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

چکیده: بازیابی اطلاعات بین زبانی به فراوانی گونه می‌شود که طی آن یک کابری، جستجوی (یک واقع، عبارت، یا حتی جمله) را به یک زبان جستجو می‌کند. در حالیکه اکثریت جستجو در زبان یکی در زبان دیگر درست نمی‌شود. این امر از همان ابزار استفاده که می‌تواند در فضاهای سایر، عدم امکان بازیابی موضوعات مورد نظر است که این مسأله تا حد زیادی به حجم کم اطلاعات در زبان فارسی در این فضا برمی‌گردد. استفاده از فرهنگ لغت نیز بهدلیل عدم توالی در ارائه پاسخ مناسب به ترکیبات چندانی رایج در زبان‌ها کمتر در این زمینه راه‌اندازی شده است.

طرح‌ی حاضر که به هدف بایدن راهحل مناسب برای این مسئله با بهره‌برداری یک مدل تفسیری یا ککس الگوریتم‌های راهحل منطقی و مفروض به‌صرفه‌برای برای این مشکل ارائه نماید. به‌منظور آزمون کیفیت کار سامانه طراحی، در این طرح، آزمایش بر روی تعداد 100 ترکیب از زبان فارسی و انگلیسی انجام شد که بر اساس سلامتی یک اطلاعات برای این مجموعه از ترکیبات سپاه رضایت بخش بوده است. یکی از دستاوردهای اصلی این طرح، بالا بردن دقت سامانه‌های بازیابی اطلاعات در موتورهای جستجو است که با استفاده از پیکره و پانک اطلاعات، ترکیب‌بدن واقع‌کننده قابل دسترس است.

کلیدواژه‌ها: بازیابی اطلاعات دوزبانه، پیکره‌های زبانی، مدل‌های وارگانتی، ترجمه خودکار، فهم‌شناسی
1. مقدمه

طی سال‌های اخیر، گسترش استفاده از رایانه و در بی‌آن افراش قابل توجه و چشمگیر منابع و مطالب علمی و غیرعلمی به زبان‌های مختلف از جمله فارسی در فضاهای اینترنت و سابیر، باعث توجه بیشتر محققان و نیز عوام مدرم 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 تهیه بازسازی‌های اطلاعات و دانش از داده‌های موجود در رایانه‌های است، رشد چشمگیری یافته است. در این مقاله، نیاز مورد به جستجو و بافت اطلاعات از نظر سایر با استفاده از نرم‌افزارهای رایانه‌ای شاخه جدید و میان رشته‌ای به نام "بازسازی اطلاعات" را به وجود آورده که مورد توجه متخصصان فناوری اطلاعات و زبان‌شناسی رایانه‌ای قرار داشته است.

بازسازی اطلاعات بین زبانی به فرانسه گفته می‌شود که از آن یک کاربر، جستاری (یک واژه، عبارت، با حاکم جمله‌ای) را به یکی از زبان‌های جستجو می‌کند و در حالی که انتظار دارد نتایج جستجوی خود را به زبان دیگری دریافت نماید. از آنجا که کاربران برای یافتن مطالب مورد نظر خود به طور معمول، توaneousی لازم در ترجمه عبارات فارسی به انگلیسی را دارند و با هنگام خواستاری فارسی می‌باشند، می‌توان به یک علمی در زبان انگلیسی را نمی‌شناسند، امکان دسترسی به منابع موجود را به‌دست می‌دهند.

۲. بیان مسئله و ضرورت پژوهش

یکی از مشکلات عده‌ای کاربران فارسی زبان در استفاده از منابع موجود در فضای سایبر، عدم امکان بازسازی موضوعات مورد نظر است که این مسئله تا حد زیادی به حجم کم اطلاعات به زبان فارسی در این فضای برمی‌گردد. استفاده از فرهنگ لغت نیز به‌دلیل عدم توانایی در ارائه پاسخ مناسب به ترکیبات چندتایی رایج در زبانها که در این زمینه را گفتمان. با بررسی Mلاحظه می‌شود که طور معمول Google کلیدواژه‌های جستجوشده به زبان فارسی در سایت ملاحت می‌شود که به‌طور معمول، کاربران توانستند مطالب مورد نظر خود انتخاب و ترجمه نمایند.

در واقع مشکل اصلی، عدم آگاهی کاربران از معادل مناسب این ترکیبات در زبان فارسی است. Google به عنوان مثال، با یک جستجوی "عبارت فارسی" در سایت مترجم "دوره‌های مقدماتی" در سایت متترجم وب‌گاه "USB" به عنوان 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
نظر کاربر ترجمه شده است. در زبان فارسی، تعداد ترکیبات پرکاربرد وازههای فراوان است و
کمتر فرهنگی توسط مترجمان به متن فارسی ترجمه نمی‌شود.

