

جستارهای بازیابی اطلاعات بین-زبانی

مترجمان:

حمید علنیراده

استادیار علوم کتابداری و اطلاع رسانی

محمد رضا قانع

استادیار علوم کتابداری و اطلاع رسانی

سارا کلینی

کارشناس ارشد مهندسی کامپیوتر



عنوان و نام پدیدآور : جستارهای بازیابی اطلاعات بین-زبانی / مترجمان سارا کلینی،

محمدرضا قانع، حمید علیزاده

مشخصات نشر : تهران: چاپار، ۱۳۹۰.

مشخصات ظاهری : ۱۱۴ص: جدول، نمودار.

شابک : ISBN: 978-600-5294-83-5

وضعیت فهرست‌نویسی : فیپا

یادداشت : کتابنامه.

موضوع : بازیابی اطلاعات بین زبانی.

موضوع : زبان‌شناسی کامپیوتری.

شناسه افزوده : کلینی، سارا، مترجم.

شناسه افزوده : قانع، محمدرضا، ۱۳۳۸-، مترجم.

شناسه افزوده : علیزاده، حمید، مترجم.

رده‌بندی کنگره : ۱۳۹۰ ج ۵ / Z۶۶۷

رده‌بندی دبیری : ۰۲۵/۰۴

شماره کتابشناسی ملی : ۲۳۲۶۸۸۳



تهران: ۸۸۸۹۹۶۸۰ - ۹۱۲۱۹۶۰۲۱۴

جستارهای بازیابی اطلاعات بین - زبانی

مترجمان: سارا کلینی، محمدرضا قانع، حمید علیزاده

طراح جلد و صفحه‌آرا: صابر ۰۹۱۲۲۴۰۵۸۶۶

ناظر فنی چاپ: مریم تفضلی

چاپ و صحافی: چاپخانه آرین

چاپ اول: ۱۳۹۰

شمارگان: ۱۰۰۰ نسخه

شابک: ۹۷۸-۶۰۰-۵۲۹۴-۸۳-۵ ISBN: 978-600-5294-83-5

حق هرگونه چاپ و انتشار محفوظ است.

قیمت ۲۵۰۰ تومان

فهرست مندرجات

۹	مقدمه مترجمان
۱۳	۱ آشنایی با بازیابی اطلاعات بین - زبانی
۱۵	۱- مقدمه
۱۶	۱-۱ منابع
۱۷	۲-۱ ارزیابی
۱۸	۲ عبور از مانع زبان
۲۰	۱-۲ ترجمه پرس و جو
۲۰	۲-۲ ترجمه مفارک
۲۱	۳-۲ ترجمه عبارتی
۲۱	۴-۲ انتخاب ترجمه ها
۲۲	۱-۴-۲ کیفیت فهرست اصطلاحات دوزبانه
۲۳	۲-۴-۲ عدم توازن
۲۳	۳-۴-۲ پرس و جوهایی متوازن
۲۴	۴-۴-۲ پرس و جوهایی ساخت یافته
۲۵	۵-۴-۲ زبان میانجی
۲۶	۵-۲ ویرایش ترجمه ها
۲۶	۳ راهکارهای بازیابی بین - زبانی
۲۷	۱-۳ مدل های زبان برای CLIR
۲۳	۲-۳ راهکارهای مجموعه های دوزبانه
۳۴	۱-۲-۳ موازی سازی
۳۵	۲-۲-۳ پیکره قابل مقایسه
۳۵	۳-۳ راهکارهای پیکره های قابل مقایسه
۳۶	۱-۳-۳ استخراج فهرست اصطلاحات دوزبانه
۳۷	۲-۳-۳ مدل های پنهان مارکوف
۳۸	۳-۳-۳ مدل های ربط بین - زبانی
۳۹	۴-۳-۳ مدل فضای برداری تعمیم یافته
۴۰	۵-۳-۳ نمایه سازی معنایی پنهان

