
بسم الله الرحمن الرحیم

دانشگاه یزد

دانشکده مهندسی برق و کامپیوتر

گروه مهندسی کامپیوتر

پایان‌نامه

برای دریافت درجه کارشناسی ارشد

مهندسی کامپیوتر- هوش مصنوعی و رباتیک

ترجمه هوشمند اسامی فارسی در بازیابی اطلاعات بین زبانی

استاد راهنما : دکتر علی محمد زارع بیدکی

استاد مشاور : دکتر علیرضا یاری

پژوهش و نگارش : زهره حق‌اللهی

شهریور ۹۱

تعهدنامه اصالت اثر

اینجانب زهره حق‌اللهی تایید می‌کنم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب است و به دست آوردهای دیگران که در این نوشته از آنها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم سطح یا بالاتر ارائه نشده است.

نام و نام خانوادگی دانشجو: زهره حق‌اللهی

امضای دانشجو:

تقدیم به پدر، مادر و همسرم

تشکر و قدردانی

در ابتدا از استاد محترم آقای دکتر زارع بیدکی که استاد راهنمای اینجانب در انجام این پایان‌نامه بوده‌اند و کمک شایانی در مسیر انجام پروژه داشته‌اند، تشکر و قدردانی ویژه داشته باشم ونیز جناب آقای دکتر یاری که در کسوت استاد مشاور پشتیبان اینجانب بوده‌اند، کمال تشکر را دارم. همچنین از اعضای هیأت داوران که متقبل زحمت داوری پایان‌نامه شدند، نهایت امتنان و سپاسگزاری را دارم.

بخشی از حمایت مالی این پایان‌نامه توسط مرکز تحقیقات مخابرات ایران انجام شده است.

چکیده

با گسترش روزافزون استفاده از اینترنت، کاربران و جستجوگران اطلاعات، دیگر تنها به منابع اطلاعاتی که به زبان آنها نوشته شده اکتفا نمی‌کنند. بنابراین سیستم‌های بازیابی اطلاعات بین زبانی، این نیاز را برای کاربران برآورده می‌سازند. به این صورت که می‌توان تنها با دادن پرس‌وجو در یک زبان، اسناد با زبان‌های مختلف را بازیابی کرد. از بین پرس‌وجوهای کاربران، اسامی افراد از اهمیت خاصی برخوردار هستند. زیرا تعداد زیادی از پرس‌وجوها به اسامی افراد اختصاص می‌یابد. همچنین کاربران زیادی صفحات شخصی خود را به زبان‌های دیگر طراحی می‌کنند. بنابراین در صورت تشخیص این اسامی در پرس‌وجوها و تبدیل آنها به زبان‌های دیگر (که نویسه‌گردانی نام دارد)، می‌توان به صفحات شخصی آنها مستقل از زبانی که هستند، دست یافت.

در این پایان‌نامه، مروری کوتاه بر الگوریتم‌های تشخیص نام، سیستم‌های بازیابی اطلاعات بین زبانی و نیز به طور خاص نویسه‌گردانی کلمات انجام شده است. همچنین چالش‌های مطرح شده در این زمینه و روش عملکرد و نقاط قوت و ضعف آنها مورد بررسی قرار گرفته‌اند. در ادامه روشی برای تشخیص اسامی افراد به زبان فارسی با استفاده از موتور جستجو و همچنین روشی نوین به منظور نویسه‌گردانی این اسامی با استفاده از محتوای صفحات وب فارسی پیشنهاد شده است. مطابق آزمایشات انجام شده با استفاده از روش‌های پیشنهادی، می‌توان به دقت بالایی در زمینه تشخیص و ترجمه اسامی خاص و یا به عبارتی بازیابی اطلاعات بین زبانی اسامی افراد دست پیدا کرد.

واژه‌های کلیدی : موتور جستجو، بازیابی اطلاعات بین زبانی، نویسه‌گردانی، اسامی افراد

فهرست مطالب

عنوان | شماره صفحه

فهرست شکل‌ها

عنوان	شماره صفحه

فهرست جدول‌ها

عنوان	شماره صفحه

فصل ۱. مقدمه

۱-۱ مقدمه

بازیابی اطلاعات بین زبانی یکی از شاخه‌های پرکاربرد در حوزه هوش مصنوعی می‌باشد که با ترجمه پرس‌وجو و اسناد موجود در وب، دسترسی تمامی افراد به اطلاعات با زبان‌های مختلف را میسر می‌سازد. در مقوله بازیابی اطلاعات بین زبانی، به دلیل تعدد قابل توجه اسامی افراد در پرس‌وجوها، این نوع از اسامی از اهمیت ویژه‌ای برخوردار هستند. بنابراین با تشخیص اسم افراد در پرس‌وجو، صفحات شخصی افراد که شامل این نام می‌باشند، در رتبه بالاتری قرار می‌گیرند[SWZ+08]. همچنین با توجه به اینکه افراد زیادی صفحات شخصی خود را به زبان دیگری طراحی می‌کنند، می‌توان با نویسه‌گردانی آوایی[1] پرس‌وجوها، این صفحات را در نتایج پرس‌وجو لحاظ کرد و بازیابی اطلاعات بین زبانی را انجام داد.