با انگلیسی، ساخت ساختگی اطلاعاتی و پیکرهای که حاوی ترجمه دقیق‌تری از ترکیبات
چندتایی زبان فارسی به انگلیسی باشد برای استفاده در موتورهای پژوهش اطلاعاتی بین زبانی
ضروری است؛ ایجاد یک روش جدید در دستبندی منابع و معادل‌سازی عبارات ترجمه‌ای به
بکارگیری یک بانک اطلاعات شامل همه ترکیبات چندتایی فارسی و معادل انگلیسی آنهاست.
که می‌تواند ضمن استفاده در ترم‌افزارهای جستجو در زبان‌ها، به عنوان یک بانک اطلاعاتی
جامع و کامل در موتورهای ترجمه مانندین استفاده شود.

3. سوالات پژوهش

همان‌گونه که در بیان مساله گفته شد، هدف این پژوهش پیدا کردن
بین زبانی فارسی–انگلیسی است که بتواند ایهامات ترجمه‌ای در عبارات چندتایی را حل کند.
از این رو در این پژوهش سعی خواهیم داشت به سوالات زیر پاسخ داده شود:

(1) آیا در توان سامانه پژوهشی اطلاعات به درستی تفسیر و تحلیل
کرده‌ایم؟

(2) آیا در توان از عامل‌های تأثیرگذار جمع‌آوری اطلاعات به منظور
تشکیل یک پیکر به‌کار گرفته‌ایم؟

(3) نظر پژوهش در سامانه پژوهشی اطلاعات بین زبانی مانندین استفاده شده است؟

4. روش پژوهش

در این پژوهش که با هدف استخراج به‌کار گیری روش‌های فارسی به انگلیسی
گرفته شد، ابتدا با استفاده از نوع ترجمه‌های مانند "عامل اقدام به جمع آوری شده و
زاویه‌های مختلف از ویدیو-نیم‌های مختلف شده در چندین نوع استفاده از
فکنون دست‌رسیده در گروه‌های مختلف قرار داده شدند.

گروه‌های به‌دست آمده در هر نوع موضوع دارای خواص مختلفی هستند که
استفاده در ترجمه مانندین است پژوهش حاضر از نوع توصیعی و کاربردی
است. در این پژوهش، ضمین تهیه یک ترم‌افزار آزمایشگاهی نمونه‌های واقعی به عنوان واحدهای
محاسبه‌شده و صحت الگوریتم و مدل ساخته شده مرور آزمایش قرار گرفت.

1. Corpora
متنی اصلی این پژوهش تحلیلی - تجربی است که بر یا ایندیسی نمونه آزمایشگاهی
مدل ابداعی قرار دارد.

5. پژوهش‌های گذشته

با پژوهش‌های اندازه‌گیری افراد استادی که محیط آنها مرتب می‌باشد یا گزارش‌های زیان طبیعی به حساب می‌آید. بنابراین، مرتبی بودن از مسائل مهم دستبندی در رشته‌های اندازه‌گیری زیان‌ها با به کارگیری رویکرد متفاوت سعی بر ارتقای سامانه‌های این کاربری بالا نموده‌اند. اما بیشتر پژوهش‌هایی که تاکنون در زمینه بازیابی اندازه‌گیری انحراف شده است بسیاری و در انجام برای نمونه

پژوهش‌های طیف‌پذیر در حوزه بازیابی اطلاعات زیان‌های است و در انجام برای نمونه

به‌ماورادی‌اشتراک می‌شود. یک مطالعه در این زمینه با روش‌های سامانه‌های داد که تجربه وازه

به‌واژه‌عبارت‌های سمجتی می‌تواند منجر به کاهش بین 40٪ کارآمدی بازیابی در

مقایسه با بازیابی اندازه‌گیری با استفاده از مباحث متوسط‌دان

با دریافتند که تجربه‌ای در مقایسه با تجربه وازه به‌واژه‌عبارت‌های سمجتی، منجر

به‌نتیجه‌ی بهتری می‌شود (Hull and Grefenstette 1996).

1. Morphologic
پیش از ترجمه آنها در مقایسه با عدم انجام این پردازش، و (3) کار آمادی پیشتر استفاده از شیوه ترجمه عبارت در مقایسه با ترجمه وازه به وازه در هنگام ترجمه عبارت جستجوهای فارسی (علیزاده و همکاران 1388).

