بسمه تعالی

دانشگاه تهران

پردیس دانشکده های فنی

دانشکده مهندسی برق و کامپیوتر

پیشنهاد و فرم حمایت از پایان‌نامه تحصیلات تکمیلی

کارشناسی ارشد ■ **دکتری** □

شماره مرجع * :

* شماره مرجع، توسط معاونت پژوهشی پردیس دانشکده‌های فنّی هنگام صدور ابلاغ درج خواهد شد.

۱- خلاصه اطلاعات پایان نامه

عنوان پایان نامه به زبان فارسی:

بازیابی اطلاعات بین‌زبانی انگلیسی-فارسی مبتنی بر لغت‌نامه

عنوان پایان نامه به زبان انگلیسی:

Dictionary Based English-Persian Cross-Language Information Retrieval

نوع پایان نامه: ☐ بنیادی ■ کاربردی ☐ توسعه‌ای

پردیس/دانشکده: فنی **دانشکده/گروه: برق و کامپیوتر**

مقطع تحصیلی : کارشناسی ارشد **رشته و گرایش تحصیلی: مهندسی‌کامپیوتر-نرم‌افزار**

تاریخ پیشنهاد:۹۲/۷/۳ **تاریخ تصویب:**

۲- اطلاعات اساتید راهنما و مشاورین

نوع مسئولیت	نام و نام‌خانوادگی	مرتبه علمی	محل خدمت	امضاء
استاد راهنما (مجری)	**دکتر آزاده شاکری**	استادیار	دانشگاه تهران	
استاد راهنمای دوم (حسب نیاز)	دکتر هشام فیلی	استادیار	دانشگاه تهران	
استاد مشاور				
استاد مشاور دوم (برای دکتری)				

۳- اطلاعات دانشجو

نام و نام‌خانوادگی: جاوید داداش‌کریمی **شماره دانشجویی:** ۸۱۰۱۹۱۴۱۰

رشته و گرایش تحصیلی: مهندسی‌کامپیوتر-نرم‌افزار **دانشکده:** برق و کامپیوتر **مقطع تحصیلی:** کارشناسی ارشد

پست الکترونیک :dadashkarimi@ut.ac.ir **تلفن ثابت:** ۰۴۱۲-۷۲۶۱۱۵۱ **تلفن همراه:**۰۹۱۴۱۷۶۴۰۷۲

۴- مشخصات موضوعی پایان نامه

تعریف مسأله، هدف و ضرورت اجرای (حداکثر سه صفحه)

گسترش وب در سال‌های اخیر و حجم انبوه اطلاعات در این فضا باعث گردیده‌است که محققان علم بازیابی اطلاعات روش‌های نوینی را چه در جهت بهبود کارایی[۱] و چه در جهت افزایش فراخوان[۲] ارایه‌دهند. حال آن که با گذشت زمان زمان سهم زبان‌های غیر انگلیسی در اسناد این فضا قابل توجه است. در یکی دو دهه گذشته و همزمان با تولد وب روش‌های متفاوت ذخیره‌سازی داده‌ها و بازیابی آن‌ها از مسایل بسیار مهم محققان بوده‌است. بازیابی اطلاعات به صورت تک زبانه مورد توجه جدی آن‌ها بوده و حال آن که بازیابی بین‌زبانی به دلیل وجود برخی داده‌های مورد نیاز کاربر در زبان غیر بومی وی، چالشی جدید برای علم بازیابی اطلاعات به شمار می‌رفت.

بازیابی اطلاعات بین‌زبانی به عمل بازیابی اسنا در زبان مقصد اطلاق می‌شود که در پاسخ به نیاز کاربر بصورت به اصطلاح کیسه کلمات[۳] به زبان کاربر به سیستم داده می‌شود به طوری که زبان اسناد و کاربر متفاوت باشد. لذا ساده‌ترین راه برای پیاده‌سازی چنین سیستمی ترجمه است. برخی روش‌های پیشین در این حوزه از پیکره‌های موازی و تطبیقی برای ترجمه بهره می‌برده‌اند و این در حالی بود که استفاده از لغت‌نامه بعنوان یک منبع جامع همواره مورد توجه بوده‌است. لغت‌نامه در اصل به منبعی گفته می‌شد که فهرستی از کلمات در زبان مقصد را بعنوان ترجمه یک کلمه در زبان منبع دارا باشد. روش‌هایی که تاکید داشتند تنها از لغت‌نامه برای ترجمه استفاده کنند به دلیل اضافه نمودن اختلال[۴] در فهرست کلمات ترجمه شده و همین‌طور ابهام در حین ترجمه و یا ابهام در زبان مقصد [۲۰] با مشکلات عدیده‌ای روبرو بوده‌اند که نمود آن در کاهش کارایی روش‌های بازیابی اطلاعات بین‌زبانی نسبت به بازیابی تک‌زبانی مشهود است.

