافت	ترجمهٔ واژه به واژه	ترجمهٔ عبارتی
در ۵ مدرک	۰/۴۸۵	۰/۵۹۲
در ۱۰ مدرک	۰/۴۴۱	۰/۵۴۱
در ۲۰ مدرک	۰/۳۹۹	۰/۵۱۲
در ۳۰ مدرک	۰/۳۱۸	۰/۴۸۱

نتایج اطلاعات موجود در جدول ۶ نشانگر برتری نسبی ترجمهٔ عبارتی در مقایسه با ترجمهٔ واژه به واژه در سطوح مختلف بازیابی است. این یافته‌ها در تکمیل نتایج پیشین، مبین سودمندی بهره‌گیری از برچسب‌زنی انواع نقش‌های دستوری و شناسایی و ترجمهٔ عبارت‌ها در فرایند بازیابی اطلاعات بین‌زبانی است.

٤. استفاده از روش دگرنویسی اصطلاح‌های ترجمه‌ناپذیر یا خارج از واژه‌نامهٔ فارسی با حروف انگلیسی و بازیابی بر اساس آن، در مقایسه با حذف آن‌ها از عبارت‌های جستجو، باعث کارآمدی بیش‌تر بازیابی اطلاعات بین زبانی فارسی- انگلیسی می‌گردد.

واژه‌هایی که ترجمه‌ناپذیر و خارج از واژه‌نامه هستند (مثل نام‌های خاص) اغلب از مهم‌ترین اجزای یک عبارت جستجو به‌شمار می‌روند. «پیرکولا» مشکل ترجمه‌ناپذیری آن‌ها را از بزرگ‌ترین چالش‌های پیش روی نظام بازبین می‌داند (Pirkola et al 2001). در این پژوهش برای حل مشکل عدم‌ترجمهٔ این گونه واژه‌ها، شیوهٔ دگرنویسی (نوشتن

واژه‌های ترجمه‌ناپذیر به الفبای زبان دیگر) پیشنهاد شده است. این شیوه در ادامهٔ این قسمت با شیوهٔ حذف این واژها از عبارت جستجو، مقایسه شده است.

جدول ۷. مقایسهٔ میانگین متوسط دقت بازیافت بازبین با استفاده ازدگرنویسی، و بازبین بدون استفاده از دگرنویسی

نوع اجرا	میانگین متوسط دقت بازیافت	درصد تغییر
بازبین با استفاده از دگرنویسی	۰/۲۲۳	٪۷۴
بازبین بدون استفاده از دگرنویسی	۰/۱۲۸	–

همان‌گونه که در جدول فوق مشاهده می‌شود، استفاده از روش دگرنویسی واژه‌های ترجمه‌ناپذیر در مقایسه با حذف آن‌ها از عبارت‌های جستجو باعث ارتقاء عملکرد به میزان ٪۷۴ می‌شود. این یافته‌ها با نتایج پژوهش‌های «کو» و «گرفنستت»، و «چنگ» و همکاران همخوانی دارد (Qu and Grefenstette 2003; Cheng et al 2004). در پژوهش «چنگ» و همکاران نیز دگرنویسی واژه‌های ترجمه‌ناپذیر در مقایسه با حذف آن‌ها، باعث افزایش کارآمدی نظام بازبین به میزان ٪۸۰ گردیده بود. مقایسهٔ این دو روش در سطوح مختلف بازیابی نیز مؤید همین برتری است. جدول ۸ این نکته را نشان می‌دهد.

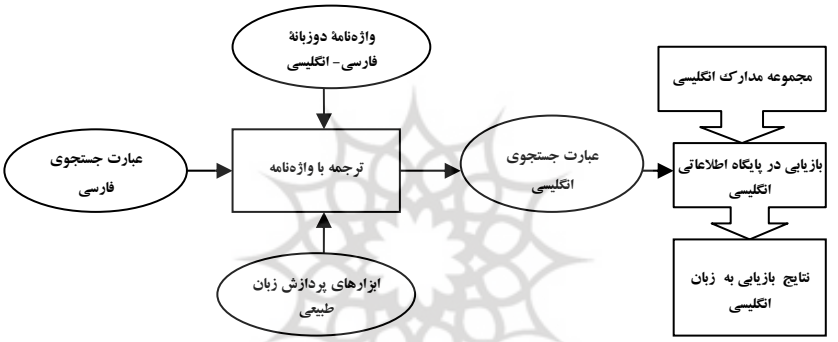
جدول ۸. مقایسهٔ میانگین دقت بازیافت بازبین با استفاده از دگرنویسی، و بدون استفاده از دگرنویسی در سطوح مختلف بازیابی