در پژوهشی دیگر در زمینه بایزیافت اطلاعات بین زبانی برای هر دو زبان فارسی و انگلیسی، فقط از پیکره دوزی‌های برای استخراج معادله‌ای مناسب برای وازه ای استفاده شده است. این طرح، در موقعیت کوئیشی است در جهت بهبود روش ارائه‌شده در آن پژوهش با اضافه‌نامدن پیکره یک‌زبانی با بی‌بندان و همچنین به عنوان از هر پایداری برای ترکیبات نیز برآمید (Mosavi Miangah 2008). می‌توگه به پژوهش‌های انجام‌شده در این است، روشی است که هیچ سامانه بازیافت اطلاعات دوزی‌های بدون دسترسی به بانک داده‌ای ذهنی نبوده و به اندازه‌ای که فقط وازه‌ها مجزا را در خود جای می‌دهند، نمی‌تواند در جهت معلوپذایی ترکیبات چندنابا کارآمد باشد. این بانک داده‌ای ذهنی جز پیکره‌های یک‌زبانی و دوزی‌های که در این طرح برای اولین بار از قابلیت‌های آنها استفاده شده است.

6. بایزیاف اطلاعات

پایگاه اطلاعات به فاکس و دانش پیچیده جستجو و استخراج اطلاعات، داده‌ها، و فرآیندها در انواع گوناگون منابع اطلاعاتی مثل بانک اطلاعاتی مجموعه‌ای از مجموعه‌های تصاویر و گفتگو می‌شود (Mosavi Miangah 2008). می‌توگه با افزایش ورودی‌وار و حجم اطلاعات ذخیره‌شده در منابع قابل دسترس و گوناگون، فرآیند بایزیاف و استخراج اطلاعات به‌هر‌یک‌داده‌ای ویژه‌ای یافته است. اطلاعات موردنظر ممکن است شامل هر نوع منبع مانند متن، تصویر، صوت، و ویدئو باشد. برخی از پایگاه‌های داده‌های آنالیزی در منابع اطلاعاتی بزرگ مانند وب و زیرمجموعه‌های آن مانند شبکه‌های اجتماعی از ساختار مشخصی پریوی کرده و در هر دیار معانی تعیین شده و مشخصی نیستند.

امروز، دنیا تغییر کرده است و میلیون‌ها فناوری نامیده به صورت روزمره از طریق موتورهای جستجو از فن بایزیاف اطلاعات استفاده می‌کنند. بایزیاف اطلاعات علاوه بر این می‌تواند سایر مشکلات موجود در داده‌ها و اطلاعات که ناشی از به‌روزرسانی در آنهاست را حل نماید. موضوع "ساختارنافذیگی داده‌ها" و بزرگان شفافیت در آنها بر می‌گردد و همچنین، عدم داشتن منطق واضح و ساختار مشخص کامپیوتری نیز از عضلات دیگر بایزیاف اطلاعات است.

بایزیاف اطلاعات، همچنین بمنظور ساماندهی و ساختاردهی مجدد اطلاعات بایزیاف شده به‌کار برده می‌شود. این کار شامل دست‌بنیت موضوعی اطلاعات به‌دست آمده براساس محیط‌های متنی با سندی است.
موضوعی که در اینجا مطرح می‌شود به دست‌بندی آمده است. این روش‌کردهای این‌طور صورت

است که با توجه به استاد جدید بازیابی شده، این‌دان دست‌بندی‌هایی شامل گردد و سپس، استاد

جدید پس از پردازش به صورت خودکار در موضوع مربوط قرار داده می‌شود.

در این روش‌کردهای بازیابی شده به صورت خودکار و با استفاده از روش‌های مختلفی می‌تواند

موضوع هر یک از متن‌ها تشخیص دهد و آن سند را در گروه‌ها با سخت‌سازی جامع‌تری

نمازید. یکی از شاخه‌های کاربردی در بازیابی اطلاعات، موضوع استخراج متن از داده‌ها است

که در سال‌های اخیر پیشرفت‌های قابل توجهی در آن صورت گرفته است.

هدف استخراج متن استخراج و با کارگیری اطلاعاتی است که در مدارک متین وجود

دارد. این روش‌های مختلف انجام می‌پذیرد که عبارتند از: جستجو و کشف الگوهاي

درون داده‌ها، پیدا کردن روابط میان بخش‌ها و داده‌ها، کشف قوانین به‌ینی کننده، و

شناسایی کلمات جدیدی (Lewis and Ringuette 1994).