ابهام درون زبانی[۵] که خود می‌تواند نحوی[۶] یا معنایی[۷] باشد زمانی اتفاق می‌افتد که به ترتیب ساختار جمله برداشت‌های متفاوتی را القاء کند و یک کلمه حالت‌های[۸] متفاوتی داشته‌باشد. ابهام بین‌زبانی هم می‌تواند لغوی[۹] و ساختاری باشد که وقوع کلمات خارج از لغت‌نامه[۱۰]، کلمات با چندین حالت و کلمات مرکب[۱۱] از جمله مثال‌های این دو دسته هستند. رفع ابهام از منظر لغت‌نامه و به‌کار بستن هم‌رخدادی ترجمه‌ها در پیکره‌های غیر تراز شده[۱۲]، برچسب کلمات در جمله، استفاده از شبکه کلمات[۱۳] و بهره‌گیری از پیکره‌های موازی برای امر ترجمه از راهکارهای پیشنهادی گذشته بوده‌است. حال آن که فنون متفاوتی نیز برای پیاده‌سازی هر یک از آن ها ارایه گردیده‌است.

لغت‌نامه خود نیز می‌تواند ویژگی‌های متفاوتی داشته‌باشد که در کارایی بازیابی موثر است. برخی از روش‌ها اندازه[۱۴] لغت‌نامه برای هر کلمه ورودی که میانگین تعداد ترجمه‌های یک کلمه در لغت‌نامه تعریف می‌شود را بعنوان یک ویژگی برای بررسی انتخاب‌کرده‌اند و برخی نیز پوشش[۱۵] یک لغت‌نامه که درصد لغت‌های پرس‌وجو که در لغت‌نامه وجود دارد را تحلیل کرده‌اند. استفاده از این دو معیار در کنار هم نیز نتایج قابل توجهی داشته‌است. بحث ترکیب لغت‌نامه‌های مختلف برای دستیابی به یک روش کارآمد نیز جزو کارهای مرتبط در این حوزه بوده‌است که در برخی پژوهش‌ها نتایج آن‌ها قابل توجه است. لذا چنین روش‌هایی نشان از آن دارد که می‌توان با بررسی ویژگی‌ها و روش‌های مبتنی بر لغت‌نامه مشکلات ابهام و اضافه کردن کلمات نامرتبط در فرایند ترجمه را کاهش‌داد.

---

[1] Performance
[2] Recall
[3] Bag of Words
[4] noise
[5] Within-Language Ambiguity
[6] syntactic
[7] semantic
[8] sense
[9] lexical
[10] Out Of Vocabulary
[11] phrase
[12] Unaligned Corpora
[13] Word Net
[14] scale
[15] coverage

در زبان فارسی که محور اصلی این پیشنهاد به شمار می‌رود، روش‌های مبتنی بر لغت‌نامه برای بازیابی اطلاعات بین زبانی انگلیسی-فارسی معمولا یا انجام نشده‌است و یا از روش‌های بسیار ساده برای استخراج ترجمه‌ها استفاده شده‌است. چنین روش‌هایی یا همه ترجمه‌های موجود در لغت‌نامه برای یک کلمه را در ترجمه نهایی آورده‌اند و یا ترجمه‌های با رتبه بالا[16] را انتخاب می‌کنند. لذا کارایی چنین روش‌هایی در زبان فارسی کم‌تر از ۳۱٪ میزان مشابه بدست‌آمده در روش‌های تک‌زبانه می‌باشد[۲۱]. حال آن که مقایسه آن با زبان‌های دیگری هم چون زبان عربی (۶۸٪ میزان مشابه در تک زبانه)، که دو زبان هم ریشه[17] به شمار می‌روند، قابل توجه است [۱۸]. لذا نیاز به یک مطالعه عمیق و جدی، مخصوصا در زبان فارسی برای بهبود کارایی چنین روش‌هایی کاملا احساس می‌شود. این که آیا مشکلات موجود در خود زبان فارسی مسئول چنین مشکلاتی است، آیا لغت‌نامه‌ها ترجمه‌های متناسبی را برای ماشین فراهم نمی‌کنند، آیا روش‌های موجود و اشاره شده می‌توانند راه‌گشای بازیابی اطلاعات بین زبانی در زبان فارسی باشند و آیا روش‌های دیگری برای زبان فارسی بایستی به کار برده‌شود از سوال‌هایی هستند که در این پژوهش مورد بررسی قرار خواهند گرفت.

