Downloaded from 4:24 IRST on Friday November 16th 2018

طراحی و پیاده‌سازی یک سامانه بازیابی اطلاعات دوزبانه با استفاده از پیکره‌های زبانی

امین نظارات[1]
مهندسی فنّاوری اطلاعات، دانشگاه آزاد اسلامی، واحد یزد

طیبه موسوی میانگاه*
دانشیار، دانشگاه پیام نور، استان یزد

دریافت: ۱۳۸۹/۰۷/۲۰ | پذیرش: ۱۳۹۰/۰۱/۲۱

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۸۲۲۳- ۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱- ۲۲۵۱
نمایه در LISA، SCOPUS و ISC
http://jist.irandoc.ac.ir
ویژه‌نامهٔ ذخیره، بازیابی و مدیریت اطلاعات
ص-ص ۱۹۷- ۲۱۲ زمستان ۱۳۹۰

نوع مقاله: پژوهشی

چکیده: بازیابی اطلاعات بین زبانی به فرایندی گفته می‌شود که طی آن یک کاربر، جستاری (یک واژه، عبارت، یا حتی جمله‌ای) را به یک زبان جستجو می‌کند درحالی‌که انتظار دارد نتایج جستجوی خود را به زبان دیگری دریافت نماید. یکی از مشکلات عمده کاربران فارسی‌زبان در استفاده از منابع موجود در فضای سایبر، عدم امکان بازیابی موضوعات مورد نظر است که این مسأله تا حد زیادی به حجم کم اطلاعات به زبان فارسی در این فضا برمی‌گردد. استفاده از فرهنگ لغت نیز به‌دلیل عدم توانایی در ارائه پاسخ مناسب به ترکیبات چندتایی رایج در زبان‌ها کمتر در این زمینه راه‌گشاست. طرح حاضر که با هدف یافتن راه‌حلی مناسب برای این مشکل با تهیه نرم‌افزار آزمایشگاهی مرتبط تعریف شده است، سعی دارد که با استفاده از پیکره‌های یک‌زبانه و دوزبانه و با کمک الگوریتم‌های رایانه‌ای راه‌حل منطقی و مقرون به‌صرفه‌ای برای این مشکل ارائه نماید. به‌منظور آزمودن کیفیت کار سامانه طراحی‌شده در این طرح، آزمایشی بر روی تعداد ۱۰۰ ترکیب از زبان فارسی و انگلیسی انجام شد که برونداد سامانه بازیابی اطلاعات برای این مجموعه از ترکیبات بسیار رضایت‌بخش بوده است. یکی از دستاوردهای اجرایی این طرح، بالا بردن دقت سامانه‌های بازیابی اطلاعات در موتورهای جستجو است که با استفاده از پیکره و بانک اطلاعاتی، ترکیب‌بندی واژه‌ها قابل دسترس است.

کلیدواژه‌ها: بازیابی اطلاعات دوزبانه، پیکره‌های زبانی، معادل‌های واژگانی، ترجمه خودکار، عامل‌های هوشمند

1. aminnezarat@gmail.com
*mosavit@pnu.ac.ir

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

۱. مقدمه

طی سالیان اخیر، گسترش استفاده از رایانه و در پی آن افزایش قابل توجه و چشمگیر منابع و مطالب علمی و غیرعلمی به زبان‌های مختلف از جمله فارسی در فضای اینترنت و سایبر، باعث توجه بیشتر محققان و نیز عموم مردم به استخراج اطلاعات و داده‌های مورد نیاز خود از محیط مجازی شده است.

با توجه به این موضوع، شاخه علوم رایانه‌ای از علم زبان‌شناسی که در جهت تسهیل بازشناسی و بازیابی اطلاعات و دانش از داده‌های موجود در رایانه‌هاست، رشد چشمگیری یافته است. درواقع، نیاز مردم به جستجو و یافتن اطلاعات از محیط سایبر با استفاده از نرم‌افزارهای رایانه‌ای شاخه جدید و میان‌رشته‌ایی به‌نام "بازیابی اطلاعات"[1] را به‌وجود آورده که مورد توجه متخصصان فنّاوری اطلاعات و زبان‌شناسی رایانه‌ای قرارگرفته است.

بازیابی اطلاعات بین زبانی به فرایندی گفته می‌شود که طی آن یک کاربر، جستاری (یک واژه، عبارت، یا حتی جمله‌ای) را به یک زبان جستجو می‌کند درحالی‌که انتظار دارد نتایج جستجوی خود را به زبان دیگری دریافت نماید. از آنجا که کاربران برای یافتن مطالب مورد نظر خود به‌طور معمول، توانایی لازم در ترجمه عبارات فارسی به انگلیسی را ندارند و یا حتی معادل‌های فارسی یک علم در زبان انگلیسی را نمی‌شناسند، امکان دسترسی به منابع موجود را از دست می‌دهند.

۲. بیان مسأله و ضرورت پژوهش

یکی از مشکلات عمده کاربران فارسی‌زبان در استفاده از منابع موجود در فضای سایبر، عدم امکان بازیابی موضوعات مورد نظر است که این مسأله تا حد زیادی به حجم کم اطلاعات به زبان فارسی در این فضا برمی‌گردد. استفاده از فرهنگ لغت[2] نیز به‌دلیل عدم توانایی در ارائه پاسخ مناسب به ترکیبات چندتایی[3] رایج در زبان‌ها کمتر در این زمینه راه‌گشاست. با بررسی کلیدواژه‌های[4] جستجوشده به زبان فارسی در سایت Google ملاحظه می‌شود که به‌طور معمول، کاربران نتوانسته‌اند معادل‌های مناسبی برای یافتن مطالب مورد نظر خود انتخاب و ترجمه نمایند. درواقع مشکل اصلی، عدم آگاهی کاربران از معادل مناسب این ترکیبات در زبان فارسی است. به‌عنوان مثال، با جستجوی عبارت فارسی "دوره‌های مقدماتی" در سایت مترجم Google، عبارت "foundation courses" به‌عنوان معادل ارائه می‌شود که این عبارت (که معادل صحیح عبارت فارسی مورد نظر نیست) با استفاده از فرهنگ لغت و بدون در نظر گرفتن نوع متن مورد

1. Information Retrieval 2. Dictionary 3. Chunks 4. Keywords

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

نظر کاربر ترجمه شده است. در زبان فارسی، تعداد ترکیبات پرکاربرد واژه‌ها فراوان است و کمتر فرهنگ لغتی در این زمینه می‌تواند به مترجم کمک نماید.