<table>
<thead>
<tr>
<th>ساختمان و جمع‌آوری بانک اطلاعاتی</th>
</tr>
</thead>
</table>
| پیشرفت سامانه‌های بازیابی اطلاعاتی مبتنی بر بانک‌های اطلاعاتی رابطه‌ای هستند. به‌طور
| معقول این سامانه‌ها از مجموعه‌ای از متن‌ساختار‌های بازیابی می‌شوند که به عنوان داده‌های
| بدون مشخصه نامیده می‌شوند.

| منابع اصلی جهت جمع‌آوری داده‌های ساختاری به‌طور عمده شاخص کتابخانه‌های
| دیجیتالی، بانک‌های اطلاعاتی و پایگاه‌های تخصصی است که به‌صورت معقول دست‌بندی شده و
| مشخص هستند.

| پیشینه | 8  
|--------|
| یکعمال هیوپنده | یکعمال می‌تواند یک شکست، یک مشکلی، یک کلمه، یک صدایی چیز
| یک گیگی باشد و تعیین گام‌تیم‌های آن عبارت است از چهار چیز که هر یک
| شامل، یک سامانه نرم‌افزاری است دارای ویژگی‌هایی در مهندسی و ایجاد تغییر در آن،
| خودکار و تطبیق‌یابی (و باکی به تغییرات مکاتبه‌ای بوده)، اجتماعی،
| یک عمده در جهت "اطلاعات قابلی از مهندسی"، "تجربه‌های قابلی که می‌تواند از آنها

| 1. Classification |
| 2. Text Mining |
| 3. Chunk |
پادگینی را انجام دهد"، "همدیدی که بازی برای رسدیدن به آن تلاش کند"، و "اطلاعات و مشاهداتی از خود و محیط اطراف" را داراست و عملی را انجام می‌دهد.

پژوهشگران هوش مصنوعی معتقدند عامل، یک سامانه کامپیوتری است که علاوه بر یک گی های کلی اشاره شده، به خیال خصوصیات انسانی مانند دانش، اعتقاد، اراده، و تعهد را نیز داراست. در این دیدگاه، عامل دارای یک گی های اضافی زیر است:

جایگاهی: عامل می‌تواند در یک شبکه الکترونیکی تغییر مکان دهد;

صداقت: عامل با استفاده از اطلاعات نادرست منقل نمی‌کند;

خیز‌خواهی: اهداف عامل با یکدیگر در تضاد نیستند و هر عامل در تلاش است تا فقط، وظیفه نحل شدن خود را به انجام برساند.

علقانیت: یک عامل در راستای تحقیق هدف‌شده، رفتار می‌کند (Mohammadian 2004).

عامل از دو بخش برنامه و معاصر تشکیل شده است. برنامه، تابعی است که رفتار عامل را پیاده‌سازی می‌کند و به عبرت دیگر، عمل نگاشت از ادراکات عامل به یک رفتار خاص را بر عهده دارد. وظیفه هوش مصنوعی طراحی برنامه عامل است. درخواست‌های محاسباتی که برنامه عامل بر روی آن اجرا می‌شود، معمولاً عامل نام دارند. معمولاً می‌تواند یک کامپیوتر ساده باشد و یا اینکه در عین حال، تجهیزات سخت‌افزاری خاص ماینی از ابزارهای پردازش صوت و تصویر را هم شامل شود. همچنین، سخت‌افزار وابسته شماری از برنامه‌های پادگینی باشند که میان کامپیوتر و برنامه عامل قرار گرفته و امکان برنامه‌نویسی سطح بالا را فراهم می‌نماید.

معماری از طریق سنجش‌های مشاهداتی را در اختیار عامل قرار می‌دهد، برنامه را اجرا می‌کند و رفتار عامل را از طریق اندازه‌گیری به محیط و سایر عناصری بتکنیک می‌کند (Luck and Padgham 2008).

در این پژوهش، به می‌توان به تکمیل بانک اطلاعاتی مورد نیاز در نرم‌افزاری بازی‌های اطلاعاتی، یک عامل هوشمند نرم‌افزاری نهی شده است که می‌تواند ضمن جایگذاری در محیط و وب، اقدام به جمع‌آوری مینو و داده‌هایی مورد نیاز نماید. در بخش بعد، ساختار اولیه این عامل هوشمند بررسی می‌شود.

9. عامل هوشمند طراحی شده برای این پژوهش

به منظور تاکید بر شرکت حکم عامل نرم‌افزار، سایت روزنامه همشهری به عنوان مبدا ورودی در نظر گرفته شد. با توجه به اینکه در ساختار طراحی شده برای پیکر، فارسی، نوع متون نیز یادی شده است، حکم عامل هربار از یک دسته صورت گرفت (سیاسی، دینی،...
پس از خوانش هر یک از متن موجود در صفحات وب مربوط به بخش تعیین شده، متن در یکه وارد شد و نوع جمله نیز مشخص گردید. از آنجا که هر یک از اخبار DAREX برای نمایش گذاری در HTML و XML با استفاده از خوانش ساختار وب مربوط وارد پنوند، جزء بعدی گردید و محتوا متنی آن جز نیز به یک گره افزوده شد و این عمل به‌همین صورت با خوانش کامل کل سایت ادامه پیدا گردید.