هدف از اجرای این پژوهش یافتن روش‌های نوین و اعمال روش‌های اشاره شده، بررسی راهکارهای حل مشکلات بازیابی بین زبانی انگلیسی-فارسی مبتنی بر لغت‌نامه، مطاللعه ویژگی‌های لغت‌نامه‌های انگلیسی-فارسی موجود و ارائه یک راهکار نوین در روش‌های بازیابی اطلاعات بین‌زبانی مبتنی بر لغت‌نامه در چهارچوب زبان‌های انگلیسی-فارسی است.

روشها و فنون اجرایی طرح

لغت‌نامه یک منبع قابل اطمینان برای ترجمه به حساب می‌آید. تقریبا می‌توان گفت که یک لغت‌نامه جامع همه ترجمه‌های ممکن در حوزه‌های مختلف را برای یک کلمه ورودی آن فراهم می‌آورد. روش‌های گذشته معمولا از لغت‌نامه دوزبانه خوانا توسط ماشین برای ترجمه بهره می‌بردند. برخی از روش‌ها در کنار لغت‌نامه از پیکره‌های موازی یا تطبیقی نیز برای ترجمه یا توسعه پرس‌وجو[18] استفاده کرده‌اند[۶، ۷، ۱۱، ۱۵]. این گونه روش‌ها از روش جایگذاری کلمه به کلمه[19] برای ترجمه اولیه استفاده می‌کنند. لذا برای رفع ابهام ترجمه، گاهی هم‌رخدادی[20] کلمات پرس‌وجو در پیکره‌ها به کار گرفته‌می‌شود و ترجمه‌هایی که احتمالا تناسب بیش‌تری با متن پرس‌وجو دارند انتخاب می‌شوند [۱۵].

ریشه‌یابی[21]، هنجارسازی[22]، حذف کلمات رایج[23] ، استفاده از برچسب کلمات در جمله[24]، بررسی کلمات مرکب[25] و همچنین عبارت‌ها[26] نیز از روش‌ها رایج پردازش متن هستند که معمولا در شاخه فنون مبتنی بر لغت‌نامه به کار گرفته می‌شوند که در بیش‌تر مواقع باعث بهبود کارایی می‌گردند [۱۵، ۲۰]. لذا به‌کاربستن لغت‌نامه‌ای از عبارات و کلمات مرکب [۱۹]، بهره‌گیری از پیکره‌های موازی [۲۰] و ماشین‌حالت [۵] از جمله روش‌های شناسایی این نوع کلمات به شمار می‌روند.

[16] Select top N
[17] cognate
[18] Query Expansion
[19] Word by Word Replacement
[20] Co-Occurrence
[21] Stemming
[22] Normalization
[23] Stop Words
[24] Part of Speech Tagging
[25] Compound
[26] Phrase

عدم تمرکز روش‌های بین‌زبانی گذشته در زبان فارسی-انگلیسی بر روش‌های مبتنی بر لغت‌نامه توجه جدی بر این روش ساده و در عین حال کارآمد را نیازمند است. این که بتوان از ویژگی‌های مختلف لغت‌نامه‌های گوناگون برای بهبود کارایی استفاده‌کرد نیز از روش‌های پیش‌روی این شاخه به خصوص در زبان فارسی است. تعریف نوین از ابهام در کلمات ترجمه [۲۰] نیز می‌تواند وزن‌دهی ترجمه‌های یک کلمه ورودی لغت‌نامه را تحت‌تاثیر قرار داده و افزایش چشم‌گیری در کارایی بازیابی اطلاعات را موجب شود. استفاده از پرس‌وجوهای ساخت‌یافته که اولین بار توسط [۸] ارایه شد، راهکار نوینی در ایجاد تعادل در وزن دهی کلمات پرس‌وجو -که تعداد ترجمان‌های متفاوتی در یک لغت‌نامه ممکن است داشته‌باشد- بوده‌است. این گونه روش‌ها با ارایه بسامد تکرار مجموعه ترجمان‌های یک کلمه در یک سند[۲۷] و عکس بسامد وقوع مجموعه در اسناد پیکره[۲۸] [۸]، سعی در تعدیل اثر تعداد ترجمان‌های یک کلمه در یک لغت‌نامه بر روی بازیابی اسناد داشته‌اند. چنین روش‌هایی مخصوصا در زبان فارسی یا کم‌تر به کار گرفته‌شده‌اند و یا بهبود قابل توجهی در کارایی ندشاته‌اند و در زبان‌های غیر-فارسی نیز با چالش‌های جدی روبرو هستند [۱۱، ۲۴].