در این بخش، در ابتدا در مورد بازیابی اطلاعات بحث خواهد شد. پس از آن مفهوم بازیابی اطلاعات بین زبانی و کاربردهای آن وسپس چالش‌هایی که این سیستم‌ها با آن روبرو هستند، شرح داده می‌شود. همچنین مختصری در مورد تشخیص اسامی خاص در پرس‌وجوی کاربران اشاره شده و نیز نویسه‌گردانی کلمات شرح داده می‌شود. در نهایت نیز معیارهای ارزیابی سیستم‌های تشخیص نام و نویسه‌گردانی معرفی می‌گردند.

۱-۲ موتور جستجو و بازیابی اطلاعات

وب جهان‌گستر[2] به علت توزیع‌شدگی و هزینه پایین تولید محتوا، با چالش‌های جدیدی از جمله حجم زیاد اطلاعات، رشد نمایی، پویایی زیاد، ناهمگنی و غیرساختاریافته بودن اطلاعات (شامل صوت و تصویر و غیره)، مواجه شده است [زارع۸۸]. برای مثال حجم اطلاعات نمایه‌سازی شده در سال ۱۹۹۴ از ۱۱۰,۰۰۰ صفحه [McB94] به ۱۱/۵ میلیارد صفحه در سال

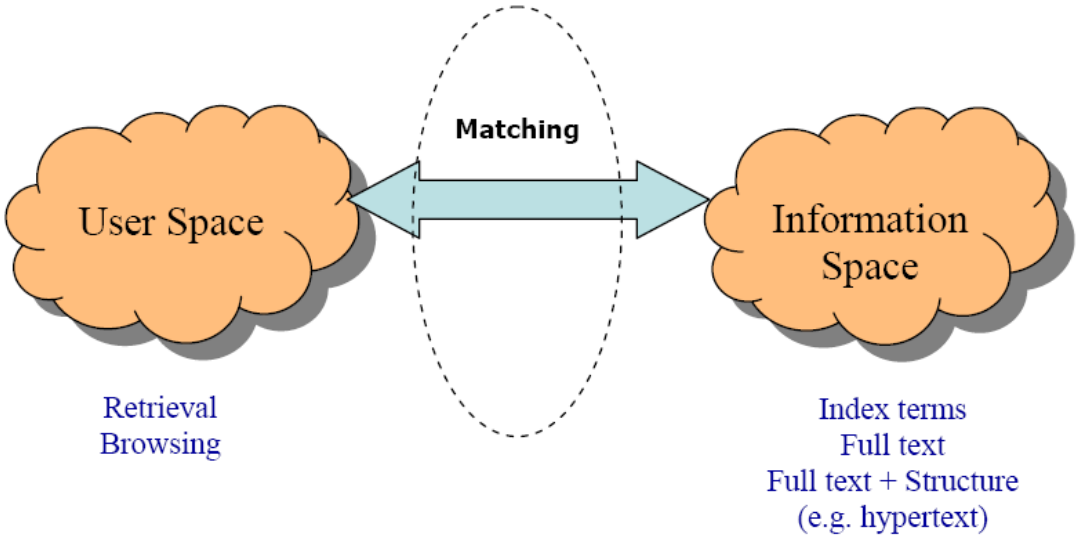
[1] Transliteration
[2] World Wide Web (WWW)

۲۰۰۵ [GS05] و پس از آن به ۲۰ میلیارد صفحه در سال ۲۰۱۱ [Web11] رشد یافته است. در حال حاضر، بهترین ابزار کارآمد برای مدیریت، بازیابی و استخراج اطلاعات مهم از این مجموعه عظیم داده‌ای، موتورهای جستجو می‌باشند. لذا حدود ۸۰٪ از کاربران اینترنت سایت‌های جدید را از طریق موتور جستجو بازیابی و استخراج می‌کنند [Inews].