بنابراین، ساخت بانک اطلاعاتی و پیکره‌ای[1] که حاوی ترجمه دقیق‌تری از ترکیبات چندتایی زبان فارسی به انگلیسی باشد برای استفاده در موتورهای بازیابی اطلاعات بین زبانی ضروری است؛ ایجاد یک روش جدید در دسته‌بندی متون و معادل‌سازی عبارات ترجمه‌ای به یکدیگر. یک بانک اطلاعاتی شامل همه ترکیبات چندتایی فارسی و معادل انگلیسی آنهاست که می‌تواند ضمن استفاده در نرم‌افزارهای جستجوی بین زبانی، به‌عنوان یک بانک اطلاعاتی جامع و کامل در موتورهای ترجمه ماشینی استفاده شود.

۳. سؤالات پژوهش

همان‌گونه که در بیان مسأله گفته شد، هدف این پژوهش پیاده‌سازی یک سامانه بازیابی بین زبانی فارسی – انگلیسی است که بتواند ابهامات ترجمه‌ای در عبارات چندتایی را حل کند. از این رو، در این پژوهش سعی خواهد شد به سؤالات زیر پاسخ داده شود:

۱) آیا می‌توان سامانه بازیابی اطلاعات بین زبانی را فقط با کمک فرهنگ لغت پیاده‌سازی کرد؟

۲) آیا می‌توان از عامل‌های هوشمند جمع‌آوری اطلاعات به‌منظور تشکیل یک پیکره دوزبانه استفاده کرد؟

۳) نقش پیکره در سامانه بازیابی اطلاعات بین زبانی تا چه اندازه است؟

٤. روش پژوهش

در این پژوهش که با هدف دستیابی به یک پیکره دوزبانه فارسی به انگلیسی صورت گرفته است، ابتدا با استفاده از نوعی نرم‌افزار هوشمند به نام "عامل" اقدام به جمع‌آوری متون و واژه‌های مختلف از وب‌سایت‌های متعدد شد. درادامه، داده‌های جمع‌آوری‌شده با استفاده از فنون دسته‌بندی در گروه‌های مختلف قرار داده شدند.

گروه‌های به‌دست‌آمده ضمن تفکیک موضوعی دارای خواص متعددی هستند که قابل استفاده در نرم‌افزارهای ترجمه ماشینی هم است. پژوهش حاضر از نوع توسعه‌ای و کاربردی است. در این پژوهش، ضمن تهیه یک نرم‌افزار آزمایشگاهی، نمونه‌های واقعی به‌عنوان پارامتر محاسبه شد و صحت الگوریتم و مدل ساخته‌شده مورد آزمایش قرار گرفت.

---

1. Corpora

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

مبنای اصلی این پژوهش تحلیلی – تجربی است که بر پایه پیاده‌سازی نمونه آزمایشگاهی مدل ابداعی قرار دارد.

۵. پژوهش‌های گذشته

بازیابی اطلاعات یعنی یافتن اسنادی که محتوای آنها مرتبط با جستار مورد نیاز کاربر باشد و یکی از شاخه‌های مهم پردازش زبان طبیعی به‌حساب می‌آید. بنابراین، مرتبط بودن از مسائل مهم دسته‌بندی در رشته بازیابی اطلاعات است. تاکنون پژوهشگران زیادی با به‌کارگیری روش‌های مختلف سعی در ارائه سامانه‌هایی با کارایی بالا نموده‌اند، اما بیشتر پژوهش‌هایی که تاکنون در زمینه بازیابی اطلاعات انجام شده است براساس واژه‌نامه‌های دوزبانه بوده است.

پژوهش‌های غیرفارسی در حوزه بازیابی اطلاعات بسیار زیاد است و در اینجا برای نمونه به مواردی اشاره می‌شود. یک مطالعه در این زمینه با رویکرد واژه‌نامه نشان داد که ترجمه واژه به واژه عبارت‌های جستجو می‌تواند منجر به کاهش بین 40 ٪ تا 60 ٪ کارآمدی بازیابی در مقایسه با بازیابی اطلاعات یک‌زبانه براساس همان عبارت‌های جستجو گردد. در این پژوهش که روی هر دو زبان فرانسه و انگلیسی انجام شد، پژوهشگران با استفاده از محاسبه متوسط دقت بازیافت دریافتند که ترجمه عبارتی در مقایسه با ترجمه واژه به واژه عبارت‌های جستجو، منجر به نتیجه بهتری می‌شود (Hull and Grefenstette 1996).

چن نیز با انجام پژوهشی به بررسی کارآمدی ترجمه عبارتی در بازیافت اطلاعات بین زبانی انگلیسی – چینی پرداخت. وی که با استفاده از روش پژوهش ارزیابی برنامه، این پژوهش را انجام داد، ترجمه عبارتی را موفق‌تر از ترجمه واژه به واژه دانست. یافته‌های او نشان داد که در مقایسه با میانگین دقت بازیافت در بازیابی یک‌زبانه، ترجمه عبارتی به 53 ٪ کارآمدی دست می‌یابد. این در حالی است که این میزان برای ترجمه واژه به واژه به 42 ٪ رسید. وی اشاره می‌کند که با بهره‌گیری از منابع اضافی و کامل‌تر برای ترجمه عبارت‌ها می‌توان این میزان از کارآمدی را افزایش داد و به 83 ٪ کارآمدی به‌دست‌آمده برای بازیابی یک‌زبانه رسید (Chen 2002).

