بسم الله الرحمن الرحیم

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکدهٔ مهندسی برق و کامپیوتر

استفاده از پیکره‌های تطبیقی برای بازیابی اطلاعات بین زبانی فارسی-انگلیسی

نگارش
هما برادران هاشمی

استاد راهنما
دکتر آزاده شاکری

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد در رشتهٔ
مهندسی کامپیوتر - گرایش نرم‌افزار

بهمن ۱۳۸۹

چکیده

امروزه با افزایش حجم اطلاعات به زبان‌های مختلف در وب، تقاضا برای سیستم‌های بازیابی اطلاعات که بتوانند از مرز بین زبان‌ها عبور کنند، رشد فزاینده‌ای داشته است. با استفاده از چنین سیستم‌هایی می‌توان تنها با دادن پرس‌وجو در یک زبان، اسنادی با زبان‌های مختلف را بازیابی کرد. از آنجایی که پرس‌وجو و اسناد در زبان‌های متفاوتی هستند، پیش از انجام تطابق باید از نوعی ترجمه استفاده کرد. به این منظور معمول‌ترین منابع ترجمه‌ای که استفاده می‌شوند، واژه‌نامه‌ها، ماشین‌های ترجمه، پیکره‌های موازی و پیکره‌های تطبیقی می‌باشند. واژه‌نامه‌ها، ماشین‌های ترجمه و پیکره‌های موازی منابع پرهزینه‌ای هستند و برای تعداد زیادی از جفت زبان‌های اقلیت وجود ندارند. این در حالی است که پیکره‌های تطبیقی بین زبان‌ها را می‌توان بسیار راحت‌تر فراهم نمود.

در این پایان‌نامه، مسأله بازیابی اطلاعات بین زبانی با استفاده از پیکره‌های تطبیقی، بین دو زبان فارسی و انگلیسی بررسی شده است. پیکره‌های تطبیقی، اسنادی با موضوع‌های مشابه در دو زبان متفاوت را شامل می‌شوند. به عنوان مثال، اخبار منتشر شده بر روی وقایع مشترک در زبان‌های مختلف، نمونه‌ای از این نوع پیکره‌ها می‌باشند. برترین مزیت این گونه از منابع در دسترس بودن آنها است که باعث شده ایجاد و استفاده از پیکره‌های تطبیقی به عنوان زمینه پژوهشی ارزشمندی در بازیابی اطلاعات بین زبانی مطرح گردد. همچنین لازم به ذکر است که تا کنون پژوهشی در زمینه‌ی استفاده از پیکره‌های تطبیقی به منظور بازیابی اطلاعات بین زبانی در زبان فارسی صورت نگرفته است.

در این پژوهش، نخستین پیکره تطبیقی بزرگ فارسی-انگلیسی (UTPECC) با هم‌ترازی اخبار فارسی روزنامه همشهری و اخبار انگلیسی خبرگزاری بی‌بی‌سی، بر اساس تاریخ انتشار و شباهت اسناد، ساخته و ارزیابی شده است. در بخش دیگر پژوهش، از پیکره‌ی تطبیقی ساخته شده برای بازیابی اطلاعات بین زبانی فارسی-انگلیسی استفاده می‌گردد. به منظور استخراج دانش ترجمه از پیکره‌های تطبیقی، روشی مبنی بر شبکه وابستگی اصطلاحات ارائه شده است که هم از ارتباطات اصطلاحات در یک زبان و هم از ارتباطات آنها بین دو زبان بهره می‌برد. از آنجایی که اصطلاحات هم می‌توانند کلمات و هم عبارات را شامل شوند، در نتیجه ترجمه‌های بدست آمده از این روش می‌توانند برای ترجمه پرس‌وجوها در بازیابی اطلاعات بین زبانی مؤثر باشند. همچنین در این پژوهش نشان داده شده است که می‌توان صحّت ترجمه‌های استخراج شده را توسط کشف برون‌هشته‌ها بررسی کرد. نتایج آزمایش‌ها بر اساس مدل زبانی پرس‌وجو، نشان می‌دهند که ترجمه‌های استخراج شده از پیکره‌های تطبیقی روشی مؤثر برای بازیابی اطلاعات بین زبانی هستند و می‌توانند با ترجمه‌ی کلماتی که در واژه‌نامه‌ها نیستند و همچنین با استخراج کلمات مرتبط به آنها، دقت بازیابی را بهبود دهند.

واژه‌های کلیدی: *بازیابی اطلاعات بین زبانی، پیکره تطبیقی فارسی-انگلیسی، شبکه ارتباطی کلمات.*

فهرست مطالب

۱ مقدمه ۱

۲ بازیابی اطلاعات بین زبانی ۹

۱.۲ روش‌های بازیابی اطلاعات بین زبانی ۱۰

۱.۱.۲ روش‌های مبنی بر واژه‌نامه ۱۲

۲.۱.۲ ماشین ترجمه ۱۴

۳.۱.۲ روش‌های مبنی بر پیکره ۱۴

۲.۲ پیکره‌های تطبیقی ۱۵

۱.۲.۲ ساخت پیکره‌های تطبیقی ۱۷

۲.۲.۲ استخراج دانش ترجمه از پیکره‌های تطبیقی ۱۹

۳.۲.۲ استفاده از دانش ترجمه استخراج شده ۲۵

۳.۲ پیکره‌های زبان فارسی ۲۶

۴.۲ بازیابی اطلاعات بین زبانی فارسی و انگلیسی ۲۷

۳ ساخت پیکره تطبیقی فارسی-انگلیسی ۳۱

۱.۳ ساخت و ترجمه نمایش اسناد ۳۲
۲.۳ هم‌ترازی اسناد ۳۴
۳.۳ ارزیابی پیکره تطبیقی ساخته شده ۳۴
۱.۳.۳ معیار ارزیابی ۳۵
۲.۳.۳ مشخصات مجموعه اسناد فارسی و انگلیسی ۳۵
۳.۳.۳ نتایج ارزیابی پیکره‌های تطبیقی ۳۷
۴.۳ مقایسه با کارهای پیشین ۴۱

۴ بازیابی اطلاعات بین زبانی با استفاده از پیکره‌های تطبیقی ۴۳
۱.۴ استخراج دانش ترجمه از پیکره‌های تطبیقی ۴۴
۱.۱.۴ COCOT: روش پایه استخراج دانش ترجمه ۴۵
۲.۱.۴ TAN: مدل شبکه‌ای وابستگی اصطلاحات ۴۶
۳.۱.۴ بررسی صحّت ترجمه ۵۰
۴.۱.۴ استخراج ترجمه عبارات ۵۴
۲.۴ ساخت مدل زبانی پرس‌وجو ۵۶
۱.۲.۴ انتخاب تعداد همسایه ثابت ۵۶
۲.۲.۴ انتخاب تعداد همسایه پویا ۵۷
۳.۲.۴ استفاده از واژه‌نامه ۵۹

۵ ارزیابی نتایج آزمایش‌ها ۶۱
۱.۵ مجموعه‌های اسناد ۶۲
۱.۱.۵ پیکره تطبیقی UTPECC ۶۲

۲.۱.۵ پیکره همشهری ۶۲

۳.۱.۵ پیکره INFILE ۶۳

۲.۵ معیارهای ارزیابی ۶۳

۳.۵ نتایج بازیابی اطلاعات بین زبانی ۶۵

۱.۳.۵ نتایج بازیابی اطلاعات تک زبانی ۶۶

۲.۳.۵ COCOT با تعداد همسایه ثابت ۶۷

۳.۳.۵ شبکه وابستگی اصطلاحات با تعداد همسایه ثابت ۶۹

۴.۳.۵ شبکه وابستگی اصطلاحات با تعداد همسایه پویا ۷۲

۵.۳.۵ بررسی صحّت ترجمه ۷۳

۶.۳.۵ ترجمه عبارات ۷۴

۷.۳.۵ ترکیب با واژه‌نامه ۷۷

۸.۳.۵ گسترش پرس‌وجو توسط شبه بازخورد ارتباطی ۷۹

۹.۳.۵ استفاده از پیکره INFILE ۸۰

۶ نتیجه‌گیری ۸۵

۱.۶ جمع‌بندی ۸۵

۲.۶ کارهای آتی ۸۷

آ پرس‌وجوهای پیکره INFILE ۹۱

مراجع ۹۴

واژه‌نامه فارسی به انگلیسی ۱۰۵

فهرست جدول‌ها

۱.۳ مشخصات مجموعه اسناد انگلیسی (بی‌بی‌سی) و فارسی (همشهری) ۳۶

۲.۳ نمونه‌ای از هم‌ترازی اسناد فارسی و انگلیسی ۳۶

۳.۳ ارزیابی کیفیت هم‌ترازی‌ها در یک ماه، ژانویه ۲۰۰۲. بر اساس مدل بازیابی KL-divergence. (آ) استفاده از تمام ترجمه‌های هر کلمه. (ب) استفاده از ۳ بالاترین ترجمه هر کلمه. ۳۸