بنابر این در این پژوهش سعی خواهد شد تا بصورت کاملا عمیق ویژگی‌های لغت‌نامه‌های موجود انگلیسی به فارسی را مورد بررسی قراردهیم. لذا با تعریف برخی ویژگی‌های مجزا تاثیر ترجمه‌های مختلف را روی کارایی بازیابی بین‌زبانی مطالعه‌خواهیم‌کرد. این که چگونه می‌توان از روش‌های پردازش متن[۲۹] نیز در این شاخه استفاده‌کرد از چالش‌های پیش روی ماست. ایجاد پرس‌وجوهای ساخت‌یافته[۳۰] برای وزن‌دهی پرس‌وجوها به شکلی که ابهام در ترجمه را به حداقل برساند نیز از کارهای محوری این مطالعه خواهدبود. این که چگونه عدم ترجمه مناسب عبارت‌ها و کلمات مرکب می‌تواند کارایی یک روش را تحت تاثیر قرار دهد و نیز راه‌کارهای رفع ابهام آن‌ها در راستای این مطالعه بررسی خواهندشد.

به طور کلی چهارچوب این پژوهش و فنون اجرایی آن را می‌توان به صورت زیر خلاصه‌کرد:

- بررسی روش‌های گذشته بازیابی بین زبانی انگلیسی-فارسی به خصوص جایگذاری کلمه‌به‌کلمه ترجمه‌ها بعنوان معیار پایه[۳۱].
- مطالعه و بررسی ویژگی‌ها و تفاوت‌های لغت‌نامه‌های انگلیسی-فارسی موجود به خصوص اندازه[۳۲] و پوشش[۳۳] لغت‌نامه‌ها و تاثیر آن‌ها در کارایی عمل بازیابی.
- ادغام لغت‌نامه‌ها و تاثیر آن در کاهش ابهام و افزایش کارایی.
- استفاده از شکبه کلمات برای توسعه پرس‌وجوها و انتخاب ترجمه‌های مرتبط.
- بررسی درجه‌ابهام و اعمال آن در وزن‌دهی ترجمه‌ها و کلمات پرس‌وجو.
- استفاده از پرس‌وجوهای ساخت‌یافته در کنار وزن‌دهی‌های مبتنی بر درجه ابهام ترجمه‌ها.

معیار ارزیابی روش‌های ارایه شده در این پیشنهاد و روش مورد مطالعه بر اساس میانگین متوسط دقت[۳۴] می‌باشد و مجموعه داده‌ای همشهری و INFILE مجموعه داده‌ای هستند که کارایی روش پیشنهادی بر روی آن بررسی خواهدشد [۲۲، ۲۵]. لغت‌نامه‌های آنلاین موجود برای زبان فارسی نیز از منابع ما برای لغت‌نامه دوزبانه خواهند بود. لغت‌نامه آریان‌پور که بعنوان مبنای کار روش‌های قبلی قرار گرفته‌بود نیز مورد توجه جدی این مطالعه خواهدبود [۲۱، ۲۴]. برای محاسبه معیار میانگین متوسط دقت ابتدا مقدار متوسط دقت برای هر پرس‌وجو محاسبه شده و از

[27] Term Frequency
[28] Inverse Document Frequency
[29] Natural Language Processing
[30] Structured Query
[31] Baseline
[32] scale
[33] coverage
[34] Mean Average Precision

میانگین این مقادیر (برای همه پرس‌وجوها) برای محاسبه میانگین متوسط دقت استفاده می‌شود. معیار میانگین متوسط دقت در صورتیکه فایل داوری ارتباط اسناد تنها بصورت ارزیابی دودویی باشد می‌تواند مورد استفاده قرار گیرد در صورتیکه ارزیابی صورت گرفته دودویی نباشد می‌توان از معیار NDCG استفاده کرد.