پس از اتمام کار این بخش و مشاهده تجربی متن بازی‌یابی شده، ایرادات عامل نرم‌افزاری مشخص و رفع گردید. لازم به توضیح است که این عامل همانند یک روبوت "عمل می‌کند و پس از رهاسازی در محیط وب، اقدام به جمع آوری داده‌های مورد نیاز و انتخاب بازی‌یابی‌شده را به پیکره متمایز ارسال می‌کند. در اینجا به‌منظور افزایش حجم متن نیاز به پیکره متنی، فهرستی از سایت‌های معترض توجه نگارش متن به‌صورت رسمی، انتخاب شد و به عناوین "فهرست هادی" به عامل نرم‌افزاری داده شد. سپس، با استفاده از دستور GO فرمان حرکت به عامل صادر شد و نرم‌افزار طبق برنامه از قبل تعیین شده اقدام به خوانش و ثبت و حرکت به صفحه وب و وبسایت بعدی می‌نماید. به توجه به همان‌گونه انجام شده در عامل نرم‌افزاری، این کار به‌صورت تمام خودکار و بدون دخالت دست صورت گرفت.

انواع متن دست‌بیدنی شده در پیکره متنی فارسی در گروه‌های زیر فارغ گرفتند:

- سیاست - پژوهشی - ادبیات - ورزشی - هنری - دینی
- علمی - حوادث - اجتماعی - اقتصادی - سایر

در حین خواندن صفحات وب، حجم زیادی از داده‌های غیرمتغیر از قبل عکس‌ها، جداول، و پنوند به‌صورت صفحات بافت شده که پس از تغییر نحو آنها اقدام به حذف داده‌ها گردید. بدین منظور، ابتدا ساختار خوانششده در قابل استفاده XML با نگهداری مشخص شده در آوردگاه پس از درج در پیکره متنی، تغییری اصلی استفاده‌شده در ساختار XML متن به‌صورت جدول 1 درآمد.

جدول 1. فهرست فیلد‌های پیکره

<table>
<thead>
<tr>
<th>ID</th>
<th>شماره جمله</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>متن</td>
</tr>
<tr>
<td>Type</td>
<td>نوع متن</td>
</tr>
<tr>
<td>Link</td>
<td>آدرس صفحه</td>
</tr>
<tr>
<td>Date Time</td>
<td>زمان و ساعت</td>
</tr>
</tbody>
</table>

با توجه به حجم زیاد اطلاعات جمع‌آوری‌شده به منظور افزایش سرعت در پییم‌رسانی اطلاعاتی SQL Server 2005 این سرویس دهنده بانک اطلاعاتی قابلیت بالای آن در پردازش پرسش‌ها و سوال‌های 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 بصورت خوش‌ساخته‌شده است، از دیگر دلایل انتخاب آن بوده است.

10. پیکره‌های متنی مورد نیاز

با توجه به اینکه روی 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 تکنیک‌های SQL Server 2005 و انجام عملیات متینی نکذدانه فارسی و انگلیسی ساخته و استفاده شود، در پژوهش دیگری که توسط موسی میانگاه صورت گرفته است، یک پیکره متین فارسی با بیش از 24700000 جمله و 14900000000000000000 لغت بهره‌مندی شده است (موسی میانگاه 1388) که در این پژوهش با استفاده از روش‌های نرم‌افزاری و الگوریتم عام‌الهای هوشمند و نرم‌افزاری اقدام به تکمیل و گسترش پیکره متنی نکذدانه و دوزبانه شد. در نتیجه، قبل از اینکه روی استفاده از این عام‌الهای هوشمند در پژوهش توضیح داده شود، مقداری درباره عامل‌های بین می‌شود و سپس ساختار نرم‌افزار پیاده‌سازی‌شده در جمع‌آوری و تکمیل پیکره متنی نکذدانه توضیح داده خواهد شد.

10-1. پیکره متین دوزبانه انگلیسی-فارسی

پیکره دیگری که در این پژوهش مورد نزد است، پیکره دوزبانه فارسی-انگلیسی است که در آن متن ترجمه‌شده و معادل آن در فارسی به انگلیسی درج شده‌اند. علاوه بر آن، نوع هر یک از متن نیز مشخص شود. با توجه به اینکه استخراج متن معادل سازی‌شده با استفاده از نرم‌افزار بطور خودکار، خارج از محصولات این پژوهش است، این کار به‌صورت دستی و با جمع‌آوری متن فارسی و انگلیسی ترجمه‌شده صورت گرفت.