دقت بازیافت	بازبین با دگرنویسی	بازبین بدون دگرنویسی
در ۵ مدرک	۰/۴۸۵	۰/۲۶۵
در ۱۰ مدرک	۰/۴۴۱	۰/۲۲۱
در ۲۰ مدرک	۰/۳۹۹	۰/۲۰۲
در ۳۰ مدرک	۰/۳۱۸	۰/۱۸۵

مجموع یافته‌ها در این مرحله از بازیابی مؤید این است که روش دگرنویسی واژه‌های ترجمه‌ناپذیر و خارج از واژه‌نامه، در مقایسه با عدم استفاده از این شیوه و حذف این‌گونه واژه‌ها از عبارت‌های جستجو، باعث کارآمدی نظام بازیابی اطلاعات بین زبانی می‌گردد.

٦. مدل پیشنهادی نظام بازبین فارسی- انگلیسی

بر اساس مفاهیم نظری که پیش‌تر ذکر گردید و با تکیه بر دانش برگرفته از یافته‌های این پژوهش، مدل پیشنهادی نظام بازبین فارسی- انگلیسی به شرح زیر ارائه می‌گردد:



شکل ۱. مدل پیشنهادی نظام بازبین فارسی- انگلیسی

در تحلیل این مدل باید گفت که نظام بازبین فارسی- انگلیسی نظام خودکاری است که با استفاده از واژه‌نامهٔ الکترونیکی و هم‌افزایی ابزارهای پردازش زبان طبیعی، پس از پردازش و ترجمهٔ عبارت‌های جستجوی زبان اصلی (فارسی)، مدارک را از مجموعهٔ زبان هدف (انگلیسی) بازیابی می‌کند. مدل پیشنهادی، انعطاف‌پذیری مناسبی برای استفاده در جفت‌های زبانی دیگر نیز دارد و بر اساس آن می‌توان از زبان فارسی به مجموعه مدارک در زبان‌های دیگر دست یافت. این مدل از سویی کاربردی و از سوی دیگر، ارتباطی است. ارتباطی از آن جهت که اجزای کاربردی نظام در جهت عملکرد بهینه، باید ارتباط نزدیکی داشته باشند. نظام بازبین پیشنهادی، کارکردهای زیر را پشتیبانی می‌نماید:

◇ ترجمهٔ عبارت جستجوی فارسی

◇ تحلیل‌های پردازش زبان طبیعی
◇ بازیابی مدارک در زبان انگلیسی

ویژگی بارز این مدل، استفاده از ابزار پردازش زبان طبیعی است که موجب افزایش دقت در ترجمه و کارآمدترشدن بازبین می‌گردد. این مدل اولین مدلی است که برای فرایند بازیابی اطلاعات بین‌زبانی فارسی مطرح شده و می‌تواند با انجام پژوهش‌های دیگر تکمیل‌تر گردد.

۷. نتیجه‌گیری

دلایل بسیاری وجود دارد که یک نظام بازبین به همان کارآمدی که بازیابی یک‌زبانه نائل می‌شود، دست پیدا نمی‌کند. اولین و مهم‌ترین دلیل، وجود دو زبان در بازبین و ساختارهای متفاوت آن‌ها است. این دوگانگی موجب بسیاری ابهام‌ها می‌گردد که بازیابی اطلاعات یک‌زبانه هرگز با آن مواجه نمی‌شود. دلیل دیگر آن است که متن و زمینهٔ عبارت‌های جستجو در هنگام ترجمه، چندان در نظر گرفته نمی‌شوند؛ چه‌بسا در مواردی مهجورترین برابرنهادهٔ یک واژه، با توجه به مفهوم عبارت جستجو مناسب‌ترین انتخاب باشد. در این پژوهش با استفاده از فنون پردازش زبان طبیعی از قبیل تحلیل مورفولوژیک، برچسب‌زنی انواع نقش دستوری و دگرنویسی، مشخص گردید که استفاده از ابزار زبانشناسی در فرایندهای بازیابی اطلاعات می‌تواند منجر به کارآمدترشدن این نظام‌ها گردد. شناسایی الگوهای زبانی و ساختار واژه‌ها، یک نظام خودکار بازبین را قادر می‌سازد که مدارک را با دقت بیش‌تری بازیابی کند و رضایتمندی کاربر را افزایش دهد. فنونی که در این پژوهش استفاده شدند اغلب فنون نحوی بودند و علی‌رغم اهمیت، محدودیت‌هایی نیز دارند. بهره‌گیری از فنون پیشرفته‌تر هوش مصنوعی و استفاده از تحلیل‌های معناشناسی می‌تواند در آینده، سامانه‌های بازبین را نسبت به نمونه‌های موجود، سامانه‌های موفق‌تری نشان دهد.