ازجمله کارهایی که در مورد بازیابی اطلاعات بین زبانی برای زبان فارسی انجام شده است مطالعه علیزاده و همکارانش است که در آن، سامانه‌ای را مورد ارزیابی قرار دادند که فقط از واژه‌نامه‌های دوزبانه ماشین‌خوان استفاده می‌نمود. از جمله یافته‌های پژوهش آنها می‌توان به این موارد اشاره نمود: ۱) کارآمدی بیشتر استفاده از رویکرد ترجمه "اولین برابر نهاده" در مقایسه با رویکرد "همه برابر نهاده‌ها"، ۲) پردازش صرفی[1] واژه‌های عبارت جستجوی فارسی

1. Morphologic

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

پیش از ترجمه آنها در مقایسه با عدم انجام این پردازش، و ۳) کارآمدی بیشتر استفاده از شیوه ترجمه عبارتی در مقایسه با ترجمه واژه به واژه در هنگام ترجمه عبارت جستجوهای فارسی (علیزاده و همکاران ۱۳۸۸).

در پژوهشی دیگر در زمینه بازیافت اطلاعات بین زبانی برای هر دو زبان فارسی و انگلیسی، فقط از پیکره دوزبانه برای استخراج معادل‌های مناسب برای واژه‌ها استفاده شده است. این طرح، درواقع کوششی است در جهت بهبود روش ارائه‌شده در آن پژوهش با اضافه نمودن پیکره یک‌زبانه تا بتوان از عهده معادل‌یابی برای ترکیبات نیز برآمد (Mosavi Miangah 2008).

با توجه به پژوهش‌های انجام‌شده در این راستا، روشن است که هیچ سامانه بازیافت اطلاعات دوزبانه‌ای بدون دسترسی به بانک داده‌ای غنی‌تری نسبت به واژه‌نامه‌ها که فقط واژه‌های مجزا را در خود جای می‌دهند، نمی‌تواند در جهت معادل‌یابی ترکیبات چندتایی کارآمد باشد. این بانک داده‌ای غنی چیزی نیست جز پیکره‌های یک‌زبانه و دوزبانه که در این طرح برای اولین بار از قابلیت‌های آنها استفاده شده است.

٦. بازیابی اطلاعات

بازیابی اطلاعات به فنّاوری و دانش پیچیده جستجو و استخراج اطلاعات، داده‌ها، و فراداده‌ها در انواع گوناگون منابع اطلاعاتی مثل بانک اسناد، مجموعه‌ای از تصاویر، و وب گفته می‌شود (Mosavi Miangah 2008).

با افزایش روزافزون حجم اطلاعات ذخیره‌شده در منابع قابل دسترس و گوناگون، فرایند بازیابی و استخراج اطلاعات اهمیت ویژه‌ای یافته است. اطلاعات مورد نظر ممکن است شامل هر نوع منبعی مانند متن، تصویر، صوت، و ویدئو باشد. برخلاف پایگاه داده‌ها، اطلاعات ذخیره‌شده در منابع اطلاعاتی بزرگ مانند وب و زیرمجموعه‌های آن مانند شبکه‌های اجتماعی از ساختار مشخصی پیروی نمی‌کنند و در کل، دارای معانی تعریف‌شده و مشخصی نیستند.

امروزه، دنیا تغییر کرده است و میلیون‌ها نفر از مردم به‌صورت روزمره از طریق موتورهای جستجو از فنّ بازیابی اطلاعات استفاده می‌کنند. بازیابی اطلاعات علاوه بر این می‌تواند سایر مشکلات موجود در داده‌ها و اطلاعات که ناشی از ابهام در آنهاست را حل نماید. موضوع "ساختارنیافتگی داده‌ها" به عدم شفافیت در آنها برمی‌گردد و همچنین، عدم داشتن منطق واضح و ساختار مشخص کامپیوتری نیز از معضلات دیگر در بازیابی اطلاعات است.

بازیابی اطلاعات، همچنین به‌منظور ساماندهی و ساختاردهی مجدد اطلاعات بازیابی‌شده به‌کار برده می‌شود. این کار شامل دسته‌بندی موضوعی اطلاعات به‌دست‌آمده براساس محتویات هر متن یا سندی است.

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

موضوعی که در اینجا مطرح می‌شود بحث دسته‌بندی[1] است. این رویکرد بدین صورت است که با توجه به اسناد جدید بازیابی‌شده، ابتدا دسته‌هایی شناسایی می‌گردد و سپس، اسناد جدید پس از پردازش به‌صورت خودکار در موضوع مربوط قرار داده می‌شود.

در این رویکرد، سامانه به‌صورت خودکار و با استفاده از روش‌های مختلفی می‌تواند موضوع هر یک از متون را تشخیص دهد و آن سند را در گروه یا دسته متناسب جاگذاری نماید. یکی از شاخه‌های کاربردی در بازیابی اطلاعات، موضوع استخراج متن از داده‌ها[2] است که در سال‌های اخیر پیشرفت‌های قابل توجهی در آن صورت گرفته است.

هدف استخراج متن، استخراج و به‌کارگیری اطلاعاتی است که در مدارک متنی وجود دارد. این فرآیند با روش‌های مختلفی انجام می‌پذیرد که عبارتند از: جستجو و کشف الگوهای درون داده‌ها، پیدا کردن روابط میان بخش‌های داده‌ها، کشف قوانین پیش‌بینی‌کننده، و شناسایی کلمات چندتایی[3] (Lewis and Ringuette 1994).

۷. ساخت و جمع‌آوری بانک اطلاعاتی

بیشتر سامانه‌های بازیابی اطلاعات مبتنی بر بانک‌های اطلاعاتی رابطه‌ای هستند. به‌طور معمول این سامانه‌ها، از مجموعه‌ای از متون ساختارنیافته بازیابی می‌شوند که به‌عنوان داده‌های بدون مشخصه نامیده می‌شوند.

منابع اصلی جهت جمع‌آوری داده‌های ساختاریافته به‌طور عمده شامل کتابخانه‌های دیجیتال، بانک اختراعات، و وبلاگ‌های تخصصی است که به‌صورت معمول دسته‌بندی‌شده و مشخص هستند.

بنابراین، در این پژوهش نیز همان‌گونه که در بخش‌های بعد توضیح داده می‌شود جهت تکمیل پیکره‌های مورد نیاز از همین منابع استفاده شده است. یک مثال از این منابع، پایگاه‌های نمایه‌گذاری پایان‌نامه‌ها و مقالات دانشجویی است؛ این پایگاه‌ها به‌طور عمده دوزبانه هستند.