۱-۲-۱ بازیابی اطلاعات[1]

براساس [زارع۸۸]، بازیابی اطلاعات شامل استانداردها و پروتکل‌های نمایش، ذخیره‌سازی، سازماندهی و دسترسی به اقلام اطلاعاتی با هدف بازیابی کلیه‌ی اسنادی که با پرس‌وجوی کاربر مرتبط است، می‌باشد. مطابق شکل ۱-۱، بازیابی اطلاعات به دو حوزه کاربر و اطلاعات و یک حوزه میانی بنام حوزه بازیابی تقسیم می‌شود. در حوزه کاربر، نیاز اطلاعاتی کاربر بیان می‌شود که در اکثر موارد باید با توجه به زبان سیستم بازیابی این نیاز بیان شود یا این‌که سیستم با توجه به رفتار کاربر، تاریخچه‌ی رفتار وی و یا اطلاعات مستقیمی که از تخصص یا موارد شخصی وی دارد، او را مدل کرده و نیاز اطلاعاتی او را پیش‌بینی کند. در حوزه اطلاعات باید اطلاعات و دانش نهفته در اسناد و داده‌ها مدل‌سازی و سازماندهی شوند. حوزه بازیابی نیز فصل مشترک این دو و تطبیق دهنده نیاز اطلاعاتی کاربر با اسناد اطلاعاتی است. نیازهای اطلاعاتی یا به صورت پرس‌وجوهای برخط[2] به حوزه بازیابی ارسال می شوند که در این حالت از مجموعه‌ی اسناد موجود، موارد مرتبط بازیابی می‌شود و یا این‌که پرس‌وجویی وجود دارد که در معرض جریان اطلاعات (مثلاً اخبار) قرار می گیرد و اسناد مرتبط را فیلتر و جدا می‌کند.

[1] Information Retrieval
[2] Online



شکل ۱-۱) نمای کلی سیستم بازیابی [BYRN99].

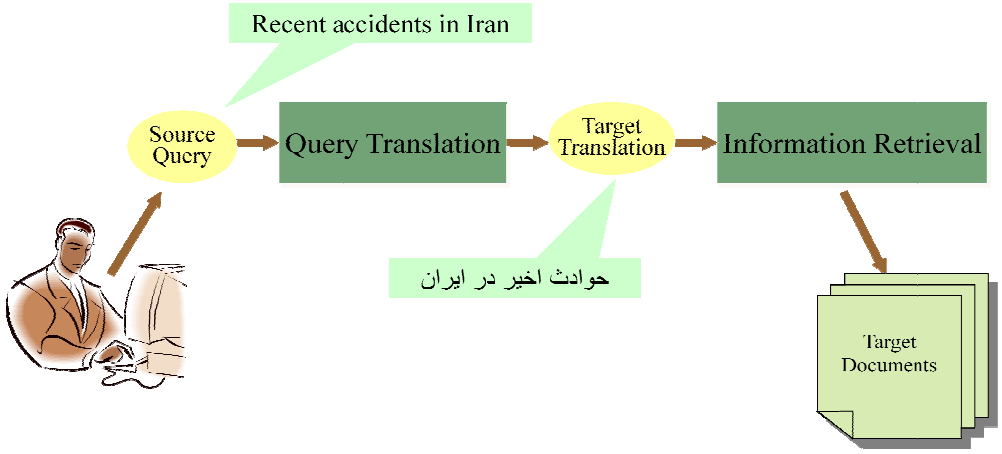
در فضای کاربر برای دستیابی به اطلاعات مفید، دو سیاست جستجو و مرور[1] وجود دارد. روش جستجو برای حالتی که کاربر دقیقاً هدف خود را می‌داند، مفید می باشد. روش مرور وقتی مناسب است که کاربر با محتوای مورد نظر خود نا آشنا باشد. دو روش جستجو و مرور، مکمل یکدیگر بوده و تاثیر عمده‌ی خود را در حالتی که با هم استفاده شوند، خواهند داشت [زارع۸۸].

۱-۳ بازیابی اطلاعات بین زبانی

بازیابی اطلاعات بین زبانی، به نوعی از بازیابی اطلاعات گفته می‌شود که نتایج بدست آمده از پرس‌وجوی کاربر در مقایسه با پرس‌وجو دارای زبان متفاوتی است. مطابق شکل ۱-۲ هدف اصلی در بازیابی اطلاعات بین زبانی، پیدا کردن اطلاعات در زبان هدف[2] (همانند فارسی) در پاسخ به پرس‌وجوهایی به زبان اصلی[3] (همانند انگلیسی) می‌باشد. امروزه با رشد وسیع فناوری وب و

[1] Browsing
[2] Target language
[3] Source language

افزایش روزافزون حجم اطلاعات الکترونیکی با زبان‌های گوناگون، بازیابی اطلاعات مستقل از زبان اسناد، اهمیت ویژه‌ای پیدا کرده است.