۴.۳ نتایج ۳ بالاترین ترجمه‌ها با و بدون نویسه گردانی و مدل بازیابی KL-divergence ۳۹

۵.۳ نتایج ۳ بالاترین ترجمه‌ها با و بدون نویسه گردانی و مدل بازیابی Okapi ۳۹

۶.۳ توزیع کیفیت هم‌ترازی‌ها برای دو مجموعه متفاوت از حدود آستانه ۴۰

۱.۵ خصوصیات پیکره تطبیقی ساخته شده ۶۲

۲.۵ کارآیی بازیابی تک زبانی فارسی-فارسی ۶۷

۳.۵ نمونه‌ای از ترجمه‌های استخراج شده انگلیسی-فارسی با بیشترین امتیاز شباهت ۶۷

۴.۵ ترجمه‌های فارسی استخراج شده از پیکره تطبیقی برای چهار کلمه انگلیسی ۶۸

۵.۵ کارآیی روش پایه COCOT ۷۰

۶.۵ کارآیی مدل شبکه TAN بدون در نظر گرفتن همسایگی ۷۱

۷.۵ کارآیی مدل شبکه TAN با در نظر گرفتن همسایگی ۷۲

۸.۵ کارآیی روش‌های مختلف ترجمه پرس‌وجو با استفاده از پیکره تطبیقی ۷۳
۹.۵ نمونه‌هایی از ترجمه عبارات استخراج شده از پیکره تطبیقی توسط نگرش اول ۷۶
۱۰.۵ نمونه‌هایی از ترجمه عبارات استخراج شده از پیکره تطبیقی توسط نگرش دوم ۷۷
۱۱.۵ کارآیی ترجمه عبارات ۷۷
۱۲.۵ کارآیی روش‌های مختلف ترجمه پرس‌وجو ۸۰
۱۳.۵ کارآیی گسترش پرس‌وجو توسط شبه بازخورد ارتباطی ۸۰
۱۴.۵ کارآیی بازیابی تک زبانی انگلیسی-انگلیسی ۸۱
۱۵.۵ ارزیابی روش‌های مختلف ترجمه پرس‌وجو در پیکره INFILE ۸۴

آ.۱ پرس‌وجوهای پیکره INFILE ۹۱

فهرست شکل‌ها

۱.۲ روش‌های اصلی بازیابی اطلاعات بین زبانی [۳۳] ۱۲

۲.۲ مدل پیشنهادی در مرجع [۷۳] به منظور استخراج ترجمه کلمات و عبارت ۲۳

۱.۳ مراحل ساخت پیکره‌ی تطبیقی ۳۲

۱.۴ مراحل بازیابی اطلاعات بین زبانی ۴۴

۲.۴ ساختار شبکه وابستگی اصطلاحات ۴۸

۳.۴ دو مثال برای بررسی صحّت ترجمه‌ها بر اساس کشف برون‌هشته‌ها. ۵۳

۴.۴ انتخاب تعداد همسایه پویا برای کلمات انگلیسی "sanction" و "atomic" ۵۸

۱.۵ میانگین متوسط دقت برای روش‌های مختلف مبنی بر پیکره تطبیقی ۶۹

۲.۵ میانگین متوسط دقت با استفاده از کشف کلمات برون‌هشته ۷۵

۳.۵ میانگین متوسط دقت برای روش‌های مختلف مبنی بر واژه‌نامه ۷۹

۴.۵ میانگین متوسط دقت برای روش‌های مختلف مبنی بر پیکره تطبیقی در پیکره INFILE ۸۲

۵.۵ میانگین متوسط دقت با استفاده از کشف کلمات برون‌هشته در پیکره INFILE ۸۳

۶.۵ میانگین متوسط دقت برای روش‌های مختلف مبنی بر واژه‌نامه در پیکره INFILE ۸۴

فصل ۱

مقدمه

هدف بازیابی اطلاعات تامین راه‌هایی برای پیدا کردن اسناد مربوط به نیاز اطلاعاتی کاربران است. در یک سیستم معمول بازیابی اطلاعات، کاربر نیاز اطلاعاتیش را به صورت یک پرس وجو بیان کرده و سیستم، مجموعه‌ای از داده‌ها را برای پیدا کردن اسناد مربوط به آن پرس‌وجو جستجو می‌کند.

با ظهور وب، سیستم‌های بازیابی اطلاعات اهمیت ویژه‌ای پیدا کرده‌اند و موتورهای جستجوی حاضر کم و بیش قادر به پاسخگویی نیاز کاربران هستند. با افزایش حجم اطلاعات به زبان‌های مختلف در وب، تقاضا برای سیستم‌های بازیابی اطلاعات که بتوانند از مرزهای بین زبان‌ها عبور کنند، رشد فزاینده‌ای داشته است. با چنین سیستم‌هایی می‌توان تنها با دادن پرس‌وجویی در یک زبان، اسناد مربوط حتی با زبان‌های دیگر را بازیابی کرد.

مقصد بازیابی اطلاعات بین زبانی[1] دست‌یابی به این هدف است که بتوان اسناد مرتبط نوشته‌شده به زبانی متفاوت از پرس‌وجو را پیدا کرد. در این میان، زبان پرس‌وجو به عنوان *زبان مبدأ* و زبان اسناد به عنوان *زبان مقصد* در نظر گرفته می‌شوند. از آنجایی که پرس‌وجو و اسناد در زبان‌های متفاوتی هستند، تطابق مستقیم پرس‌وجو و اسناد امکان‌پذیر نمی‌باشد. بنابراین پیش از انجام تطابق باید از نوعی ترجمه استفاده کرد. سه روش کلی برای ترجمه مطرح است: ترجمه پرس‌وجو، ترجمه اسناد و ترجمه توأم پرس‌وجو و اسناد. از این میان ترجمه پرس‌وجو متداول‌تر است و از پرس‌وجوی ترجمه شده به زبان مقصد برای بازیابی اسناد زبان مقصد استفاده می‌شود. یک مثال معمول در کاربرد بازیابی اطلاعات بین زبانی هنگامی است که زبان مبدأ، زبان مادری کاربر در نظر گرفته می‌شود در حالی که زبان مقصد می‌تواند زبانی

[1] Cross-Language Information Retrieval

باشد که کاربر تنها کمی با آن آشنا است. ممکن است کاربر نتواند به راحتی متنی را به زبان مقصد بنویسد و حتی نوشتن یک پرس‌وجوی کوتاه هم ممکن است برایش بسیار دشوار باشد، ولی امکان دارد قادر به خواندن اسناد بازیابی شده باشد. یا کاربری ممکن است به چندین زبان مسلط باشد و بخواهد نتیجه‌ی پرس‌وجوی مورد نظرش را به تمام آن زبان‌ها در یک زمان مشاهده کند و لازم نباشد که چندین بار به چندین زبان آن پرس‌وجو را تکرار کند. واضح است که وجود یک سیستم بازیابی اطلاعات با قابلیت بازیابی بین زبانی برای این چنین کاربرانی بسیار سودمند است.

معمول‌ترین منابع ترجمه‌ای که به منظور ترجمه پرس‌وجو به کار می‌روند، واژه‌نامه‌ها، ماشین‌های ترجمه، پیکره‌های موازی و پیکره‌های تطبیقی [۳۸] می‌باشند. در روش مبنی بر واژه‌نامه، کلمات پرس‌وجوی زبان مبدأ با ترجمه معادلشان در واژه‌نامه دوزبانه جایگزین می‌شوند. اگر چه این روش آسان است، ولی مشکلات خاص خودش را دارد که مهمترین آن کلمات غیر قابل ترجمه (کلماتی که در واژه‌نامه نیستند) و ابهام ترجمه (انتخاب از میان معناهای متفاوت در ترجمه یک کلمه) است. به عنوان مثال، کلمه انگلیسی bank هم به معنای بانکداری است و هم معنای کناره‌ی رودخانه را دارد. بدون داشتن اطلاعات متن تشخیص اینکه کدام کلمه ترجمه صحیح می‌باشد، تقریباً غیر ممکن است.