۴. ابزارهای سودمند (تسهیل کننده) بین - زبانی ۴۴
- ۴-۱. بازخورد ربط بین - زبانی ۴۴
- ۴-۲. ریشه سازی ۴۶
- ۴-۲-۱. ترجمه Backoff ۴۶
- ۴-۲-۲. ساخت خودکار ریشه ساز ۴۷
- ۴-۲-۳. ریشه سازهای عربی ۴۹
- ۴-۲-۴. زبان های آسیایی ۵۰
- ۴-۳. این - گرام ها ۵۱
- ۴-۴. دگر نویسی ۵۱
- ۴-۴-۱. رویکرد آوایی ۵۳
- ۴-۵. برچسب زدن موجودیت ها ۵۴
- ۴-۶. هم جوشی ۵۶
۵. خلاصه ۵۷
- منابع ۵۸
- ۴ پرسش و پاسخ تعاملی، آیا جستجوی بین - زبانی دشوارتر از جستجوی تک زبانه است؟ ۶۳
- چکیده ۶۵
۱. مقدمه ۶۵
۲. طرح آزمایش ۶۹
- ۲-۱. داده های آزمایشی ۶۹
- ۲-۲. پروفایل کاربر ۷۰
- ۲-۳. رابط های جستجو ۷۱
- ۲-۴. جلسات جستجو ۷۳
- ۲-۵. مقیاس های ارزیابی ۷۳
۳. نتایج بحث ۷۴
- ۳-۱. آیا جستجوی بین - زبانی از جستجوی تک زبانه مشکل تر است؟ ۷۵
- ۳-۱-۱. تفاوت ها در دقت ۷۵
- ۳-۱-۲. تفاوت ها در رفتار جستجو ۷۷
- ۳-۱-۳. پرسشنامه ها و مطالعات مشاهده ای ۷۸
- ۳-۲. آیا متن بیشتری جهت فراهم نمودن پاسخ های دقیق مورد نیاز است؟ ۸۰
- ۳-۳. تحلیل عدم موفقیت ۸۳
۴. آیا شرایط آزمون برای پرسش و پاسخ محاوره ای مناسب است؟ ۸۸

۹۱.....	۵. کارهای مربوط.....
۹۲.....	۶. نتایج.....
۹۴.....	منابع.....
۹۵.....	۳ بازیابی اطلاعات چندرسانه‌ای بین-زبانی.....
۹۷.....	چکیده.....
۹۷.....	۱. مقدمه.....
۹۸.....	۲. برنامه کاربردی بازیابی چندرسانه‌ای بین-زبانی.....
۱۰۱.....	۳. روش‌های ترجمه.....
۱۰۳.....	۴. ارزیابی.....
۱۰۹.....	منابع.....
۱۱۱.....	واژه‌نامه انگلیسی به فارسی.....
۱۱۳.....	واژه‌نامه فارسی به انگلیسی.....

مقدمه مترجمان

با گسترش روزافزون استفاده از اینترنت و غلبه بر محدودیت‌های دسترسی به اطلاعات که به مدد توسعه فناوری اطلاعات و ارتباطات حاصل شده است، کاربران دیگر تنها به منابع اطلاعاتی که به زبان آنها نوشته شده است اکتفا نمی‌کنند بلکه دسترسی به اطلاعات مرتبط در سایر زبان‌ها را حق خویش می‌دانند. امروزه وظیفه بازیابی اطلاعات به فرایندهای سختی آن محدود نمی‌شود، بلکه هدف‌های گسترده‌تر، یعنی غلبه بر موانع زبانی در هنگام جستجو و بازیابی اطلاعات نیز در این حوزه مطرح شده است.

تعداد زبان‌های زنده دنیا را چیزی حدود ۴۵۰۰ زبان تخمین می‌زنند، که از میان آنها در حدود ۳۰ زبان وجود دارد که هر کدام توسط حداقل ۳۰ میلیون نفر استفاده می‌شود. بدیهی است که برای تبادل اطلاعات در جامعه اطلاعاتی چندزبانه، دیگر مطلوب نیست که به اطلاعات یک زبان خاص محدود شد.

اینترنت به عنوان محل ظهور این زبان‌ها بیشترین نمود این گوناگونی را به خود اختصاص داده است. آمارها نشان می‌دهد که استفاده از اینترنت در چند سال اخیر رشد قابل ملاحظه‌ای داشته است. این نرخ رشد به ویژه در خاورمیانه، آمریکای جنوبی و آفریقا بسیار چشمگیر است. این تنوع جغرافیایی با تنوع زبانی نیز همراه است. بطوری که با رشد منابع اینترنتی، مشکلات و همچنین سودمندی دسترسی و بهره‌گیری از منابع به زبان‌های دیگر نیز بیشتر شده است. راه حل اساسی غلبه بر مشکلات زبانی در بهره‌گیری از منابع در محیط‌های چندزبانه، بازیابی اطلاعات بین-زبانی است.