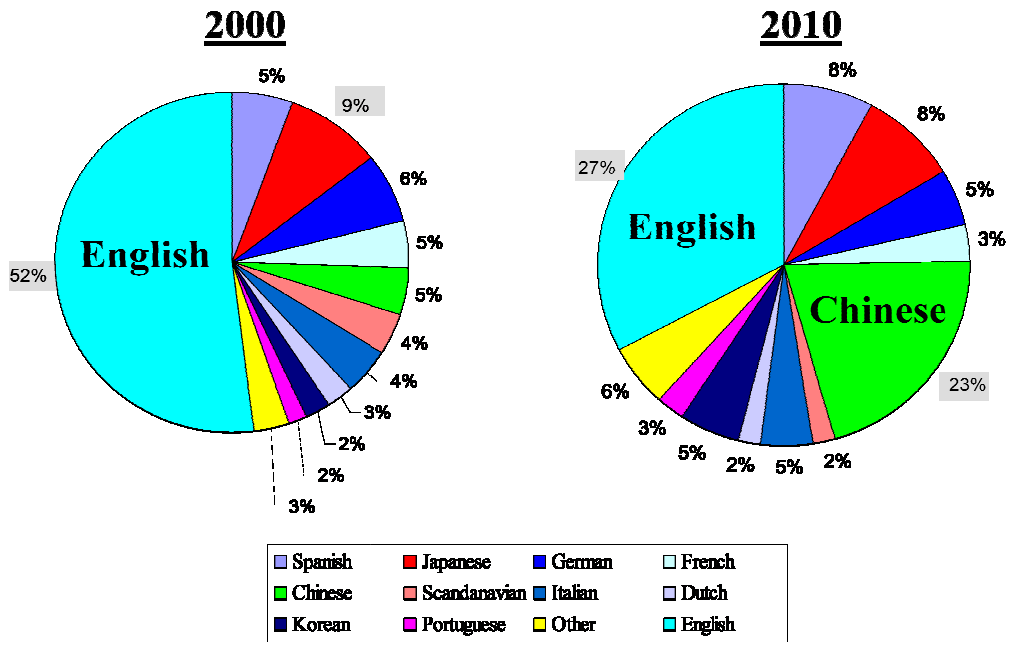


شکل ۱-۲) بازیابی اطلاعات بین زبانی.

در شکل ۱-۳، اهمیت روزافزون بازیابی اطلاعات بین زبانی مشهود است. همان‌طور که دیده می‌شود، یکی از مهم‌ترین دلایل آن محیط چندزبانه بودن[1] وب می‌باشد که در سال‌های اخیر، به بستری برای تمامی افراد با زبان‌های مختلف تبدیل شده است.

بازیابی اطلاعات بین زبانی به منظور تبادل و اشتراک اطلاعات، به کار می‌رود. این موضوع، به عنوان نمونه در امنیت ملی[2] و همچنین در دسترسی به اطلاعات مربوط به بیماری‌ها، نقش خود را بهتر نشان می‌دهد [RGB+03].

[1] Multilingual environment
[2] National security



شکل ۱-۳) چندزبانه بودن محیط وب.

این نوع سیستم‌ها زمانی کاربرد دارد که کاربرانی که توانایی خواندن چندین زبان را دارند، می‌توانند از زبانی برای نوشتن پرس‌وجویشان استفاده نمایند که برای آنها روان‌تر باشد. همچنین می‌توانند بفهمند که آیا اسنادی مرتبط به پرس‌وجوی آنها به زبان‌های دیگر وجود دارد یا خیر. از دیگر کاربردها می‌توان به بازیابی تصویر و یا ویدئو از سایت‌ها با زبان‌های گوناگون اشاره نمود [Ord97].

بازیابی اطلاعات بین زبانی شامل مقوله‌هایی از قبیل بازیابی اطلاعات، پردازش زبان طبیعی[۱]، ترجمه ماشینی و خلاصه‌سازی[۲]،پردازش گفتار[۳]، ارتباط متقابل رایانه و انسان[۴] می‌باشد.