روش ترجمه ماشینی به احتمال زیاد بهترین نتایج ترجمه را بر می‌گرداند. ولی از طرف دیگر، این روش بر این فرض استوار است که ورودی یک واحد منظم طبق قواعد صرف و نحوی است، همانند جملات اخباری "*به دنبال اطلاعاتی درباره نظریه گفتگوی تمدن‌ها و نقش رئیس جمهور خاتمی می‌گردم*" و یا یک سوال "*نقش رئیس جمهور خاتمی در نظریه گفتگوی تمدن‌ها چه بود؟*". ماشین‌های ترجمه برای اینکه محتمل‌ترین ترجمه را انتخاب کنند تا حدود زیادی بر روی اطلاعات زبان‌شناختی متن تکیه می‌کنند. مشکل اینجاست که بیشتر کاربران پرس‌وجوی خود را نه تنها به صورت یک جمله با قاعده نمی‌نویسند، بلکه معمولاً به صورت لیستی از کلمات کلیدی، همانند *رئیس جمهور خاتمی*، *گفتگوی تمدن‌ها*، و بدون توجه به هیچ‌گونه ترتیب کلمات و یا محدودیت‌های نحوی دیگر می‌نویسند. از جمله اشکالات دیگر اکثر ماشین‌های ترجمه، به ویژه سیستم‌های تجاری، در بازیابی اطلاعات بین زبانی این است که آنها تنها محتمل‌ترین ترجمه را برمی‌گردانند، در حالی که محتمل‌ترین ترجمه بر اساس الگوریتم داخلی ماشین ترجمه تعریف شده است و این بدین معنا نیست که ترجمه‌های خوب دیگری - و یا حتی از منظر دیگری ترجمه‌های بهتری- وجود نداشته است [۵۲].

همچنین ممکن است همیشه منابع زبان‌شناسی غنی (از جمله واژه‌نامه‌ها و ماشین‌های ترجمه) بین دو زبان در دسترس نباشد. در نتیجه برای رسیدگی به این مسأله مهم از روش‌های مبنی بر پیکره‌های چند زبانه به منظور استخراج آماری دانش ترجمه، استفاده می‌شود. پیکره‌ها که مجموعه‌ای از مستندات هستند، می‌توانند بر اساس ارتباط متون به دو دسته موازی و تطبیقی تقسیم شوند. پیکره موازی، مجموعه‌ای است که متون زبان مبدأ به ترجمه‌های دقیقشان در زبان مقصد نگاشت شده‌اند. به عنوان مثال، قوانین اتحادیه اروپا یک پیکره موازی است، زیرا که هر قانون به تمامی زبان‌های رسمی اروپا

نوشته شده است. از طرف دیگر در پیکره تطبیقی متون ترجمه‌ای از یکدیگر نیستند، ولی از طریق موضوع تا حدودی به هم مرتبط هستند [۸۰]. به عنوان مثال، اسناد هم‌تراز شده در این پیکره‌ها می‌توانند اخبار منتشر شده در کشورهای مختلف بر روی وقایع مشترک با زبان‌های متفاوت باشند.

از آنجایی که در پیکره‌های موازی و تطبیقی، زوج کلمات در زبان‌های مختلف در متون مشابهی ظاهر می‌شوند، می‌توان از این گونه پیکره‌ها برای بدست‌آوردن دانش ترجمه استفاده کرد. به عنوان مثال زوج کلماتی همانند *انتخابات* و *رأی دادن* احتمالاً در اخبار فارسی مربوط به انتخابات ایران ظاهر شده و به طور مشابه کلمات متناظرشان، *(election)* و *(to vote)*، نیز در اخبار انگلیسی درباره‌ی هر انتخابات دیگری نیز ظاهر می‌شوند.

طبیعتاً ترجمه برخی واژگان را نمی‌توان از پیکره موازی استخراج کرد، به عنوان مثال، نمی‌توان با اطمینان کامل واژگان ورزش فوتبال را بر اساس پیکره موازی که شامل قوانین اتحادیه اروپا است ترجمه کرد. این مسأله به عنوان مشکل اصلی پیکره‌های موازی مطرح است. اگرچه مجموعه‌های مستندات موازی ساخت بشر به تعداد زیادی وجود دارد، ولی این مجموعه‌ها عموماً کمبودهایی در پوشش حوزه‌های خاص دارند و همیشه در دسترس نیستند. این در حالیست که برترین مزیت پیکره‌های تطبیقی در دسترس بودن آنها است. از این‌رو امروزه ایجاد و استفاده از پیکره‌های تطبیقی به عنوان یک زمینه تحقیقاتی معتبر در بازیابی اطلاعات بین زبانی مطرح است.

دستاوردهای این پژوهش

ما در این پایان‌نامه بازیابی اطلاعات بین زبانی را با فرض داشتن منابع زبان‌شناسی محدود، به طور خاص پیکره‌های تطبیقی بین دو زبان فارسی و انگلیسی، بررسی می‌کنیم. پیکره‌های تطبیقی، اسنادی در دو زبان متفاوت هستند که موضوعات مشابهی را می‌پوشانند. به عنوان مثال اخبار منتشر شده بر روی وقایع مشترک در دو زبان مختلف نمونه‌ای از این گونه پیکره‌ها می‌باشند. اگر چه ممکن است منابع زبان‌شناسی غنی همانند واژه‌نامه‌ها و یا پیکره‌های موازی بین دو زبان نداشته باشیم، ولی معمولاً پیکره‌های تطبیقی بین دو زبان وجود دارد. همچنین لازم به ذکر است که بر اساس دانسته‌های نگارنده، تا کنون تحقیقاتی در زمینه پیکره‌های تطبیقی به ویژه به منظور بازیابی اطلاعات بین زبانی در زبان فارسی صورت نگرفته است.

هدف از این پروژه تمرکز بر روی روش‌های مبنی بر پیکره‌های تطبیقی در بازیابی اطلاعات بین زبانی است و بررسی چگونگی تأثیر این گونه از پیکره‌ها بر روی کارآیی سیستمی که از پیکره‌ها به عنوان منبع ترجمه استفاده می‌کند. همان‌طور که گفتیم پیکره‌های چندزبانی، پیکره‌های موازی و یا تطبیقی، بین زوج زبان‌های زیادی وجود دارند. ولی زبان فارسی،

به عنوان زبانی که به طور گسترده در خاور میانه صحبت می‌شود، منابع غنی زبان‌شناسی در این زمینه ندارد. دلیل عمده این کمبودها ویژگی‌های خاص و مشکلات ساخت پیکره‌ها در این زبان است [۲۷، ۵]. بدین منظور در گام نخست این پژوهش پیکره‌ی تطبیقی فارسی-انگلیسی UTPECC[1] را می‌سازیم که بنابر اطلاعات ما نخستین پیکره بزرگ تطبیقی فارسی-انگلیسی است. در این پژوهش، پیکره فارسی-انگلیسی را توسط اخبار فارسی روزنامه همشهری[2] و اخبار انگلیسی خبرگزاری بی‌بی‌سی[3] می‌سازیم. در مجموع حدود ۵۳،۰۰۰ سند انگلیسی با ۱۹۰،۰۰۰ سند فارسی هم‌تراز شده و نتیجه پیکره تطبیقی با بیش از ۱۰،۳۶۵ زوج سند حاصل شده است. بسیاری از مطالعاتی که بر روی پیکرهای تطبیقی برای استخراج وابستگی کلمات بین زبانی صورت گرفته‌اند فرض کرده‌اند که پیکره‌های تطبیقی به راحتی می‌توانند با استفاده از هم‌ترازی تاریخ انتشار اخبار در زبان‌های مختلف بدست آیند [۸۹، ۹۳، ۹]. زیرا که اخباری که در یک بازه زمانی منتشر می‌شوند عموماً رویدادهای مهم بین‌المللی را در موضوعات متفاوتی از جمله سیاسی، علمی و ورزشی، پوشش می‌دهند. اگرچه ممکن است این فرض برای بعضی از زبان‌ها درست باشد، ولی در مورد زبان فارسی این گونه نیست. اخبار خبرگزاری‌هایی که روزانه مقالاتی به هر دو زبان فارسی و انگلیسی منتشر می‌کنند بدین منظور مناسب نیستند، زیرا که مقالات انگلیسی آنها معمولاً بسیار کوتاه هستند و در اکثر مواقع ترجمه‌ی خلاصه شده‌ای از مقالات فارسی می‌باشند و علاوه بر این آرشیو طولانی مدتی به صورت آنلاین ندارند که به راحتی قابل دستیابی باشد. از طرف دیگر، از آنجایی که اخبار مرتبط به یک واقعه‌ی مشترک در تاریخ‌های مختلفی در خبرگزاری‌های متفاوت منتشر می‌شوند، نمی‌توان به راحتی آنها را توسط تنها تاریخ انتشار هم‌تراز کرد. به عنوان مثال اخبار علمی که به نقل از خبرگزاری‌های خارجی در روزنامه‌های ایران منتشر می‌شوند معمولاً با چند روز تأخیر همراه هستند. بنابراین در این پژوهش برای ساخت پیکره‌ی تطبیقی از دو مجموعه اخبار متفاوت از دو مبدأ، مجموعه اخبار انگلیسی بی‌بی‌سی و مجموعه اخبار فارسی همشهری، استفاده کرده‌ایم و هم‌ترازی اسناد را بر اساس تاریخ انتشار و شباهت محتوای متنی اخبار آنها انجام داده‌ایم.