تکرار ترجمه‌ها در فهرست ترجمه یک کلمه ورودی لغت‌نامه از موضوعات کلیدی خواهدبود که قرار است مورد بررسی قرارگیرد. این که تکرار مجاز باشد یا برای برخی ترجمه‌های با ابهام کمتر صورت پذیرد و یا به طور کلی نباشد و نیز اندازه کم لغت‌نامه و تاثیر آن بر از دست‌دادن ترجمه‌های بسیار مهم در راستای این کار قرار می‌گیرند.

بررسی وزن‌دهی کلمات پرس‌وجو قبل از عمل ترجمه و همین‌طور وزن‌دهی ترجمه‌ها بعد از عمل ترجمه بر اساس درجه ابهام و درجه اهمیت (اولویت ترجمان‌ها) از کارهای اساسی خواهدبود که کارایی این‌گونه روش‌ها مورد بررسی قرارخواهدگرفت.

روش‌های بازیابی مبتنی بر لغت‌نامه که از فنون پرس‌وجوهای ساخت‌یافته بهره می‌جویند نیز به کار گرفته‌خواهدشد. این‌گونه روش‌ها از ساختار هم‌خانواده [۳۵] و ساختار عبارت برای مجموعه ترجمه‌ها به جای جایگذاری همه ترجمه‌ها استفاده می‌کنند. لذا بررسی این روش‌ها در زبان فارسی و لغت‌نامه‌های موجود جزو کارهای ضروری خواهدبود.

پیشینه تحقیق (همراه با ذکر منابع اساسی)

امروزه بخش عمده‌ای از اسناد الکترونیکی موجود، در فضای وب قرار دارد. موتورهای جست‌وجو با فراهم‌آوردن امکان بازیابی این اسناد، امر دسترسی به آن‌ها را تسهیل نموده‌اند. این اسناد معمولا به زبان‌های مختلف نگاشته شده‌اند و کاربر اسناد مرتبط با پرس‌وجوی خود را براساس میزان ارتباط سند و پرس‌وجو به صورت یک فهرست مشاهده می‌کند. در دهه‌ی اخیر علم بازیابی‌اطلاعات پس از ارایه مقالاتی در زمینه‌ی حجم وب و سهم زبان‌های مختلف، وارد عرصه نوینی شد[۱، ۲، ۳]. [۲] نشان داد که سهم اسناد انگلیسی در فضای وب بیش‌تر از سایر زبان‌ها می‌باشد.

دستیابی به اسناد غیربومی نیز در این سال‌ها همواره از موضوعات مهم بازیابی اطلاعات بوده‌است. کاربر پرس‌وجوی خود را به زبان بومی وارد سیستم می کند و اسناد به زبان انگلیسی و یا غیربومی توسط سیستم نمایش داده‌می شوند. روش‌های مرسومی را که برای پیاده‌سازی چنین سیستمی در این سال‌ها به کار برده شده‌اند به ۳ دسته می‌توان طبقه‌بندی‌کرد: ۱- ترجمه پرس و جو به زبان مقصد. ۲- ترجمه اسناد به زبان مبدا و ۳- تبدیل پرس‌وجو و اسناد به یک فضای مشترک مستقل از زبان. به دلیل زمان‌بر بودن روش اول معمولا از روش‌های بعدی برای عملیات بازیابی استفاده‌می‌شود [۴، ۸، ۱۶، ۱۸، ۱۹، ۲۰].

برای این که پرس‌وجوی کاربر ترجمه شود، روش‌های متفاوتی ممکن‌است به کار برده‌شوند. استفاده از ماشین ترجمه، ترجمه‌های مبتنی بر پیکره و روش‌های بر مبنای لغت‌نامه از جمله این روش‌هاست [۴، ۸، ۱۸، ۱۹، ۲۰]. ماشین‌های ترجمه معمولا کارآمدترین ابزار برای ترجمه جمله‌های ساختاری (کامل) محسوب می‌شوند. ماشین ترجمه به دلیل کمیاب بودن بین جفت‌زبان‌ها (که به دلیل سختی ایجاد این ابزار است) و کم‌تر بودن جمله‌های ساختاری در پرس‌وجوها معمولا با اقبال کم‌تری بازیابی اطلاعات بین‌زبانی روبرو بوده‌است [۱۷]. پیکره‌های موازی و تطبیقی نیز از دیگر روش‌های ترجمه به شمار می‌روند که با به کارگیری روش‌های آمار و احتمالاتی، احتمال ترجمه کلمات مبدا به کلمه‌های متناظر در زبان مقصد را محاسبه کرده و پرس‌وجو را بر اساس آن ترجمه می‌کنند. این گونه روش‌ها