۸. فهرست منابع

آزادنیا، محمد. ۱۳۸۳. طرح مقدماتی سند توسعهٔ بومی‌سازی در ایران. در *مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه*. تهران: طیف‌آرا.

علیزاده، حمید. ۱۳۸۳. مشکلات دسترسی به اطلاعات در جهان شبکه‌ها. *فصلنامه کتاب* ۱۵(۲): ۱۲۱-۱۱۵.

مهراد، جعفر و مریم ناصری. ۱۳۸۷. *پردازش زبان طبیعی و بازیابی اطلاعات*. تهران: چاپار، شیراز: مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری.

Adriani, M. 2000. Using statistical term similarity for sense disambiguation in cross- language information retrieval. *Information retrieval* 38(2): 69-80.

Aljlayl, Mohammad, and Phir Frieder. 2001. *Effective Arabic- English Cross- Language Information Retrieval via Machine- Readable Dictionaries and Machine Translation*. Oral presented at the ACM Tenth Conference on Information and Knowledge Management, Atlanta.

Ballesteros, L, and B. Croft. 1998. *Resolving Ambiguity for Cross- Language Retrieval*. Presented at the SIGIR, Australia, Melbourne.

Chen, H. H. 2002. *Chinese information extraction techniques*. Presented at the SSIMIP, Singapore.

Cheng, Pu-Jen, Pan, Yi-Cheng Lu, Lu Wen-Hsiang and Lee Feng Chien. 2004. Creating Multilingual Translation Lexicons with Regional variations Using Web Corpora. In *Procceedings Of ACL*. www.aclweb.org/anthology/P/P04/P04-1068.pdf. (accessed may 15, 2008).

Davis, M, and T. Dunning. 1995. *Query Translation Using Evolutionary Programming for Multilingual Information Retrieval*.Presented at the 4th Evolutionary Programming Conference.

Edwards, J. 1994. *Multilingualism*. London, England: Penguin.

Hull, D. and G. Grefenstette. 1996. Querying Across Languages; A Dictionary –Based Approach to Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM Sigir*, 49-57. Zurich, Switzerland.

Pirkola, A., H. Hedlund, T. Keskustaloh and K. Jarvelin. 2001. Dictionary – Based Cross-language Information Retrieval: Problems, Methods and Research Findings. *Information Retrieval*4C3/4: 209-230.

Porter, M. F.1980. An algorithm for suffix stripping. *Program*14: 130-137.

Qu, Yan, and Gregory Grefenstette. 2003. *Automatic transliteration for Japanese-to-English text retrieval*. Presented at SIGIR, Canada, Torento. http://dblp.uni-trier.de/db/conf/sigir/sigir2003.html#QuGE03 (accessed August 23, 2009).

Voorhees, Ellen M. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In C. J.Van rijs Bergen, W. Bruce Croft and Alistair Moffat, Proceedings of the 21st Annual International ACM Sigir Conference on Research and Development in Information Retrieval, 315-323. ACM Press.

Voorhees, Ellen M. 2003. Overview of TREC2002. http://trec.nist.gov/pubs/trec11/papers/OVERVIEW.11.pdf (accessed Jan. 21, 2008)

Wang, Jianqiang, and Douglas W. Oard. 2005. *Document and Query Expansion Using Side Collections and Thesauri.* Presented at the CLEF 2005, Vienna, Austria. http://www.informatik.uni-trier.de/~ley/db/conf/clef/clef2005.html#WangO05 (accessed August 23, 2009).