پس از ساخت پیکره تطبیقی، به منظور انجام بازیابی اطلاعات بین زبانی باید ترجمه‌ی کلمات به گونه‌ای از پیکره استخراج شوند و پرس‌وجوها توسط آنها ترجمه گردند. به منظور استخراج دانش ترجمه، در این پژوهش روش‌های متفاوتی را ارائه کرده‌ایم که به عنوان روش پایه، ترجمه را بر مبنای رخداد همزمان کلمات در هم‌ترازی‌های پیکره تطبیقی استخراج می‌کنیم. در گام بعدی، روشی را مبنی بر شبکه وابستگی اصطلاحات[4] معرفی می‌کنیم که از ارتباط کلمات

---

[1] University of Tehran Persian-English Comparable Corpus

[2] www.hamshahrionline.ir

[3] www.bbc.co.uk

[4] TAN: Term Association Network

در داده‌های تک زبانی علاوه بر ارتباط کلمات بین دو زبان بهره می‌برد. ایده اصلی این روش استفاده از اطلاعات همسایگی‌های کلمات است. به طور شهودی می‌توان گفت که دو کلمه در دو زبان متفاوت به احتمال زیاد ترجمه یکدیگر هستند اگر کلمات همسایگی اطرافشان با یکدیگر ارتباط زیادی داشته باشند. به عبارت دیگر، اگر کلمات همسایگی آنها در زبان بومی آنها با یکدیگر همبستگی زیادی نداشته باشند، آنگاه می‌توان گفت که دو کلمه نمی‌توانند ترجمه یکدیگر باشند. در شبکه ساخته شده، یال‌های بین کلمات در یک زبان اطلاعات متقابل[۱] بین کلمات را نشان می‌دهد و یال‌های بین کلمات در زبان‌های مختلف ارتباط رخداد همزمان آنها در هم‌ترازی‌ها هستند که توسط روش پایه بدست می‌آیند. با استفاده از این شبکه، در حقیقت از میزان شباهت همسایگی کلمات برای بهبود ترجمه‌های استخراج شده استفاده می‌کنیم.

ترجمه‌های استخراج شده از پیکره‌ها توسط روش‌های آماری این امکان را می‌دهد که از امتیاز ترجمه‌ی محاسبه شده‌ی آنها برای ساخت پرس‌وجو استفاده شود. بدین صورت که به ترجمه‌های مختلف یک کلمه بر اساس ارتباطش، وزنی اختصاص داده می‌شود. همچنین مزیت دیگر استفاده از روش‌های مبنی بر پیکره این است که می‌توان از ترجمه‌های استخراج شده برای گسترش پرس‌وجو[۲] استفاده کرد. بدین صورت که هر چه وزن ارتباطی ترجمه‌های استخراج شده برای کلمه‌ای کمتر می‌شود، اگر چه ممکن است دیگر ترجمه‌ی دقیقی از آن کلمه نباشند ولی کماکان کلماتی مرتبط می‌باشند و می‌توان از آنها برای گسترش پرس‌وجو و افزایش دقت بازیابی اطلاعات بین زبانی استفاده کرد. در ادامه به منظور استفاده از ترجمه‌های استخراج شده برای بازیابی اطلاعات بین زبانی، مدل زبانی پرس‌وجوها را بر اساس امتیازهای ترجمه‌ی استخراج شده در زبان مقصد می‌سازیم. نتایج آزمایش‌ها بر روی مجموعه همشهری با پرس‌وجوهای CLEF 2008, 2009[۳] برای بازیابی انگلیسی-فارسی، نشان می‌دهند که با روش پیشنهادی می‌توانیم به ۴۳/۳% میانگین متوسط دقت، ۴۵/۴% دقت در ۵ سند اول و ۴۴/۷% دقت در ۱۰ سند اول در مقایسه با کارآیی بازیابی تک زبانی فارسی-فارسی دست یابیم.

از آنجایی که پیکره تطبیقی در این پژوهش را از اخبار روزنامه همشهری ساخته‌ایم و از ترجمه‌های استخراج شده از آن برای بازیابی اطلاعات بین زبانی بر روی پیکره همشهری استفاده کرده‌ایم، آزمایش‌هایمان را بر روی پیکره‌ی دیگری که مبدأ کاملاً متفاوتی دارد نیز انجام داده‌ایم تا بتوانیم تحلیل درستی از نتایج بدست آمده داشته باشیم. از پیکره INFILE[۴]

---

[۱] mutual information

[۲] query expansion

[۳] www.clef-campaign.org

[۴] http://www.infile.org/

به صورت بازیابی بین زبانی فارسی-انگلیسی استفاده کرده‌ایم. به طور کلی، نتایج بدست آمده از اعمال روش استخراج ترجمه توسط شبکه، امید بخش هستند و نسبت به روش پایه بهبود قابل توجهی کسب کرده‌اند.

با در نظر گرفتن زوج کلمات کنار هم در یک زبان به عنوان عبارت، می‌توان روش‌های ارائه شده برای استخراج کلمات را برای بدست آوردن ترجمه‌ی عبارات نیز استفاده کرد. نتایج آزمایش‌ها نشان می‌دهند که ترجمه عبارات بدست آمده به بازیابی اطلاعات بین زبانی کمک می‌کنند.

در ادامه پژوهش، به منظور بررسی صحّت ترجمه‌های استخراج شده، از روشی بر اساس کشف برون‌هشته‌ها[۱] استفاده شده است تا بتوان کلماتی که ترجمه درستی برایشان استخراج نشده‌اند را مشخص کرد. نتایج آزمایش‌ها نشان می‌دهند که با حذف این برون‌هشته‌ها می‌توان کارآیی سیستم بازیابی اطلاعات را افزایش داد. در آزمایش‌های دیگری که بر اساس تنها ترجمه‌ی پنج کلمه از برون‌هشته‌ها توسط واژه‌نامه انجام داده‌ایم، به میانگین متوسط دقت ۴۵.۸٪ بازیابی تک زبانی دست یافته‌ایم.

در این پژوهش نشان می‌دهیم که حتی هنگامی که منابع غنی ترجمه، همانند واژه‌نامه‌ها، بین دو زبان وجود دارند، هنوز هم ممکن است به استخراج دانش ترجمه از پیکره‌های تطبیقی نیاز باشد. دو دلیل مطرح در این زمینه به شرح زیر می‌باشد. اول این که دائماً کلمات و عبارات جدیدی در حال اضافه شدن به زبان‌ها می‌باشند و به روز نگه داشتن واژه‌نامه‌ای که تمام آنها را شامل شود، مشکل می‌باشد. روش‌های کاوش پیکره‌های تطبیقی، همانند روش‌هایی که ما در این پژوهش ارائه داده‌ایم، می‌توانند برای استخراج دانش ترجمه برای کلمات یا عبارات جدید از پیکره‌های تطبیقی استفاده شوند و به فرهنگ نگاران[۲] برای بروز کردن واژه‌نامه‌ها کمک کنند. علاوه بر این می‌توان انتظار داشت که ترکیب نتایج استخراج شده از پیکره‌های تطبیقی با واژه‌نامه‌ای دو زبانه، کارآیی سیستم‌های بازیابی اطلاعات بین زبانی را بهبود دهد. بر این اساس با توجه به آزمایش‌هایی که بر روی ترکیب واژه‌نامه با ترجمه‌های استخراج شده از پیکره تطبیقی انجام داده‌ایم، با استفاده از شبه بازخورد ارتباطی میانگین متوسط دقت به ۶۱.۸٪ بازیابی تک زبانی رسیده است.