[35] synonym

نیز به دلیل کم‌یاب بودن بین جفت زبان‌ها و وابستگی ترجمه به دامنه پیکره، مشکلات خاص خود را دارند [۴، ۷، ۸].

همان طوری که در بخش‌های قبلی اشاره شد، برای بازیابی بین‌زبانی استفاده از پیکره‌های موازی و تطبیقی و روش‌های احتمالاتی مبتنی بر پیکره، از رایج‌ترین روش‌های به کار رفته طی یک دهه اخیر بوده‌اند [۶، ۷، ۱۱، ۱۵، ۱۶، ۲۱]. این‌گونه روش‌ها نیازمند ایجاد پیکره‌های مورد نیاز برای انجام عملیات ترجمه می‌باشند که یا وجود ندارند و یا ساخت چنین پیکره‌هایی مشکل است. لذا روش‌های مبتنی بر لغت‌نامه از اولین و ساده‌ترین روش‌های بازیابی اطلاعات بین‌زبانی بوده‌است که برخی ویژگی‌ها و مشکلات منحصر به فرد لغت‌نامه‌ها جهت این حوزه را به سمت روش‌های مبتنی بر پیکره سوق داده‌است. این حوزه به دلیل استقلال ترجمه‌های لغت‌نامه و متن پرس‌وجو و در نتیجه بازیابی اسناد نامرتبط، ویژگی‌های منحصر به فرد لغت‌نامه‌ها، وزن‌دهی نامناسب کلمات پرس‌وجو ، ابهام بین‌زبانی و درون زبانی [۲۰] و عدم ترجمه مناسب کلمات مرکب، دارای کارایی پایین‌تری نسبت به روش‌های تک‌زبانه است. [۵، ۶] از جمله روش‌های ارایه شده برای رفع مشکل کلمات مرکب و [۱۴، ۱۵، ۲۰] از روش‌های رفع ابهام با به کارگیری پیکره‌های موازی یا غیر موازی است. عمده این روش‌ها یا نتوانسته‌اند به کارآمدی روش‌های تک‌زبانه عمل‌کنند یا برای برخی زبان‌ها کارایی قابل‌قبولی نداشته‌اند و یا مبتنی بر پیکره‌های موازی بوده‌اند که در برخی زبان‌ها نایاب هستند.

روش‌های مبتنی بر لغت‌نامه به دلیل آن که لغت‌نامه دوزبانه[۳۶] بین بیش‌تر جفت‌زبان‌ها موجود است و معمولا جزو منابع جامع دسته‌بندی می‌شود، در طول یک دهه اخیر مورد توجه قرارگرفته‌است [۴، ۸، ۱۵، ۱۹، ۲۰]. این گونه لغت‌نامه‌ها که قابلیت خوانایی توسط ماشین[۳۷] را دارند، در حقیقت فهرستی از کلمات را بعنوان ترجمه یک کلمه ورودی لغت‌نامه[۳۸] معرفی می‌کنند[۱۹، ۲۰].

مطالعه‌ی ویژگی‌های یک لغت‌نامه نیز از حوزه‌های بسیار مهم در این شاخه است. پوشش و مقیاس لغت‌نامه‌ها از موضوعات مهمی بوده که تاثیر آن‌ها بر روی کارآمدی روش مبتنی بر لغت‌نامه مورد بررسی قرارگرفته‌است [۱۹]. این که چگونه لغت‌نامه‌های با مقیاس متفاوت می‌توانند در کنار هم به فرآیند ترجمه و بهبود کارایی بازیابی اطلاعات کمک کنند در این مقاله مورد آزمون قرار گرفته‌است. پیشنهاد معیار درجه ابهام[۳۹](تعداد کلمات زبان مقصد که در صورت ساخت لغت‌نامه مقصد به مبدا برای یک کلمه فهرست می‌شوند) توسط [۱۹] و حذف برخی ترجمه‌های با درجه ابهام بالاتر نیز از فنون چالش‌برانگیز این شاخه بوده‌است.