<table>
<thead>
<tr>
<th>1. Queries</th>
<th>2. Cluster</th>
<th>3. Intelligent Agent</th>
</tr>
</thead>
</table>

205
پیگیری متنی دوربین با پیش از 2009 وازه نهایه شده است (2009، پژوهش) جازی اقدام به گسترش همان پرده نموده و تعداد وازه‌های آن را به پیش از 4500000 کلمه رسند. و نوع انواع جملات را مطلق پیگیری تکنیک‌های گسترش داده است. شکل 1 نمایش از تطابق دوزبانه انگلیسی-فارسی است.

شکل 1: دوربینه تولید نشده توسعه پیگره برای جستجو

"money laundering"
11. استخراج خودکار چندتایی‌ها

به منظور استخراج همه چندتایی‌ها در زبان‌های مباد، مقصود به این صورت عمل می‌شود:

۱. ابتدا، کل ساختار یک جمله را استخراج و پس، مقدار حرکت به جلو (g) مشخص می‌شود.
۲. مقدار حرکت به جلو و عددي است که بیشترین تصویب کلماتی را که می‌توانند در یک چندتایی معتبر وجود داشته باشند مشخص می‌کند.

در این طرح، ۴ نفر گرفته شد که گفته‌های ۴ یک درمان‌بردار کامپیوتری یا با زبان برنامه نویسی C# نوشته‌شده است، اقدام به شناسایی تمامی ترکیبات ممکن از چندتایی‌های مستر در جمله شد.

12. تعیین چندتایی‌های معتبر

به منظور تعیین چندتایی‌های معتبر و حذف داده‌های اضافه اقدام به دست‌یابی هرم یکی از چندتایی‌ها شد. برای محاسبه این عدد همان گونه که بیشتر در بحث دست‌یابی متوان (میزان واکنش) توضیح داده شد، آگر نتوان میزان واکنشی و با هم آپی هر یک از عبارات را محاسبه کرد، می‌توان نسبت به حذف یا معتبر بودن چندتایی‌های هدست‌آمده تصمیم گیری نمود.

برای این هدف می‌توان از روش X² که در پیش‌های یکی از روش‌های استفاده نمود. انتخاب این روش دینی دلیل صورت گرفت که با استفاده از فرمول X² = (df) ۲ که در این روش در عبارت لازم، در تمامی جملات یک پیکره بررسی و میزان فراوانی ترکیبات مختلف وقوع و با عدم وقوع هر یک با ترکیبی از هر کدام محاسبه Mosavi می‌گردد، می‌توان با استفاده از روش X² میزان واکنشی هر کدام به هدست می‌آید (Miangah, and Nezarat ۲۰۱۰).

جمله زیر را فرض کنید:

"کلاس‌های شما یک روز در میان تنبلی می‌شود." پس از محاسبه مقدار AS برای هر یک از ترکیبات چندتایی در این جمله، جدول ۲ به دست آمد.

<table>
<thead>
<tr>
<th>X²</th>
<th>ترکیب</th>
<th>X²</th>
<th>ترکیب</th>
</tr>
</thead>
<tbody>
<tr>
<td>۵۴۴</td>
<td>شما یک روز در</td>
<td>۶</td>
<td>کلاس‌های شما ۱۰۴۸</td>
</tr>
<tr>
<td>۲۳۴.۱۱</td>
<td>یک روز در</td>
<td>۳۴۱.۵۳</td>
<td>در میان بی‌گزار</td>
</tr>
<tr>
<td>۷۵</td>
<td>یک روز در</td>
<td>۸۰۰۳۳</td>
<td>کلاس‌های شما ۱۰۳۲</td>
</tr>
<tr>
<td>۱۳۹.۴۳</td>
<td>یک روز در</td>
<td>۱۲۵</td>
<td>در میان بی‌گزار</td>
</tr>
<tr>
<td>۲۴۰.۴۴</td>
<td>یک روز در</td>
<td>۱۲۰۰۲۴</td>
<td>شما یک روز</td>
</tr>
</tbody>
</table>

جدول ۲ می‌تواند مقدار AS را برای همه عوامل ممکن
حال بمنظور انتحاب ترکیبات مناسب، نیاز به یک مقدار حد آستانه (جدول ۳) وجود دارد که مقدار آستانه ۶/۳/۶ پس از ماحاسبه مقدار مثلث توسط پژوهشگران انتحاب گردید.