در کل نتایج آزمایش‌هایمان در این پژوهش نشان می‌دهند که ترجمه‌های استخراج شده از پیکره تطبیقی روشی مؤثر برای بازیابی اطلاعات بین زبانی هستند و می‌توانند با ترجمه‌ی کلماتی که در واژه‌نامه‌ها نیستند و همچنین با استخراج کلمات مرتبط به آنها به منظور گسترش پرس‌وجوها، دقت بازیابی را بهبود دهند.

---

[۱] outliers

[۲] lexicographer

ساختار پایان‌نامه

ساختار پایان‌نامه به شرح زیر می‌باشد: در فصل ۲ به منظور آشنايي و ورود به بحث به معرفی روش‌های بازیابی اطلاعات و اهمیت پیکره‌های تطبیقی در این زمینه می‌پردازیم. همچنین روش‌های مختلف ساخت و استفاده از این نوع پیکره‌ها را بررسی می‌کنیم. سپس در فصل ۳ به معرفی روش ساخت پیکره تطبیقی فارسی-انگلیسی به طور کلی پرداخته و پیکره‌های مختلفی را بر اساس روش‌های گوناگون ساخته و آنها را ارزیابی می‌کنیم. در فصل ۴ روش‌های ارائه شده برای بازیابی اطلاعات بین زبانی با استفاده از پیکره‌های تطبیقی را معرفی می‌کنیم. به این منظور به طور کلی دو دسته مسأله اصلی استخراج دانش ترجمه و ساخت پرس‌وجوی ترجمه شده را بررسی می‌کنیم. در نهایت در فصل ۵ به تحلیل نتایج آزمایش‌های انجام شده پرداخته خواهد شد و در فصل ۶ نیز با ارائه جمع‌بندی و نتیجه‌گیری از فصل‌های قبل، کارهای آتی در این زمینه را پیشنهاد می‌کنیم.

فصل ۲

بازیابی اطلاعات بین زبانی

هدف اصلی در بازیابی اطلاعات بین زبانی[۱] پیدا کردن اطلاعات در یک زبان (همانند انگلیسی) در پاسخ به پرس‌وجوهایی به زبان دیگر (همانند فارسی) است. امروزه با رشد وسیع فناوری‌های وب[۲] و افزایش روز افزون حجم اطلاعات الکترونیکی با زبان‌های گوناگون، بازیابی اطلاعات، مستقل از زبان اسناد، اهمیت ویژه‌ای پیدا کرده است [۵۸]. به عنوان مثال از سال ۲۰۰۰ تا ۲۰۱۰، استفاده از اینترنت در جهان ۴۴۴.۸ % افزایش یافته است. به ویژه در خاور میانه، آمریکای لاتین و آفریقا بیش از ده برابر شده است و کماکان در آسیا و آفریقا این آمار در حال افزایش است[۳]. طبق آمار بدست آمده تا اواخر سال ۲۰۱۰ در مورد حجم اطلاعات موجود در اینترنت به زبان‌های گوناگون، ده زبان برترِ اینترنت تنها %۸۲.۲ حجم اطلاعات درون این شبکه جهانی را به خود اختصاص داده‌اند. نکته جالب توجه این است که اگر چه زبان انگلیسی همچنان پر استفاده‌ترین زبان در اینترنت است ولی تنها %۲۷.۳ از کل استفاده این شبکه جهانی را دارد. لذا می‌توان نتیجه‌گیری کرد که امروزه، به طور روز افزون، نیاز کاربران برای دسترسی به اطلاعات از مرز زبان‌ها گذر خواهد کرد و بدون پشتیبانی سیستم‌های بازیابی اطلاعات بین زبانی، اطلاعات هنگفتی غیر قابل دستیابی خواهد بود. اگرچه موتورهای جستجو قسمت مهمی از ساختار کنونی اینترنت شده‌اند، ولی بیشتر آنها تنها از جستجوی اطلاعات بر اساس پرس‌وجو و اسنادی در یک زبان پشتیبانی می‌کنند. در دهه گذشته، اکثر فعالیت‌های انجام شده بر روی بازیابی

---

[۱] Cross-Language Information Retrieval (CLIR)

[۲] web

[۳] http://www.internetworldstats.com

اطلاعات بین زبانی به پژوهش‌هایی در آزمایشگاه‌ها محدود بوده است [۳۳].

در این فصل به معرفی روش‌های مختلف بازیابی اطلاعات بین زبانی و پژوهش‌هایی که در این زمینه انجام شده است، می‌پردازیم. به طور کلی سه روش معمول برای ترجمه در چنین سیستم‌هایی مطرح است: ترجمه پرس‌وجو، ترجمه اسناد و ترجمه توأم پرس‌وجو و اسناد. معمول‌ترین منابع ترجمه‌ای که به این منظور استفاده می‌شوند، واژه‌نامه‌ها، پیکره‌های موازی، ماشین‌های ترجمه و پیکره‌های تطبیقی می‌باشند. ماشین‌های ترجمه، واژه‌نامه‌ها و پیکره‌های موازی منابع پرهزینه‌ای هستند و برای تعداد زیادی از جفت زبان‌های اقلیت وجود ندارند. این در حالی است که پیکره‌های تطبیقی را می‌توان بسیار راحت‌تر فراهم کرد. پیکره‌های تطبیقی، اسنادی در دو زبان متفاوت هستند که موضوعات مشابهی را می‌پوشانند، به عنوان مثال اخبار منتشر شده بر روی وقایع مشترک در دو زبان مختلف را می‌توان نوعی پیکره تطبیقی دانست. برترین مزیت این گونه از منابع در دسترس بودن آنها است و معمولاً بین هر دو زبانی وجود دارند. بنابراین اکتساب و استفاده از پیکره‌های تطبیقی به عنوان زمینه معتبری در بازیابی اطلاعات بین زبانی مطرح است. در ادامه مطالب این فصل به بررسی روش‌های مختلف ساخت و استفاده از این گونه پیکره‌های می‌پردازیم.

۱.۲ روش‌های بازیابی اطلاعات بین زبانی

پژوهش‌های ابتدایی در زمینه تحقیقاتی بازیابی اطلاعات بین زبانی از سال ۱۹۶۹ توسط سلتون[۱] [۷۶] و تحقیقات اصولی در این زمینه از اواخر دهه ۹۰ آغاز شده است. (پژوهش‌های ابتدایی در کتاب مرجع [۲۹] جمع‌آوری شده است).

فرآیند فعالیت‌های موجود در بازیابی اطلاعات بین زبانی کاملاً مشابه فرآیند بازیابی اطلاعات است، با این تفاوت که در آن به نوعی باید از مرز بین زبان‌ها گذشت. به این منظور ساده‌ترین راه‌هایی که در اولین مرحله به ذهن می‌رسند، ترجمه پرس‌وجو به زبان اسناد (*زبان مقصد*) یا ترجمه اسناد به زبان پرس‌وجو (*زبان مبدأ*) است. روش اول در ترجمه پرس‌وجو به زبان مقصد در بازیابی اطلاعات بین زبانی متداول‌تر است، به این علت که عموماً پرس‌وجوها بسیار کوتاه‌تر از مستندات موجود در زبان مقصد می‌باشند و لذا ترجمه آنها به زبان مقصد به مراتب ساده‌تر است. همچنین در ترجمه پرس‌وجو به زبان مقصد نیاز به دانستن دانش نحوی زبان وجود ندارد که امکان استفاده از الگوریتم‌ها و منابع ساده برای ترجمه را فراهم می‌آورد. علاوه بر این اگر بخواهیم مستندات موجود در زبان مقصد را به زبان مبدأ ترجمه کنیم، لازم است که مستندات مجدداً نمایه‌گذاری شوند که این امر باعث پیچیده شدن عملیات بازیابی اطلاعات بین زبانی می‌گردد.