[2] Natural language processing
[3] Machine translation and summarization
[4] Speech processing
[5] Human-Computer interaction

۱-۳-۱ چالش‌های بازیابی اطلاعات بین زبانی

همان‌طور که گفته شد، هدف از بازیابی اطلاعات بین زبانی پیدا کردن اسناد مرتبط به پرس-وجو در زبانی متفاوت با آن می‌باشد. بنابراین در نیل به این هدف، ممکن است با چالش‌های گوناگونی روبرو شود. از جمله این چالش‌ها می‌توان به ابهام کلمات در زبان‌های مختلف اشاره کرد. یک کلمه در هر زبان ممکن است دارای معانی متفاوتی باشد [هاشمی۸۹]. برای مثال کلمه python دارای سه مفهوم کاملاً متفاوت اژدها، زبان برنامه‌نویسی و غیب‌گو می‌باشد. بنابراین با داشتن چنین ابهامی در پرس‌وجوی کاربر، با مشکل مواجه می‌شویم.

از دیگر چالش‌ها می‌توان به ترجمه عبارات و اصطلاحات[1] اشاره نمود. هر زبان دارای اصطلاحات و عباراتی است که نمی‌توان با ترجمه تک‌تک کلمات آن به مفهوم آن پی برد [هاشمی۸۹]. بنابراین برای داشتن دقت بالا در یک سیستم بازیابی اطلاعات بین زبانی، می‌بایست این‌گونه اصطلاحات به درستی شناخته و ترجمه شوند.

مشکل عمده دیگری که مطرح می‌گردد، زمانی است که از واژه‌نامه دوزبانه[2] به منظور ترجمه استفاده می‌شود. در این حالت کلماتی وجود دارند که در واژه‌نامه موجود نیستند که به آنها کلمات خارج از واژه‌نامه[3] گویند[DL05]. بنابراین سیستم بازیابی اطلاعات بین زبانی برای ترجمه این‌گونه کلمات می‌بایست از روش‌های دیگری به غیر از واژه‌نامه استفاده نماید. رفع این چالش در زمینه اسامی افراد در پرس‌وجو از جمله اهداف این پایان‌نامه می‌باشد.

همچنین از دیگر مشکلاتی که در استفاده از واژه‌نامه‌ها برای ترجمه وجود دارد، عدم وجود صرف‌های گوناگون کلمات در واژه‌نامه می‌باشد [هاشمی۸۹]. به بیان دیگر، شکل پایه‌ای هر کلمه در واژه‌نامه قرار داده شده است و صرف‌ها و مشتقات گوناگون آن موجود نمی‌باشد.

---

[1] phrase
[2] Bilingual dictionary
[3] Out of dictionary

پردازش‌های ریخت‌شناسی[1] روی کلمات در بسیاری از زبان‌ها مانند عربی مشکلاتی را به همراه دارد[CLI05]. برای مثال حذف علائم تفکیک کننده[2] در این زبان ابهاماتی را به وجود می‌آورد.

۱-۴ تشخیص نام در پرس‌وجو

با گسترش اطلاعات موجود در وب نیاز به یک موتور جستجوی قوی به منظور بازیابی این اطلاعات، احساس می‌شود. یکی از مواردی که به صحت و دقت موتور جستجو کمک می‌کند، شناخت انواع پرس‌وجوها می‌باشد.

اسامی افراد یکی از پرکاربردترین پرس‌وجوها در موتورجستجو می‌باشند. براساس[SWZ+08]، با استفاده از داده‌های جمع‌آوری شده از یک موتور جستجوی تجاری، ۴~۲٪ از پرس‌وجوهای کاربران منحصراً اسامی افراد بوده و ۳۰٪ از آنها، پرس‌وجوهایی هستند که شامل اسامی افراد و کلمات دیگری غیر از این اسامی می‌باشند[AVG05]. این مقادیر در حالتی که کاربران نتایج حاصل از یک موتور جستجو را نامناسب ارزیابی کرده و موتور جستجوی دیگری را انتخاب می‌نمایند، افزایش می‌یابد. تعداد زیاد پرس‌وجوهای اسامی افراد از آنجا ناشی می‌شود که اکثر افراد، نام خود را به عنوان پرس‌وجو وارد کرده و عملکرد موتور جستجو را از طریق نتایج حاصل، ارزیابی می‌کنند. در صورت شناخت موفقیت آمیز این نوع از پرس‌وجوها، موتور جستجو می‌تواند عملکرد بهتری در رتبه‌بندی نتایج داشته باشد. به طور کلی، اگر بتوان پرس‌وجوی دارای نام را تشخیص داد، می‌توان خصیصه‌هایی هم‌چون صفحه شخصی[3] افراد و یا تعداد کلمات پرس‌وجو در صفحه، را در رتبه‌بندی نتایج اعمال کرد [SWZ+08]. همچنین با توجه به این‌که تعداد زیادی از کاربران صفحات شخصی خود را به زبان‌های دیگر طراحی می‌کنند، می‌توان پس از

[1] Morphological analyzing
[2] Punctuation mark
[3] Personal homepage