۸. عامل‌های هوشمند

یک عامل می‌تواند یک شخص، یک ماشین، یک قطعه کد نرم‌افزاری و یا هر چیز دیگری باشد و تعریف لغت‌نامه‌ای آن عبارت است از هر چیزی که توانایی عمل داشته باشد.

عامل، یک سامانه نرم‌افزاری است دارای ویژگی‌های درک محیط و ایجاد تغییر در آن، خودمختاری، تطبیق‌پذیری (واکنش به تغییرات محیط یادگیری)، و اجتماعی بودن.

یک عامل در هر لحظه "اطلاعات قبلی از محیط"، "تجربه‌های قبلی که می‌تواند از آنها

1. Classification 2. Text Mining 3. Chunk

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

یادگیری را انجام دهد"، "هدفی که باید برای رسیدن به آن تلاش کند"، و "اطلاعات و مشاهداتی از خود و محیط اطراف" را داراست و عملی را انجام می‌دهد.

پژوهگشران هوش مصنوعی معتقدند عامل، یک سامانه کامپیوتری است که علاوه بر ویژگی‌های کلی اشاره‌شده، برخی خصوصیات انسانی مانند دانش، اعتقاد، اراده، و تعهد را نیز داراست. در این دیدگاه، عامل دارای ویژگی‌های اضافی زیر است :

جابجایی: عامل می‌تواند در یک شبکه الکترونیکی تغییر مکان دهد؛

صداقت: عامل به‌طور عمده، اطلاعات نادرست منتقل نمی‌کند؛ و

خیرخواهی: اهداف عامل‌ها با یکدیگر در تضاد نیستند و هر عامل در تلاش است تا فقط، وظیفه محول‌شده خودش را به انجام برساند.

عقلانیت : یک عامل در راستای تحقق هدفش، رفتار می‌کند (Mohammadian 2004).

عامل از دو بخش برنامه و معماری تشکیل شده است. برنامه، تابعی است که رفتار عامل را پیاده‌سازی می‌کند و به‌عبارت دیگر، عمل نگاشت از ادراکات عامل به یک رفتار خاص را برعهده دارد. وظیفه هوش مصنوعی طراحی برنامه عامل است. سخت‌افزارهای محاسباتی که برنامه عامل بر روی آنها اجرا می‌شود، معماری عامل نام دارند. معماری می‌تواند یک کامپیوتر ساده باشد و یا اینکه در عین حال، تجهیزات سخت‌افزاری خاص مانند ابزارهای پردازش صوت و تصویر را هم شامل شود. همچنین، سخت‌افزار باید شامل نرم‌افزاری باشد که میان کامپیوتر و برنامه عامل قرار گرفته و امکان برنامه‌نویسی سطح بالا را فراهم نماید.

معماری از طریق حسگرها مشاهدات را در اختیار عامل قرار می‌دهد، برنامه را اجرا می‌کند، و رفتار عامل را از طریق اندام مجری به محیط و سایر عامل‌ها منتقل می‌کند (Luck and Padgham 2008).

در این پژوهش، به‌منظور تکمیل بانک اطلاعاتی مورد نیاز در نرم‌افزار بازیابی اطلاعات، یک عامل هوشمند نرم‌افزاری تهیه شده است که می‌تواند ضمن جابجایی در محیط وب، اقدام به جمع‌آوری متون و داده‌های مورد نیاز نماید. در بخش بعد، ساختار اولیه این عامل هوشمند بررسی می‌شود.

۹. عامل هوشمند طراحی‌شده برای این پژوهش

به‌منظور تعیین نقطه شروع حرکت عامل نرم‌افزاری، سایت روزنامه همشهری به‌عنوان مبداء ورودی درنظر گرفته شد. با توجه به اینکه در ساختار طراحی‌شده برای پیکره فارسی، نوع متون نیز باید مشخص شوند، حرکت عامل هربار از یک دسته صورت گرفت (سیاسی، دین،

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

اندیشه، و...). پس از خوانش هر یک از متون موجود در صفحات وب مربوط به بخش تعیین‌شده، متن در پیکره وارد شد و نوع[1] جمله نیز مشخص گردید. از آنجا که هر یک از اخبار دارای خبرهای مرتبط دیگری نیز بودند، با استفاده از خوانش ساختار XML و HTML، صفحه وب مربوط وارد پیوند[2] جزء بعدی گردید و محتوای متنی آن جز نیز به پیکره افزوده شد و این عمل به‌همین صورت تا خوانش کامل کل سایت ادامه پیدا کرد.

پس از اتمام کار این بخش و مشاهده تجربی متون بازیابی‌شده، ایرادات عامل نرم‌افزاری مشخص و رفع گردید. لازم به توضیح است که این عامل همانند یک روبوت[3] عمل می‌کند و پس از رهاسازی در محیط وب، اقدام به جمع‌آوری داده‌های مورد نیاز و اسناد بازیابی‌شده را به پیکره متنی ارسال می‌کند. درادامه به‌منظور افزایش حجم متون پیکره متنی، فهرستی از سایت‌های معتبر با نحوه نگارش متون به‌نسبت رسمی، انتخاب شد و به‌عنوان "فهرست هادی"[4] به عامل نرم‌افزاری داده شد. سپس، با استفاده از دستور GO فرمان حرکت به عامل صادر شد و نرم‌افزاری طبق برنامه از قبل تعریف‌شده اقدام به خوانش و ثبت و حرکت به صفحه وب و وب‌سایت بعدی نمود. با توجه به برنامه‌نویسی انجام‌شده در عامل نرم‌افزاری، این‌کار به‌صورت تمام‌خودکار و بدون دخالت دست صورت گرفت.

انواع متون دسته‌بندی‌شده در پیکره متنی فارسی در گروه‌های زیر قرار گرفتند :

- سیاست - پزشکی – ادبیات - ورزشی - هنری - دینی
- علمی - حوادث - اجتماعی - اقتصادی - سایر

در حین خواندن صفحات وب، حجم زیادی از داده‌های غیرمرتبط از قبیل عکس‌ها، جداول، و پیوند به سایر صفحات یافت شد که پس از تشخیص نوع آنها اقدام به حذف داده‌ها گردید. بدین منظور، ابتدا ساختار خوانده‌شده در قالب استاندارد XML با تگ‌های[5] مشخص‌شده درآورده شد و پس از درج در پیکره متنی، تگ‌های اصلی استفاده‌شده در ساختار XML متن به‌صورت جدول ۱ درآمد.