بازیابی اطلاعات بین-زبانی نوعی از بازیابی اطلاعات است که در آن حداقل دو زبان حضور دارد، زبان عبارت جستجو و زبان مجموعه مدرک. زبان عبارت جستجو را زبان اصلی و زبان مجموعه مدرک را زبان هدف یا مقصد می‌نامند. یک نظام بازیابی اطلاعات بین-زبانی، مدارکی را که با زبان عبارت جستجو متفاوت است بازیابی می‌کند. در این شیوه، کاربر عبارت جستجو را به زبان بومی خویش ارائه می‌کند، اما مدارک دریافتی به زبان مجموعه مدرک خواهد بود. این نظام کار جستجوگرانی را که به چند زبان تسلط دارند ساده می‌کند و در عین حال جستجوگرانی را که تنها به یک زبان تسلط دارند، قاهر می‌سازد عبارت جستجو را به زبان خود ارائه کرده و آنگاه با استفاده از دانش خود یا با بهره‌گیری از دیگران، بین مدارک بازیابی‌شده تمایز قائل شوند. سپس مدارکی که مربوط تشخیص داده می‌شود، با استفاده از عامل انسانی یا ماشینی ترجمه شده و مورد استفاده قرار می‌گیرد.

در بازیابی اطلاعات بین-زبانی، هم مدرک و هم عبارت جستجو می‌تواند ترجمه شود. چون ترجمه عبارت جستجو در مقایسه با ترجمه مدرک ارزان‌تر و به زمان و کار علمی کم‌تری نیاز دارد، در پژوهش‌های انجام‌شده در این حوزه بیشتر به ترجمه عبارت جستجو توجه گردیده است.

در ترجمه عبارت جستجو سه رویکرد عمده وجود دارد: استفاده از ترجمه ماشینی، استفاده از واژه‌نامه‌های دوزبانه ماشین خوان و استفاده از پیکره‌های متن. هر کدام از این رویکردها، نقاط ضعف و قوت خود را دارد. شاید پژوهش‌ها نشان داده است که میزان پیشرفت‌هایی که در ترجمه ماشینی حاصل شده است در مقایسه با هزینه‌هایی که صرف تحلیل زبانشناسی در نظام می‌شود اندک بوده است. ماشین‌های ترجمه موجود توانایی چندانی در ترجمه دقیق مفاهیم مطرح‌شده در مدرک را ندارند. بنابراین اغلب تلاش‌ها بر ترجمه عبارت جستجو متمرکز شده است؛ زیرا عبارت جستجو در مقایسه با کل یک مدرک دارای کلمات کمتری است و در نتیجه ابهام کمتری در ترجمه آن مشاهده می‌شود. استفاده از رویکرد پیکره‌های متن نیز دارای نقطه ضعف، یعنی کمبود منابع قابل استفاده در زبان‌های مختلف است. مجموعه‌های متن همیشه به راحتی در دسترس نیستند و موارد اندک موجود هم نسبتاً کوچک بوده و یا فقط حجم اندکی از موضوعات را پوشش می‌دهد.

در رویکرد ترجمه عبارت جستجو با استفاده از واژه‌نامه، عبارت‌های جستجوی زبان اصلی را بوسیله یافتن برابرنهاده هر اصطلاح در واژه‌نامه دوزبانه الکترونیکی ترجمه می‌کند. نتیجه این کار یک عبارت جستجو در زبان مجموعه مدارک (زبان هدف) است که با در نظر گرفتن بعضی یا تمام برابرنهاده‌ها تولید می‌شود. بدیهی است پذیرش این نظام توسط کاربر تا حد بسیار زیادی به کارآمدی شیوه‌ای که عبارت‌های جستجو به آن شیوه ترجمه می‌شود بستگی دارد.

اثر پیش رو، اولین کتاب در حوزه بازیابی اطلاعات بین-زبانی به زبان فارسی است و برای آشنایی اولیه علاقمندان حوزه بازیابی اطلاعات با این موضوع ترجمه شده است. مترجمان امیدوارند در گام‌های بعدی زوایای دیگر این مبحث را برای علاقمندان تشریح نمایند.

www.ketabchi.ir

www.ketab.ir