---

[۱] Salton

مورد دیگر اینکه ترجمه مستندات به زبان مبدأ بر خلاف ترجمه پرس‌وجو می‌تواند به صورت برون‌خط[1] انجام شود [۳۸]. در مقابل ترجمه پرس‌وجو دارای نقاط منفی نیز هست، از جمله اینکه ترجمه پرس‌وجو به علت خلاصه بودن پرس‌وجوها و عدم وجود دانش متن به احتمال خیلی زیاد دارای ابهام خواهد بود. در این پژوهش ما بر روی ترجمه پرس‌وجو برای بازیابی اطلاعات بین زبانی تمرکز کرده و در این بخش به بررسی روش‌های مختلف آن می‌پردازیم.

با این رویکرد، بازیابی اطلاعات بین زبانی را می‌توان همانند بازیابی اطلاعات عادی قلمداد کرد که مراحلی بیشتر در فاز پردازش پرس‌وجو دارد. بازیابی اطلاعات بین زبانی به ویژه برای کاربرانی که مهارت کمی در زبان مقصد دارند، مفید است. به عنوان مثال، ممکن است کاربری بتواند متنی را به زبانی بخواند ولی نتواند پرس‌وجوی مناسب را در آن زبان ایجاد کند.

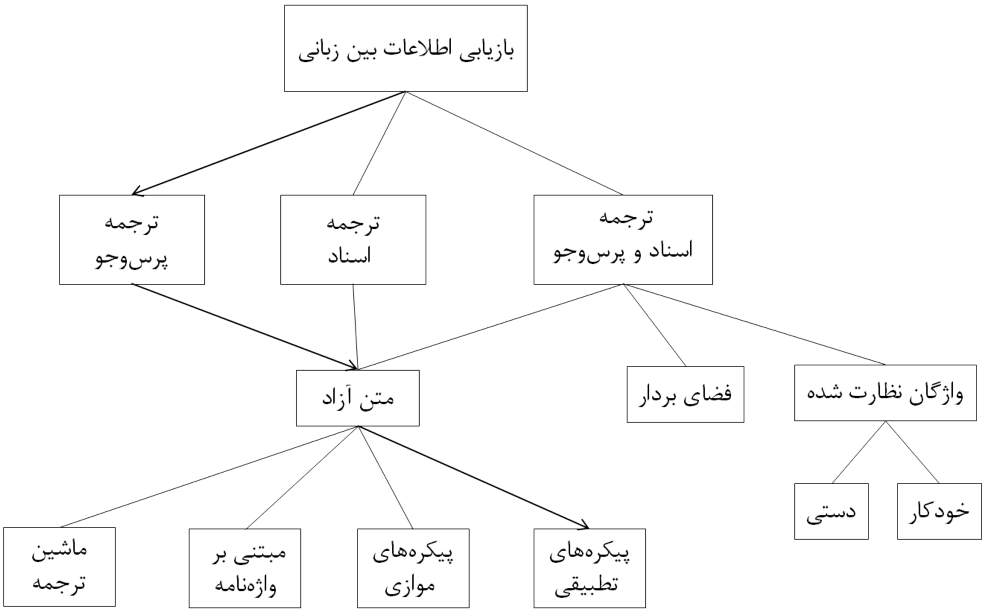
در مرجع [۴] نگارندگان آزمایش‌هایی بر روی کاربرانی با سطوح مهارتی متفاوت در زبان مقصد انجام داده‌اند و پی برده‌اند که جستجوی وب بین زبانی مبنی بر ترجمه پرس‌وجو به ویژه برای کاربرانی که مهارت کمی در زبان مقصد دارند بسیار مفید است. این کاربران نتایج بازیابی بهتری را از پرس‌وجوهای ترجمه شده نسبت به وقتی که خودشان پرس‌وجوها را مستقیماً به زبان مقصد وارد کرده‌اند، بدست آورده‌اند. اگرچه کیفیت منابع ترجمه (که در این مثال واژه‌نامه بود) نقش مهمی را در کارآیی نتایج بدست آمده ایفا می‌کند.

در ادامه این بخش روش‌های مختلف ترجمه پرس‌وجو که در بازیابی اطلاعات بین زبانی به کار می‌روند را بررسی می‌کنیم. شکل ۱.۲ نمایی از روش‌های کلی بازیابی اطلاعات بین زبانی را نشان می‌دهد. ما در این پایان‌نامه به بررسی روش‌های متن آزاد[2] می‌پردازیم. همچنین لازم به ذکر است که مسیری که در این پایان‌نامه برای حل مسأله بازیابی اطلاعات بین زبانی پیموده‌ایم در شکل با پیکان نشان داده‌ایم. همان گونه که این شکل نشان می‌دهد، به طور کلی روش‌های اصلی ترجمه، روش‌های مبنی بر واژه‌نامه، ماشین‌های ترجمه و روش‌های مبنی بر پیکره (پیکره‌های موازی و تطبیقی) هستند [۵۸، ۳۸] که در ادامه به بررسی مختصر این روش‌ها پرداخته‌ایم.

---

[1] offline

[2] free text



شکل ۱.۲: روش‌های اصلی بازیابی اطلاعات بین زبانی [۳۳]

۱.۱.۲ روش‌های مبنی بر واژه‌نامه

در روش بازیابی اطلاعات بین زبانی مبنی بر واژه‌نامه از واژه‌نامه‌های دوزبانه ماشین‌خوان[۱] به منظور جایگزینی کلمات پرس‌وجو به زبان مبدأ با مترادف‌هایشان در زبان مقصد استفاده می‌شوند. در ادامه فهرستی از چالش‌های مطرح در استفاده از واژه‌نامه را در سیستم‌های بازیابی اطلاعات بین زبانی آورده‌ایم. لازم به ذکر است که بیشتر این چالش‌ها برای تمام روش‌های بازیابی اطلاعات بین زبانی نیز صدق می‌کند.

۱. کلمات خارج از واژگان[۲]. هیچ واژه‌نامه‌ای کامل نیست. به ویژه کلماتی فنی، اسامی خاص و کلمات جدید را نمی‌توان اغلب در واژه‌نامه‌ها پیدا کرد.

۲. وجود ابهام در کلمات زبان مبدأ و مقصد. معنای دقیق کلمات را تنها می‌توان از طریق متنی که در آن آمده‌اند، مشخص کرد. به عنوان مثال، کلمه انگلیسی goal را در نظر بگیرید که هم می‌تواند یک واژه ورزشی باشد و هم معنای هدف و مقصد داشته باشد. ابهام کلمات تا حد زیادی باعث افزایش پیچیدگی مرحله ترجمه می‌شود.

۳. صرف کلمات زبان مبدأ. کلمات پرس‌وجو در زبان مبدأ معمولاً با صرف‌های مختلفی ظاهر می‌شوند، در حالی که واژه‌نامه‌ها معمولاً تنها شکل پایه‌ای هر کلمه را دارند.

---

[۱] machine readable

[۲] Out Of Vocabulary (OOV)

۴. تشخیص و ترجمه عبارات. معمولاً ترجمه عبارات در واژه‌نامه‌ها وجود ندارد. حتی اگر ترجمه عبارات در واژه‌نامه‌ها موجود باشند، باید اول آنها را بتوان به طریقی در پرس‌وجوها تشخیص داد و استخراج کرد.

دو چالش آخر در بازیابی تک‌زبانی نیز مطرح هستند و می‌توان برای بر طرف کردن آنها در بازیابی اطلاعات بین زبانی از روش‌های مشابه در بازیابی تک‌زبانی استفاده کرد. برای مثال، کلمات پرس‌وجو می‌توانند پیش از استفاده از واژه‌نامه ریشه‌یابی شده و یا تبدیل به بن‌واژه شوند. به طور مشابه، تشخیص عبارات نیز می‌تواند پیش از ترجمه انجام شود. دو چالش اول، کلمات خارج از واژگان و ابهام ترجمه، چالش‌های مطرح در بازیابی اطلاعات بین زبانی هستند و می‌توان گفت که دلایل اصلی کارآیی پایین سیستم‌های بازیابی اطلاعات بین زبانی نسبت به بازیابی اطلاعات تک‌زبانی می‌باشند. روش ترجمه مبنی بر واژه‌نامه را می‌توان به عنوان خط بنیادی[1] بازیابی اطلاعات بین زبانی در نظر گرفت. در بخش‌های آینده مروری بر روش‌های حل مشکلات ترجمه مبنی بر واژه‌نامه و به طور کلی روش‌های کلی بازیابی اطلاعات بین زبانی انجام خواهد شد.