جدول ۱. فهرست فیلدهای پیکره

شماره جمله	ID
متن	Text
نوع متن	Type
آدرس صفحه	Link
زمان و ساعت	Date Time

1. Type 2. Link 3. Robot 4. Driving list 5. Tag

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

با توجه به حجم زیاد اطلاعات جمع‌آوری‌شده به‌منظور افزایش سرعت در بازیابی اطلاعات پیکره، بانک اطلاعاتی عامل نرم‌افزاری SQL Server 2005 انتخاب شد. دلیل انتخاب این سرویس‌دهنده بانک اطلاعاتی، قابلیت بالای آن در پردازش پرسش‌ها و سؤال‌های SQL[1] است. علاوه بر این، قابلیت سازگاری این بانک اطلاعاتی با بیشتر زبان‌های برنامه‌نویسی از جمله C# و همچنین، ساختار ذخیره‌سازی بانک اطلاعاتی بر روی هارد کامپیوتر که به‌صورت خوشه‌بندی‌شده[2] است، از دیگر دلایل انتخاب آن بوده است.

۱۰. پیکره‌های متنی مورد نیاز

با توجه به اینکه روش این پژوهش مبتنی بر استفاده از پیکره‌های متنی است، لازم است که ابتدا نسبت به ساخت آنها اقدام شود. از آنجا که هدف اصلی این پژوهش استخراج ترکیبات چندتایی رایج در زبان فارسی و انگلیسی است، لازم است که پیکره‌های متنی تک‌زبانه فارسی و انگلیسی ساخته و استفاده شوند. در پژوهش دیگری که توسط موسوی میانگاه صورت گرفته است، یک پیکره متنی فارسی با بیش از ۲۶۴۰۰۰۰ جمله و ۱۴۹۰۰۰۰۰۰ لغت تهیه شده است (موسوی میانگاه ۱۳۸۸) که در این پژوهش با استفاده از روش‌های نرم‌افزاری و با کمک عامل‌های هوشمند[3] نرم‌افزاری اقدام به تکمیل و گسترش پیکره متنی تک‌زبانه و دوزبانه شد. درنتیجه، قبل از اینکه روش استفاده از این عامل‌های هوشمند در پژوهش توضیح داده شود، مقدمه‌ای درباره عامل‌ها بیان می‌شود و سپس، ساختار نرم‌افزار پیاده‌سازی‌شده در جمع‌آوری و تکمیل پیکره متنی تک‌زبانه توضیح داده خواهد شد.

۱۰-۱. پیکره متنی دوزبانه انگلیسی- فارسی

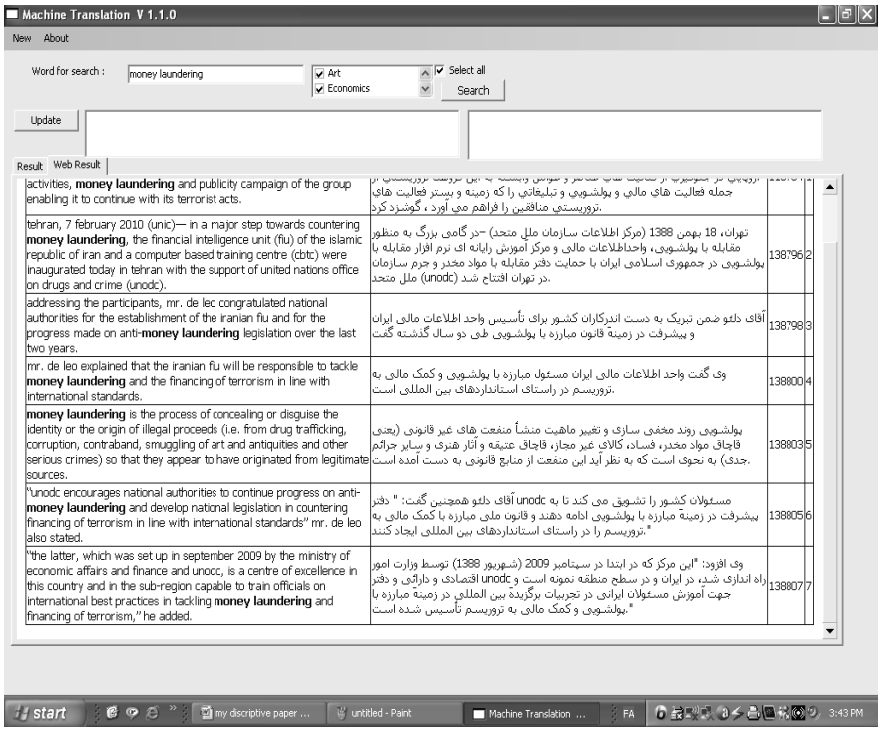
پیکره دیگری که در این پژوهش مورد نیاز است، پیکره دوزبانه فارسی – انگلیسی است که در آن متون ترجمه‌شده و معادل‌سازی‌شده فارسی به انگلیسی درج شده باشد، علاوه بر آن، نوع هر یک از متون نیز مشخص شود. با توجه به اینکه استخراج متون معادل‌سازی‌شده با استفاده از نرم‌افزار به‌صورت خودکار، خارج از محدوده این پژوهش است، این کار به‌صورت دستی و با جمع‌آوری متون فارسی و انگلیسی ترجمه‌شده صورت گرفت.

به‌منظور تسهیل و جلوگیری از بروز خطای انسانی، نرم‌افزاری جهت ثبت جملات ترجمه‌شده و انتخاب نوع جمله تهیه و برنامه‌نویسی شد. با استفاده از این نرم‌افزار کاربر می‌تواند هر دو جمله فارسی و انگلیسی را در دو بخش مجزا وارد کند و نرم‌افزار اقدام به ثبت آن در بانک اطلاعاتی می‌نماید. با توجه به اینکه در یک پژوهش دیگر توسط موسوی میانگاه، یک

1. Queries 2. Cluster 3. Intelligent Agent

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

پیکره متنی دوزبانه با بیش از ۳۵۰۰۰۰۰ واژه تهیه شده است (Mosavi Miangah 2009)، پژوهش جاری اقدام به گسترش همان پیکره نموده و تعداد واژه‌های آن را به بیش از ۴۵۰۰۰۰۰ کلمه رسانده و تنوع انواع[1] جملات را مطابق پیکره تک‌زبانه گسترش داده است. شکل ۱ نمایشی از تطابق[2] دوزبانه انگلیسی-فارسی است.