برخی پژوهش‌ها از وزن‌دهی کلمات پرس‌وجو قبل و بعد از ترجمه بعنوان راهکارهای حذف اختلال و رفع ابهام سخن گفته‌اند [۱۹]. برخی نیز مجموعه ترجمه‌ها و پرس‌وجوهای ساخت‌یافته را بعنوان راه‌حلی برای آن دسته ترجمه‌هایی که لغات پرس‌وجویی که تعداد ترجمه‌هایی بیش‌تری در لغت‌نامه دارند و در نتیجه وزن بیش‌تری در عمل بازیابی می‌گیرند معرفی کرده‌اند [۸].

ترجمه‌های Transitive نیز از موضوعات مهم بازیابی اطلاعات بین‌زبانی بوده‌است که به دلیل افزایش ابهام و درنتیجه کاهش کارایی روش بازیابی و نیز اهمیت این شاخه در ترجمه زبان‌هایی که منابع مستقیمی برای ترجمه ندارد مورد توجه بوده‌است [۱۷، ۲۰]. لذا این که ترجمه Transitive حتی بتواند توان رقابت با بازیابی بین‌زبانی داشته‌باشد و باعث کاهش ابهام بدلیل افزایش منابع ترجمه شود از موضوعات مهم در شاخه بازیابی اطلاعات در آینده به شمار می‌رود.

[36] Bilingual Dictionary
[37] Machine Readable Dictionary
[38] Dictionary Entry
[39] Degree of Ambiguity

منابع

[1] Numberg , "Languages In The Wired World, " *The Politics of Language and The Building of Modern Nations, France*, Paris, May, 1998.

[2] Xu J.L ,*"Multilingual Search on the World Wide Web," Proceedings of the Hawaii International Conference on System Science HICSS-33, Maui, January, 2000.*

[3] G. Grefenstette, J. Nioche, "Estimation of English and non-English Language Use on the WWW," 6 chemin de Maupertuis, Meylan, France, 2000.

[4] K. Kishida, "Technical issues of Cross-Language Information Retrieval: A Review," *Information Proccessing and Management, pp.433-455, June, 2004*.

[5] A. Chen, H.Jiang and F.Gey "English-Chinese Cross-Language IR using Bilingual Dictionary," in Proceeding of Text Retrieval Conference, 2000.

[6] G. Cao, J. Gao, J.Y Nie and J.Bai, "Extending Query Translation to Cross-Language Query Expansion with Markov Chain Models," in *CLKM'07*, Lisboa, Portugal, 2007.

[7] C.Monz and B.J Dorr, "Iterative Translation Disambiguation for Cross-Language Information Retrieval," in *Proceedings of The 5th Conference of SIGIR*, Salvador, Brazil, August, 2005, pp. 15-19.

[8] A. Pirkola, "The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval," *in Proceedings of The SIGIR Conference*, Melbourne Australia, 1998.

[9] L.S. Larkey and M.E. Connell, "Structured Queries, Language Modelling, and Relevance Modelling in Cross-Language Information Retrieval,".

[10] R. Sperer and D.W. Oard, "Structured Translation for Cross-Language Information Rtrieval," in ACM SIGIR, 2000.

[11] D.W. Oard and J. Wang, "NTCIR-2 ECIR Experiments at Maryland: Comparing Pirkola's Structured Queries and Balanced Translation," *Mandarin-English Informatin project at Jhons Hopkins University, Summer, 2000.*

[12] T.Hedlund, "Dictionary-Based Cross-Language Information Retrieval: Principles, System, Design and Evaluation," Academic Dissertation of Tampere University, November, 2003.

[13] R. Lehtokangas, H. Keskustalo and K.Jarvelin, "Dictionary-Based CLIR Loses Highly Relevant Documents," in Proceedings of ECIR, 2005, 421-432.

[14] K. Marko and S. Schulz, "Bootstrapping Dictionaries for Cross-Language Information Retrieval," in *Proceedings of The 5th International Conference on SIGIR*, Salvador, Brazil, 2005.

[15] Y. Liu, R.Jin and J.Y. Chai, "A Maximum Coherence for Dictionary-Based Cross-Language Information Retrieval," in Proceedings of 5th Conference of SIGIR, Salvador, Brazil, 2005, pp. 15-19.