<table>
<thead>
<tr>
<th>پ</th>
<th>(Critical value) ( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>71.2</td>
</tr>
<tr>
<td>0.05</td>
<td>84.3</td>
</tr>
<tr>
<td>0.01</td>
<td>63.6</td>
</tr>
<tr>
<td>0.005</td>
<td>88.7</td>
</tr>
<tr>
<td>0.001</td>
<td>83.10</td>
</tr>
</tbody>
</table>

اگر مقدار \( \chi^2 \) محاسبه شده برای هر یک از ترکیب‌ها کوچکتر از مقدار حد آستانه ۶/۳/۶ باشد، یعنی معنی‌است که می‌توان وایکستگی بین کلمات آن ترکیب را قبول کرد و مقدار بالاتر از حد آستانه را رد نمود.

این روال برای شناسایی همه چندنبایی‌ها در پیکره فارسی ادامه داده می‌شود و یک نانک اطلاعاتی جدید از چندنبایی‌های فارسی تشکیل می‌گردد. سپس، به منظور تعیین معادل انگلیسی هر یک از چندنبایی‌های بست‌آمده با استفاده از پیکره دوزبانه فارسی-انگلیسی امام به یافتن تمامی رکورد‌هایی که مقدار چندنبایی های مورد نظر را دارند، می‌شود. به عنوان مثال، ترکیب چندنبایی "یک روز در میان" را در نظر گرفته می‌شود و از پیکره دوزبانه رکورد‌های با شرط گفتگوهای استخراج می‌گیرد و برای هر یک از رکورد‌های بست‌آمده، انگلیسی، معادل سازی و تیزی محاسبه می‌شود:

1. تمامی ترکیب‌هایی که در دو چهارتایی جمله مناظر انگلیسی استخراج می‌شود؛
2. مقدار فرمول \( \chi^2 \) برای هر یک از این ترکیبات با استفاده از این تعدادی مناظر انگلیسی محاسبه می‌شود;
3. ترکیب‌هایی با مقدار \( \chi^2 \) کمتر از ۶/۳/۶ نگه داشته و مابقی حذف می‌شوند;
4. می‌توان با استفاده از ترکیب‌های به‌دست آمده از چهارتایی بیشتر دو چهارتایی با انگلیسی‌ی به‌دست‌آمده در مرحله قبل همین الگوریتم جدول دانست جدول بعدی تمامی ترکیبات مختلف ساخته می‌شود.

1. Every other day
اجرای و بهبود‌کنیزی یک سامانه بازیابی اطلاعات دوره‌ها | نظام‌های و موسساتیانه

مقدار AS مرتبه‌بین ترکیب‌های جدول محاسبه و مقایسه کمتر از حد آستانه نگه داشته

می‌شود. ملاحظه می‌گردد که بیشترین مقدار AS متعلق به ترکیب مورد نظر، بین عبارت other day است.

جدول ۴. محاسبه  مقدار AS ترکیب‌های مدل

<table>
<thead>
<tr>
<th>ترکیب انگلیسی</th>
<th>ترکیب فارسی</th>
<th>ترکیب انگلیسی</th>
<th>ترکیب فارسی</th>
</tr>
</thead>
<tbody>
<tr>
<td>پک روی در میان</td>
<td>The</td>
<td>پک روی در میان</td>
<td>The</td>
</tr>
<tr>
<td>002.121</td>
<td>The committee</td>
<td>22.240</td>
<td>The committee</td>
</tr>
<tr>
<td>پک روی در میان</td>
<td>The committee converes</td>
<td>پک روی در میان</td>
<td>The committee converes every</td>
</tr>
<tr>
<td>234.1</td>
<td>Every other day</td>
<td>031.7</td>
<td>Every other</td>
</tr>
</tbody>
</table>

۱۳. پاسخ به سوالات پژوهش

یا توجه به آنچه که گذشت و نمونه آزمایشگاهی پدیده‌سازی‌شده، در پاسخ به سوال اول
باید گفت که از فرهنگ لغت در ترجمه عبارات چندین باید سامانه‌های بازیابی اطلاعات بین
زبانی می‌توان فقط به منظور یک ابراز صحة‌گذاری استفاده کرد. در این پژوهش، بودن استفاده
از فرهنگ لغت، ترجمه بیشتر عبارت با استفاده از روش‌های دستبنی متون و اگزیکورم
بهبودی که ترجمه استخراج شد. در مورد پاسخ به سوال دوم پاسخ به سوال دوم پاسخ به سوال دوم که به دلیل قابلیت
یافتن جایگاه عامل‌های دو چندان و مستقل بودن از نظر اجرایی، می‌توان یا کمک یک عامل
هوشمند و جایگاه در و پیچیده‌های مختلف، همان‌طور که روبوت ایمنی اقدام به تکمیل
یک پیکر از داده‌های متنی آن و سایت‌ها نمود.