شکل ۱. رکوردهای تولیدشده توسط پیکره برای جستار"money laundering"

۱۰-۲. پیکره یک‌زبانه انگلیسی

از آنجا که برای شناخت و تشخیص عبارات چندتایی در زبان انگلیسی نیاز به یک پیکره متنی در زبان انگلیسی وجود دارد، پس از بررسی همه پیکره‌ها، پیکره Word Net که یکی از بزرگترین و بهترین پیکره‌هاست، انتخاب گردید (Miller et al. 1993). در این پیکره، علاوه بر وجود تعداد بسیار زیاد کلمات انگلیسی، نوع و معادل آنها نیز وجود داشته و به‌عنوان یک واژه‌نامه نیز قابل استفاده است.

1. Types 2. Concordance

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

۱۱. استخراج خودکار چندتایی‌ها (Chunk)

به‌منظور استخراج همه چندتایی‌ها در زبان‌های مبداء و مقصد به این صورت عمل می‌شود: ابتدا، کل ساختار یک جمله را استخراج و سپس، مقدار حرکت به جلو (g) مشخص می‌شود. مقدار حرکت به جلو و عددی است که بیشترین تعداد کلماتی را که می‌تواند در یک چندتایی معتبر وجود داشته باشد، مشخص می‌کند. در این طرح، ۴ = g درنظر گرفته شد که انتخاب عدد ۴ برمبنای تجربه زبان‌شناسی صورت گرفت. آنگاه با استفاده از یک نرم‌افزار کامپیوتری که با زبان برنامه‌نویسی C# نوشته شده است، اقدام به شناسایی تمامی ترکیبات ممکن از چندتای‌های مستتر در جمله شد.

۱۲. تعیین چندتایی‌های معتبر

به‌منظور تعیین چندتایی‌های معتبر و حذف داده‌های اضافه اقدام به دسته‌بندی هر یک از چندتایی‌ها شد. برای محاسبه این عدد همان‌گونه که پیشتر در بحث دسته‌بندی متون (میزان وابستگی) توضیح داده شد، اگر بتوان میزان وابستگی و با هم‌آیی هر یک از عبارات را محاسبه کرد، می‌توان نسبت به حذف یا معتبر بودن چندتایی‌های به‌دست‌آمده تصمیم‌گیری نمود.

برای این هدف می‌توان از روش X^2 که در بخش‌های پیشین توضیح داده شد، استفاده نمود. انتخاب این روش بدین دلیل صورت گرفت که با استفاده از فرمول X^2 (d,c) می‌توان میزان وابستگی دو عبارت c, d را در یک پیکره تعیین و درجه آن را مشخص کرد. با دقت در فرمول می‌توان دریافت که در این روش دو عبارت c, d در تمامی جملات یک پیکره بررسی و میزان فراوانی ترکیبات مختلف وقوع و یا عدم وقوع هر یک یا ترکیبی از هر کدام محاسبه می‌گردد، سپس با استفاده از روش X^2 میزان وابستگی هر کدام به‌هم به‌دست می‌آید (Mosavi Miangah, and Nezarat 2010).

جمله زیر را فرض کنید:

"کلاس‌های شما یک روز در میان تشکیل می‌شود."

پس از محاسبه مقدار AS برای هر یک از ترکیبات چندتایی در این جمله، جدول ۲ به‌دست آمد.

جدول ۲. محاسبه مقادیر AS برای همه عبارات ممکن

ترکیب	X2	ترکیب	X^2	ترکیب	X^2
کلاس‌های شما	۶.۰۲	شما یک روز در	۶.۰۰۳۴	روز در میان	۵.۲۴۱
کلاس‌های شما یک	۱۰.۲۶۸	یک روز	۳.۴۵۳	روز در میان برگزار	۲۳۴.۱۱
کلاس‌های شما یک روز	۲۵۴.۱۱	یک روز در	۸.۰۰۳۲	در میان	۷.۵
شما یک	۱۲.۰۰۳	یک روز در میان	۱.۲۵	در میان برگزار	۱۳۹.۴۳
شما یک روز	۷.۰۰۲۴	روز در	۱۳.۴۳۱	در میان برگزار می‌شود	۲۹۰.۴۴

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

حال به‌منظور انتخاب ترکیبات مناسب، نیاز به یک مقدار حد آستانه (جدول ۳) وجود دارد که مقدار آستانه ۶۳/۶ پس از محاسبه مقادیر مختلف توسط پژوهشگر انتخاب گردید.

جدول ۳. حد آستانه

P	(Critical value) X^2
1.0	71.2
05.0	84.3
01.0	63.6
005.0	88.7
001.0	83.10

اگر مقدار X^2 یا AS محاسبه‌شده برای هر یک از ترکیب‌ها کوچکتر از مقدار حد آستانه ۶۳/۶ باشد، بدین معنی است که می‌توان وابستگی بین کلمات آن ترکیب را قبول کرد و مقادیر بالاتر از حد آستانه را رد نمود.