[16] J.Y. Nie "Cross-Language Information Retrieval," *Morgan and Claypool Publisher*, vol. 7*(2)*, pp. 29-55, 2010.

[17] R. Lehtokanagas, E. Airio and K. Jarvelin, "Transitive Dictionary Translation Challenges Direct Dictionary Translation in CLIR," in Proceedings of Information Precessing and Management, 2004, pp. 973-988.

[18] G.A. Levow, D.W. Orad and P. Resnik, "Dictionary-Based Techniques for Cross-Language Information Retrieval," in *Information Processing and Management, June, 2004, pp. 523-547*.

[19] D.Nic. Gearailt, "Dictionary Characteristics in Cross-Language Information Retrieval," *P.H.D Dissertation in Cambridge University*, February, 2005.

[20] L.A. Ballesteros, "Resolving Ambiguity for Cross-Language Information Retrieval: A Dictionary Approach," in Proceedings of ECAI Conference University, September, 2001.

[21] H. Azarbonyad, A. Shakery and H. Faili, "Using Learning to Rank Approach for Parallel Corpora Based Cross Language Information Retrieval," in Proceeding of 20th European Conference on Artificial Intelligence (ECAI), Montpellier, France, 2012.

[22] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar and F. Oroumchian, Hamshahri: A standard Persian text collection, Knowledge-Based Systems 22(5), pp.382-387, 2009.
[23] H.B. Hashemi, Using Comparable Corpora for English-Persian Cross-Language Information Retrieval, M.Sc. Thesis, University of Tehran, Tehran, Iran, 2011.
[24] H.B. Hashemi, A. Shakery and H. Faili, "Creating a Persian-English Comparable Corpus," in proceedings of Conference on Multilingual and Multimodal Information Access Evaluation (CLEF), pp. 27-39.Padua,Italy, 2010.
[25] R. Besançon, S. Chaudiron, "Djamel Mostefa, Ismail Timimi, Khalid Choukri and Meriama Laïb," Overview of CLEF 2009 INFILE track, 2009.
[26] J. Wang, "Matching Meaning for Cross-Language Information Retrieval," Doctor of Philasofy in University of Maryland, 2005.

۵- مصوبه شورای پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر

۵-۱- فرم پیشنهاد و حمایت از پایان‌نامه در تاریخ ..... در شورای پژوهشی و تحصیلات تکمیلی دانشکده /گروه مطرح و نظر شورا به شرح زیر اعلام می‌شود:

☐ تصویب شد ☐ نیاز به اصلاح دارد ☐ به تصویب نرسید

۵-۲- عنوان طرح جامع تحقیقات استاد راهنما: سیستم‌های اطلاعاتی و محیط‌های هوشمند

امضاء استاد راهنما

۵-۳- آیا پایان نامه پیشنهادی مرتبط با طرح جامع تحقیقات استاد راهنما/مشاور/گروه آموزشی/ دانشکده می باشد:

☐ بلی ☐ خیر

امضاء رئیس / معاون پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر

شماره:
تاریخ:

معاون محترم آموزشی و تحصیلات تکمیلی پردیس دانشکده های فنی
با سلام و احترام،

فرم پیشنهاد و حمایت از پایان‌نامه کارشناسی ارشد آقای جاوید داداش‌کریمی **با عنوان** بازیابی اطلاعات بین‌زبانی انگلیسی-فارسی مبتنی بر لغت‌نامه **به راهنمایی** خانم دکتر آزاده شاکری **در شورای پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر مورخ ..... به تصویب رسید.**
خواهشمند است دستور فرمایید اقدامات مقتضی انجام شود.

امضاء رئیس / معاون پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر

شماره:
تاریخ:

معاون محترم پژوهشی پردیس دانشکده های فنی
با سلام و احترام ،
به پیوست فرم پیشنهاد و حمایت از پایان‌نامه تحصیلات تکمیلی با مشخصات مذکور که به تصویب شورای پژوهشی وتحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر رسیده است، جهت دستور اقدام مقتضی تقدیم می‌شود.

امضاء معاون آموزشی و تحصیلات تکمیلی پردیس دانشکده های فنّی

رونوشت: معاون محترم پژوهشی و تحصیلات تکمیلی دانشکده مهندسی برق و کامپیوتر: جهت اطلاع و پیگیری