در پاسخ به سوال سوم که استفاده از پیکره‌های کنترلی برای کاربرد در بازیابی اطلاعات
بین زبانی مورد توجه قرار گرفته است، می‌توان گفت که به دلیل پیچیدگی عبارات در زبان
فارسی، لزوم به کار بردن یک پیکر اطلاعاتی مستقل در کنار فرهنگ لغت الزامی است. این
پیکر اطلاعاتی در واقع، همان پیکر زبانی است که شامل جملات و ترجمه‌های معادل آنها در
زبان مقصد است که چپ از ترجمه کلمات با استفاده از فرهنگ لغت می‌توان از پیکره به عضو
ایزاب صحنه‌گذاری ترجمه استفاده کرد.
14. تیپه‌گری

در این پژوهش مدل شده ات فکارا بر تکمیل پیکره‌های منتی ساخته‌شده در طرح‌های پیشین (یکسانه و دوزبانه) یک روش آماری دست‌نیز جهت تعیین میزان واکنشی عبارات به یکدیگر ارائه شود و با استفاده از آن اقدام به رفع ایجاد در ترجمه عبارات جنگی گردد.

همان‌گونه که در گزارش مطرح شد، یکی از مشکلات عمده در پایداری اطلاعات بین زبانی، عدم سطح کاربران به ترجمه دقیق ترکیبات جنگی و حیات آن بر تولید مشکل در فرآیند پایداری اطلاعات و پایداری اطلاعات و متن این کسب و کار را مطرح می‌کند. از این برخی از ترکیبات جنگی در لغتنامه‌ها نیز باید مشکل افزوده‌ای است. در این پژوهش، علاوه بر تکمیل پیکره‌های منتی یک و دوزبانه، یک آمار در آنها براساس اطلاعات موجود در پیکره‌ها که ترجمه تأثیر پزشکی داشته باشد.

در این روش، با استفاده از همن‌گونه X یا همان روش میزان استفاده از زبان و احتمال آن در زبان دیگر براساس جملات موجود در پیکره‌ها ترجمه کاندید مشخص می‌شود. انتخاب ترجمه کاندید نیز با استفاده از جدول حفظ آسانه‌ای نهایتی بحراط صورت گرفته.

یکی از دستاوردهای اخیر در این طرح، بالا بردن دقیق سامانه‌های پایداری اطلاعات در موتورهای جستجو است که با استفاده از پیکره و یکسانه اطلاعاتی، ترکیب‌بندی و از این دسترس است. با توجه به اینکه در انتهای طرح، پیکره‌های دوزبانه یک در همه موارد با یکسانه اطلاعاتی بسیار غناه به وجود آمده است، می‌توان پیش بینی نمود که در طرح‌های مجزا از دیگر یا اضافه‌ای است. داشت به‌همراه مورد در این طرح، محصولی از قبیل واژه‌های منتی بر پیکره و نیز سامانه حافظه ترجمه به‌عنوان غالفی است. نمود.

در عامل‌های آن، پژوهشگران فرد دارند از پیکره تولید شده به عنوان یک مرحله دانش اضافه‌ای و یک فرهنگ لغت جدید ایجاد نمایند. بسیاری منظور استفاده از روش‌های داده‌کاوی جهت تکمیل پیکره یکسانه و دوزبانه پیشنهاد می‌شود. همچنین، به منظور افزایش

1. Corpus-based dictionary                   2. Translation memory system
سرعت محاسبات می‌توان با تغییر در روش محاسبه AS، سرعت دستبندی کلمات و معادل سازی آنها را افزایش داد.

15. منابع

علیزاده، حمید و همکاران. ۱۳۸۸. بررسی یک الگوریتم روش‌های موجود در پیام‌برداری اطلاعات بین زبانی فارسی- انگلیسی با استفاده از واسکوم دوشان ماسین خوان. فصل‌نامه علوم و فناوری اطلاعات ایران ۲۵ (۱): ۵۳-۷۰.

موسوی مینگاهی، طیبه. ۱۳۸۸. تجربه برگه‌های بی‌گروه بین‌زبانی در بهبود کیفیت ترجمه مانسی‌تی. طرح پژوهشی، دانشگاه پیام نور، زمستان ۱۳۸۸.


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

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

Tayebeh Mosavi Miangah1
Associate Professor of Applied Linguistics, Payame Noor University, Yazd

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 extent 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