این روال برای شناسایی همه چندتایی‌ها در پیکره فارسی ادامه داده می‌شود و یک بانک اطلاعاتی جدید از چندتایی‌های فارسی تشکیل می‌گردد. سپس، به‌منظور تعیین معادل انگلیسی هر یک از چندتایی‌های به‌دست‌آمده با استفاده از پیکره دوزبانه فارسی – انگلیسی اقدام به یافتن تمامی رکوردهایی که مقدار چندتایی‌های مورد نظر را دارند، می‌شود. به‌عنوان مثال، ترکیب چندتایی "یک روز در میان"[1] را در نظر گرفته می‌شود و از پیکره دوزبانه رکوردهای با شرط گفته‌شده استخراج می‌گردند و برای هر یک از رکوردهای به‌دست‌آمده، الگوریتم معادل‌سازی زیر اعمال و نتیجه محاسبه می‌شود:

۱) تمامی ترکیب‌های از دو تا چهارتایی جمله متناظر انگلیسی استخراج می‌شود؛

۲) مقدار فرمول X^2 یا AS برای هر یک از این ترکیبات با استفاده از پیکره تک‌زبانه انگلیسی محاسبه می‌شود؛

۳) ترکیب‌هایی با مقدار AS کمتر از ۶۳/۶ نگه داشته و مابقی حذف می‌شوند؛

۴) سپس، با استفاده از ترکیب‌های به‌دست‌آمده فارسی در بخش قبل و ترکیب‌های انگلیسی به‌دست‌آمده در مرحله قبل همین الگوریتم جدولی مانند جدول ۴ برای تمامی ترکیبات مختلف ساخته می‌شود؛ و

1. Every other day

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

۵) مقدار AS مرتبط با ترکیب‌های جدول محاسبه و مقادیر کمتر از حد آستانه نگه داشته می‌شود. ملاحظه می‌گردد که بیشترین مقدار AS متعلق به ترکیب مورد نظر، یعنی عبارت Every other day است.

جدول ٤. محاسبه مقدار AS ترکیب‌های معدل

ترکیب فارسی	ترکیب انگلیسی	AS	ترکیب فارسی	ترکیب انگلیسی	AS
یک روز در میان	The	54.141	یک روز در میان	.....	
یک روز در میان	The committee	002.121	یک روز در میان	.....	
یک روز در میان	The committee convenes	22.240	یک روز در میان	.....	
یک روز در میان	The committee convenes every	99.233	یک روز در میان	.....	
یک روز در میان	committee	..	یک روز در میان	.....	
یک روز در میان	committee convenes	..	یک روز در میان	.....	
یک روز در میان	committee convenes every	..	یک روز در میان	Every other	031.7
یک روز در میان	committee convenes every other	..	یک روز در میان	Every other day	234.1

۱۳. پاسخ به سؤالات پژوهش

با توجه به آنچه که گذشت و نمونه آزمایشگاهی پیاده‌سازی‌شده، در پاسخ به سؤال اول باید گفت که از فرهنگ لغت در ترجمه عبارات چندتایی در سامانه‌های بازیابی اطلاعات بین زبانی می‌توان فقط به‌منظور یک ابزار صحه‌گذاری استفاده کرد. در این پژوهش، بدون استفاده از فرهنگ لغت، ترجمه بیشتر لغات با استفاده از روش‌های دسته‌بندی متون و الگوریتم پیشنهادی پژوهش استخراج شد. در مورد پاسخ به پرسش دوم باید عنوان کرد که به‌دلیل قابلیت جابجایی عامل‌های هوشمند و مستقل بودن از بستر اجرایی، می‌توان با کمک یک عامل هوشمند و جابجایی در وب‌سایت‌های مختلف، همانند یک روبوت اینترنتی اقدام به تکمیل یک پیکره از داده‌های متنی آن وب‌سایت‌ها نمود.

در پاسخ به سؤال سوم که استفاده از پیکره‌های چندزبانه برای کاربرد در بازیابی اطلاعات بین زبانی مورد توجه قرار گرفته است، می‌توان گفت که به‌دلیل پیچیدگی عبارات در زبان فارسی، لزوم به‌کار بردن یک بانک اطلاعاتی مستقل در کنار فرهنگ لغت الزامی است. این بانک اطلاعاتی درواقع، همان پیکره زبانی است که شامل جملات و ترجمه‌های معادل آنها در زبان مقصد است که پس از ترجمه کلمات با استفاده از فرهنگ لغت می‌توان از پیکره به‌عنوان ابزار صحه‌گذاری ترجمه استفاده کرد.

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

۱۴. نتیجه‌گیری

در این پژوهش سعی شده است که علاوه بر تکمیل پیکره‌های متنی ساخته‌شده در طرح‌های پیشین (یک‌زبانه و دوزبانه)، یک روش آماری دسته‌بندی جهت تعیین میزان وابستگی عبارات به یکدیگر ارائه شود و با استفاده از آن اقدام به رفع ابهام در ترجمه عبارات چندتایی گردد.

همان‌گونه که در گزارش مطرح شد، یکی از مشکلات عمده در بازیابی اطلاعات بین زبانی، عدم تسلط کاربران به ترجمه دقیق ترکیبات چندتایی و به‌تبع آن بروز مشکل در فرآیند بازیابی اطلاعات و بازیافت اطلاعات و متون غیرمرتبط است. عدم وجود این ترکیبات چندتایی در لغت‌نامه‌ها نیز بر این مشکل افزوده است. در این پژوهش، علاوه بر تکمیل پیکره‌های متنی یک و دوزبانه، بانک اطلاعاتی‌ای از همه ترکیبات چندتایی و ترجمه آنها براساس اطلاعات موجود در پیکره‌ها ساخته شد.

در این پژوهش، با استفاده از پیکره‌های یک‌زبانه و دوزبانه فارسی و انگلیسی و ارائه یک روش دسته‌بندی متون و مقایسه ترجمه حاصل‌شده از این روش، برای عبارات چندتایی با ترجمه‌های به‌دست‌آمده از نرم‌افزار ترجمه ماشینی گوگل مشاهده شد که استفاده از پیکره در مقایسه با فرهنگ لغت (استفاده‌شده در نرم‌افزار ترجمه ماشینی گوگل) می‌تواند در بهبود کیفیت ترجمه تأثیر بسزایی داشته باشد.

در این روش، با استفاده از فرمول X یا همان روش میزان وابستگی بین یک ترکیب چندتایی و ترجمه احتمالی آن در زبان دیگر براساس جملات موجود در پیکره‌ها، ترجمه کاندید مشخص می‌شود. انتخاب ترجمه کاندید نیز با استفاده از جدول حد آستانه یا نقطه بحرانی صورت گرفت.