تطابق واژه‌های هم‌ریشه

مشکل کلمات خارج از واژگان با اعمال روش‌های تطابق واژه‌های هم‌ریشه[2] می‌تواند تا حد زیادی بر طرف شود [۳۳]. بسیاری از اسامی خاص و کلمات فنی، که به احتمال زیاد در واژه‌نامه یافت نمی‌شوند، در زبان‌های مختلف مشابه هستند. اغلب آنها هم‌ریشه[3] هستند و دنباله‌ای از حروفشان صدای مشابهی دارد. به ویژه برای زبان‌هایی که حروف پایه یکسان دارند این مورد بیشتر مشهود است. مثلاً دو کلمه انگلیسی information و فنلاندی informaatio ریشه مشابهی دارند [۳۲]. از روش‌های تطابق واژه‌های هم‌ریشه استفاده می‌شود تا دو کلمه که در دو زبان تلفظ مشابهی دارند شناسایی شود.

---

[1] baseline

[2] cognate matching

[3] cognate

۲.۱.۲ ماشین ترجمه

هدف از ماشین ترجمه[1] ایجاد ترجمه‌ای روان از متون زبان طبیعی است. واضح است که این هدف از ترجمه پرس‌وجو مشکل‌تر است، زیرا پرس‌وجوها می‌توانند کلمه به کلمه ترجمه شوند و لازم نیست کاربر ترجمه‌ها را ببیند. با این وجود ماشین ترجمه مشکلاتی نیز دارد، اول اینکه این روش بر این فرض استوار است که ورودی یک واحد منظم طبق قواعد صرف و نحوی است، همانند جملات اخباری "*به دنبال اطلاعاتی درباره نظریه گفتگوی تمدن‌ها و نقش رئیس جمهور خاتمی می‌گردم*" و یا یک سوال "*نقش رئیس جمهور خاتمی در نظریه گفتگوی تمدن‌ها چه بود؟*" . ماشین‌های ترجمه برای اینکه محتمل‌ترین ترجمه را انتخاب کنند تا حدود زیادی بر روی اطلاعات زبان‌شناختی متن تکیه می‌کنند. مشکل اینجاست که بیشتر کاربران پرس‌وجوی خود را به صورت یک جمله با قاعده نمی‌نویسند و بیشتر پرس‌وجوها به صورت لیستی از کلمات کلیدی همانند "*رئیس جمهور خاتمی، گفتگوی تمدن‌ها*" و بدون توجه به هیچ‌گونه ترتیب کلمات و یا محدودیت‌های نحوی دیگر ایجاد می‌شوند. از جمله کاستی‌های دیگر بیشتر ماشین‌های ترجمه، به ویژه سیستم‌های تجاری، برای استفاده در بازیابی اطلاعات بین زبانی این است که آنها تنها محتمل‌ترین ترجمه را برمی‌گردانند، در حالی که محتمل‌ترین ترجمه بر اساس الگوریتم داخلی ماشین ترجمه تعریف شده است و این بدین معنا نیست که ترجمه‌های خوب دیگری - و یا حتی از منظر دیگری ترجمه‌های بهتری- وجود نداشته است [۵۲].

۳.۱.۲ روش‌های مبنی بر پیکره

در روش‌های مبنی بر پیکره، دانش ترجمه می‌تواند از پیکره‌های موازی و یا تطبیقی استخراج شود. در ادامه به طور مختصر به ارائه پیکره‌های موازی موجود و کاستی‌های استفاده از آنها برای بازیابی اطلاعات می‌پردازیم. همچنین پیکره‌های موجود زبان فارسی را معرفی می‌کنیم. لازم به ذکر است از آنجایی که عمده تمرکز این پژوهش بر استفاده از پیکره‌های تطبیقی می‌باشد، این نوع از پیکره‌ها را در بخش آینده به طور کامل‌تری بررسی می‌کنیم.

پیکره‌های موازی

مناسب‌ترین منبع ترجمه برای سیستم‌های بازیابی اطلاعات بین زبانی پیکره‌های موازی هستند [۳۳]. یک پیکره موازی شامل زوج اسنادی در دو زبان است که کاملاً ترجمه یکدیگر می‌باشند. پیکره‌های موازی زیادی ساخته دست بشر از

[1] Machine Translation (MT)

منابع گوناگون وجود دارد. برای مثال قوانین سازمان ملل متحد به چندین زبان موجود است [۱۶]. اتحادیه اروپا نیز اسناد رسمی‌اش را به چندین زبان اروپایی منتشر کرده است [۴۷]. اسناد مجلس کانادا به دو زبان فرانسه و انگلیسی موجود است. در هنگ‌کنگ اخبار و قوانین به چینی و انگلیسی هستند. از کتاب‌های مذهبی همانند انجیل نیز به عنوان پیکره‌های موازی استفاده شده است . همچنین پیکره‌های موازی می توانند به طور خودکار از وب ساخته شوند [۶۹].

طبیعتاً نبود برخی از واژگان در روش ترجمه مبنی بر پیکره موازی به عنوان یک مشکل اصلی مطرح است. به طور مثال، نمی‌توان واژگان ورزش فوتبال را بر اساس پیکره موازی که شامل قوانین اتحادیه اروپا است ترجمه کرد. اگرچه مجموعه‌های مستندات موازی ساخت بشر به تعداد زیادی وجود دارند، ولی این مجموعه‌ها عموماً کمبودهایی در پوشش حوزه‌های خاص دارند و البته همیشه نیز در دسترس نیستند. این در حالیست که برترین مزیت پیکره‌های تطبیقی در دسترس بودن آنها است. پیکره‌های تطبیقی، اسنادی در دو زبان متفاوت هستند که موضوعات مشابهی را می‌پوشانند. به عنوان مثال اخبار منتشر شده درباره‌ی رویدادهای مشترک، در دو زبان مختلف، ممکن است کاملاً ترجمه یکدیگر نباشند، ولی موضوعات مشابهی دارند. اگر چه ممکن است منابع زبان‌شناسی غنی همانند واژه‌نامه‌ها و یا پیکره‌های موازی بین دو زبان نداشته باشیم، ولی معمولاً پیکره‌های تطبیقی بین آن دو زبان وجود دارد. در نتیجه برترین مزیت این گونه از منابع در دسترس بودن آنها است. بنابراین، ایجاد و استفاده از پیکره‌های تطبیقی به عنوان زمینه‌ی ارزشمندی در بازیابی اطلاعات بین زبانی مطرح است.

در ادامه این بخش به معرفی روش‌های مختلف ساخت و استفاده از پیکره‌های تطبیقی به طور خاص در بازیابی اطلاعات بین زبانی می‌پردازیم.

۲.۲ پیکره‌های تطبیقی

پیکره تطبیقی یکی از موضوع‌های تحقیقاتی چالش‌انگیز در حوزه‌های ترجمه ماشین، زبان‌شناسی مبنی بر پیکره و به طور کلی پردازش زبان‌های طبیعی است. به عنوان مثال می‌توان متونی را که در یک دسته و حوزه یکسان و در یک بازه زمانی مشترک [۴۵] در بیش از یک زبان و یا زبان‌های مختلف [۸۱] جمع‌آوری شده‌اند و اطلاعات مشترکی دارند را به عنوان پیکره تطبیقی در نظر گرفت [۵۵، ۳۵].

نمونه‌هایی از پیکره‌های تطبیقی در ادامه آورده شده‌اند:

- پیکره‌های تطبیقی انگلیسی-فنلاندی-سوئدی در حوزه اخبار (ساخته شده در دانشگاه تمپر[1]):

۱. به تعداد ۱۳۱۴۲ هم‌ترازی اسناد برای ۷۲۲۶۰ سند در مجموعه اخبار سوئدی[2] و اخبار انگلیسی[3] [۸۶]. هر سند سوئدی تنها با یک سند انگلیسی هم‌تراز شده است.