یکی از دستاوردهای اجرایی این طرح، بالا بردن دقت سامانه‌های بازیابی اطلاعات در موتورهای جستجو است که با استفاده از پیکره و بانک اطلاعاتی، ترکیب‌بندی واژه‌ها قابل دسترس است. با توجه به اینکه در انتهای طرح، پیکره‌ای دوزبانه همراه با بانک اطلاعاتی بسیار غنی‌ای به‌وجود آمده است، می‌توان پیش‌بینی نمود که در طرح‌های مجزای دیگری با استفاده از دانش به‌دست‌آمده در این طرح، محصولاتی از قبیل واژه‌نامه مبتنی بر پیکره[1] و نیز سامانه حافظه ترجمه[2] را به‌عنوان فعالیتی جدید بتوان عرضه نمود.

در فعالیت آتی، پژوهشگران قصد دارند از پیکره تولیدشده به‌عنوان یک مرجع دانش استفاده کنند و یک فرهنگ لغت جدید ایجاد نمایند. بدین منظور، استفاده از روش‌های داده‌کاوی جهت تکمیل پیکره یک‌زبانه و دوزبانه پیشنهاد می‌شود. همچنین، به‌منظور افزایش

1. Corpus-based dictionary 2. Translation memory system

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

سرعت محاسبات می‌توان با تغییر در روش محاسبه AS، سرعت دسته‌بندی کلمات و معادل‌سازی آنها را افزایش داد.

۱۵. منابع

علیزاده، حمید و همکاران. ۱۳۸۸. بررسی کارآمدی روش‌های موجود در بازیابی اطلاعات بین زبانی فارسی - انگلیسی با استفاده از واژه‌نامه دوزبانه ماشین‌خوان. *فصلنامه علوم و فنّاوری اطلاعات ایران* ۲۵ (۱): ۵۳-۷۰.

موسوی میانگاه، طیبه. ۱۳۸۸. *نقش پیکره‌های بزرگ یک‌زبانه در بهبود کیفیت ترجمه ماشینی*. طرح پژوهشی، دانشگاه پیام نور، زمستان ۱۳۸۸.

Carlson, C. N. 2004. Information overload, retrieval strategies and Internet user empowerment. In *Proceedings of COSTA Action 269*, *Helsinki*, 169-176. http://www.citeulike.org/user/aosbat/article/1644417 (accessed 10 Dec. 2009).

Chen, H .H. 2002. Chinese information extraction techniques. Summer School of Intelligent Media and Information Processing (SSIMIP), Chapter 12, National University of Singapore. http://nlg3.csie.ntu.edu.tw/conference_papers/ pretrack.pdf (accessed 6 Dec. 2009).

Douglas, O. W., and B. J. Dorr. 1996. A survey of multilingual text retrieval. *Technical Report UMIACS-TR-96-19, Institute for Advanced Computer Studies,* University of Maryland, College Park , MD, USA. xxii, 522-528.

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*, *Conference on Research and Development in Information Retrieval*, *Zurich, Switzerland,* 49-57. Zurich: Assn for Computing Machinery.

Lewis, D., and M. Ringuette. 1994. A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR94 3rd Annual Symposium on Document Analysis and Information Retrieval*, Vol. 33, 81–93. Citeseer, Las Vegas, NV, IRSI, University of Nevada, Las Vegas.

Luck, M., and L. Padgham, (eds.). 2008. Agent oriented software engineering VIII*: The 8th International Workshop on Agent Oriented Software Engineering, AOSE 2007*, Honolulu, HI, May 14, Revised Selected Papers (LNCS 4951). Berlin, Germany: Springer Verlag.

Miller, G., R. Beckwith, C. Fellbaum, D.Gross, and K. Miller. 1993. *Introduction to WordNet: an on-line lexical database. Journal of Lexicography* 3: 235-312.

Mohammadian, M. 2004. *Intelligent agents for data mining and information retrieval.* Hershey :Idea Group Publishing.

Mosavi Miangah, T. 2008. Automatic term extraction for cross-language information retrieval using a bilingual parallel corpus. In *Proceedings of the 6th International Conference on Informatics and Systems (INFOS2008), 27-29 March 2008,* 81-84. Cairo, Egypt: IEEE.

Mosavi Miangah, T. 2009. Constructing a large-scale English-Persian parallel corpus. *META* 54 (1): 181-188.

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

Mosavi Miangah, T., and A. Nezarat. 2010. A novel method for cross-language retrieval of chunks using monolingual and bilingual corpora. In *Proceedings of the International Conference on Advances in Information and Communication Technologies (ICT 2010)*, *ACEEE - Association of Computer, Electronics and Electrical Engineers,* 307–312. Cochin, India: IEEE.

Sihem, A. Y., and M. Lalmas. 2006. XML search: languages, INEX and scoring. SIGMOD record 35 (4): 16–23. DOI: doi.acm.org/10.1145/1228268.1228271.217, 519, 526 (accessed 4 Dec. 2009).

Downloaded from jipm.irandoc.ac.ir at 14:24 IRST on Friday November 16th 2018

Designing and Implementing a Cross-Language Information Retrieval System Using Linguistic Corpora

Amin Nezarat*
MS in IT, Islamic Azad University, Yazd Branch

Tayebeh Mosavi Miangah[1]
Associate Professor of Applied Linguistics, Payame Noor University, Yazd

Iranian Journal of Information Processing & Management

Iranian Research Institute
For Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in LISA, SCOPUS & ISC
special issue: on Information Storage, retrieval and Management (winter 2012)

Abstract: Information retrieval (IR) is a crucial area of natural language processing (NLP) and can be defined as finding documents whose content is relevant to the query need of a user. Cross-language information retrieval (CLIR) refers to a kind of information retrieval in which the language of the query and that of searched document are different. In fact, it is a retrieval process where the user presents queries in one language to retrieve documents in another language. This paper tried to construct a bilingual lexicon of parallel chunks of English and Persian from two very large monolingual corpora an English-Persian parallel corpus which could be directly applied to cross-language information retrieval tasks. For this purpose, a statistical measure known as Association Score (AS) was used to compute the association value between every two corresponding chunks in the corpus using a couple of complicated algorithms. Once the CLIR system was developed using this bilingual lexicon, an experiment was performed on a set of one hundred English and Persian phrases and collocations to see to what extend this system was effective in assisting the users find the most relevant and suitable equivalents of their queries in either language.

Keywords: Cross-language information retrieval, linguistic corpora, automated translation, intelligent factors

1. aminnezarat@gmail.com
*Corresponding author: mosavit@pnu.ac.ir