۲. به تعداد ۱۲۹۴۵ هم‌ترازی اسناد برای ۵۵۲۹۸ سند در مجموعه اخبار فنلاندی[4] و اخبار سوئدی [۸۵].

۳. دو پیکره تطبیقی در حوزه ژنومیک بین زبان‌های آلمانی-انگلیسی و سوئدی-انگلیسی با هم‌ترازی در سطح پاراگراف [۸۷].

۴. به تعداد ۱۲۵۷۹ هم‌ترازی اسناد برای ۷۲۲۶۰ سند در مجموعه اخبار سوئدی و اخبار انگلیسی خبرگزاری لس آنجلس [۸۴]. هر سند سوئدی می‌تواند با چندین سند انگلیسی هم‌تراز شده باشد.

- مجموعه تطبیقی از اسناد چند زبانه اخبار مالی تعاونی‌ها در اوایل دهه ۱۹۹۰ میلادی به شش زبان اروپایی: هلندی (۸.۵ میلیون کلمه)، انگلیسی(۳۰ میلیون کلمه)، فرانسوی (۱۰ میلیون کلمه)، آلمانی (۳۳ میلیون کلمه)، ایتالیایی (۱.۸۸ میلیون کلمه) و اسپانیایی (۱۰ میلیون کلمه).

- پیکره تطبیقی بلغاری-کرواتی در حوزه اخبار [۹]: ۳،۵۰۰،۰۰۰ علامت[5] (۳۹۳۰۰۰ کلمه بلغاری و ۳۱۰۰۰۰۰ کلمه کرواتی) ساخته شده از زیر مجموعه‌هایی از دو پیکره روزنامه بزرگ‌تر به زبان‌های بلغاری و کرواتی، بر اساس انتخاب متون با معیارهایی همانند سال یکسان و حوزه یکسان.

- پیکره تطبیقی انگلیسی-فرانسوی-نروژی در حوزه علم (متون دانشگاهی)، ۴۵۰ مقالات علمی بازبینی شده، ۳.۲ میلیون کلمه [۲۱].

- پیکره تطبیقی چینی-انگلیسی در حوزه اخبار. به تعداد ۲۳۱۰۲ هم‌ترازی برای ۱۷۱۸۴۵ خبر به زبان چینی و ۱۰۳۶۵۷ خبر به زبان انگلیسی از وب‌سایت خبری XinHuaNet[6] [۳۶].

---

[1] University of Tampere

[2] خبرگزاری Tidningarnas Telegrambyra

[3] خبرگزاری Los Angeles times

[4] خبرگزاری Aamulehti

[5] token

[6] http://www.xinhuanet.com

این پیکره‌های تطبیقی هر یک محدود به زبان‌های خاص و بر روی حوزه‌های معینی هستند. از آنجایی که متون اکثراً تنها بر اساس یک معیار یعنی موضوع انتخاب شده‌اند، درجه تطبیق‌پذیری این پیکره‌ها به طور قابل توجهی با یکدیگر متفاوت است.

تحقیقات بر روی پیکره‌های تطبیقی در حدود ۱۵ سال پیش با پژوهش‌های انجام شده در مراجع [۶۶، ۲۳] برای استخراج ترجمه کلمات از پیکره‌های موازی دارای اختلال شروع شده است. نگارندگان فرض کرده‌اند که اندازه داده‌های آموزش بر روی کارآیی ترجمه‌های انجام شده تأثیر می‌گذارد و یک پیکره تطبیقی می‌تواند کمبودهای پیکره‌های موازی را جبران کند. این فرضیه توسط پژوهش‌هایی که اخیراً انجام شده‌اند نیز تایید شده است [۵۶، ۴۴، ۳۵، ۲۸].

در ادامه این بخش به مطالعه و بررسی روش‌های ساخت پیکره‌های تطبیقی پرداخته و چگونگی استخراج ترجمه کلمات از این گونه پیکره‌ها را بررسی خواهیم کرد. همچنین روش‌های استفاده از ترجمه کلمات را در حوزه‌های مختلف از جمله بازیابی اطلاعات بین زبانی مورد مطالعه قرار می‌دهیم.

۱.۲.۲ ساخت پیکره‌های تطبیقی

پیکره‌های تطبیقی عموماً از اخبار [۵۶، ۱۴، ۸۳، ۹]، کتاب‌های رمان [۱۷]، پیکره‌های تحقیقاتی موجود تک‌زبانه همانند مجموعه‌های CLEF و یا TREC [۸۶، ۱۲، ۸۰] یا با جمع‌آوری صفحات و خزش وب [۸۷، ۷۸، ۹۳] ساخته می‌شوند. مسأله اصلی مطرح در ساخت پیکره‌های تطبیقی، هم‌ترازی اسناد زبان مبدأ با اسناد مربوط در زبان مقصد می‌باشد. از جمله پژوهش‌های اولیه‌ای که در زمینه ساخت خودکار پیکره تطبیقی و هم‌ترازی اسناد آن انجام شده مراجع [۱۲] و [۷۹] است. برای مثال در مرجع [۷۹] نگارندگان از توصیفات محتوا و تاریخ انتشار برای هم‌ترازی اخبار آلمانی و ایتالیایی منتشر شده توسط خبرگزاری سوئیس، استفاده کرده‌اند. اسناد این خبرگزاری، برچسب‌های توصیفی برای محتوای موضوعی اخبار (همانند مالی، فرهنگی و نظامی) و قسمتی از جهان که آن خبر مربوط به آنجاست (همانند آفریقا، آلمان، ایالات متحده آمریکا) داشتند. در پژوهش انجام شده، اسناد بر اساس تاریخ انتشارشان و محتوای توصیفاتشان هم‌تراز شده‌اند. به عنوان مثال یک سند آلمانی که در ۲۴ آگوست ۱۹۹۴ انتشار یافته و درباره موضوعی نظامی است، با سندی ایتالیایی که در همان روز منتشر شده و محتوای توصیفی نظامی داشته، هم‌تراز شده است.

نگارندگان در مرجع [۱۲]، اسناد خبرگزاری سوئیس را بر اساس اسامی خاص مشترک، اعداد و تاریخ علاوه بر توصیفات محتوا هم‌تراز کرده‌اند. همچنین آنها اخبار انگلیسی منتشر شده توسط خبرگزاری آسوشیتدپرس[۱] را با اخبار

[۱] Associated Press

آلمانی خبرگزاری سوئیس (که دو مجموعه با مبدأ کاملاً متفاوت هستند) هم‌تراز کرده‌اند. آنها کلمات بسیار عمومی و بسیار کمیاب را از اسناد آسوشیتدپرس حذف کرده و سپس کلمات باقیمانده را توسط فهرستی از کلمات دو زبانه که از منابع رایگان گوناگونی از اینترنت جمع‌آوری کرده بودند، ترجمه کرده‌اند. پرس‌وجوهای ترجمه‌شده بر روی مجموعه اسناد خبرگزاری سوئیس اعمال شده و سپس اسنادی در زبان مقصد که تاریخ انتشار نزدیکتری به سند مبدأ داشته و امتیاز شباهت بالایی داشتند به عنوان زوج هم‌ترازی‌ها انتخاب شده‌اند. همچنین حدآستانه‌ای برای امتیازها در نظر گرفته شده است که مشخص می‌کند که آیا جفت سند انتخاب شده مناسب است یا خیر.

در پژوهش متفاوت دیگری، اسناد موازی و تطبیقی از وب بدست آمده‌اند. در مرجع [۶۹] پیکره‌های موازی توسط تشخیص شباهت‌های ساختاری صفحات چند زبانه ساخته شده‌اند. عموماً هنگامی که متنی به چندین زبان در وب ارائه می‌شود، صفحاتی که ترجمه یکدیگر هستند، ساختار مشابهی (همانند پاراگراف‌ها، سربرگ‌ها[۱]، ابر پیوندها[۲] وغیره) دارند.

در مرجع [۸۹]، نگارندگان پیکره تطبیقی دو زبانه را بر اساس این مشاهده بدست آورده‌اند که کلماتی که ترجمه یکدیگرند و یا اینکه درباره موضوعی مشترک هستند، تمایل دارند که در پیکره تطبیقی در بازه‌های زمانی یکسان و یا مشابهی رخ دهند.

تلونساری در مرجع [۸۶] روشی بر مبنای بازیابی اطلاعات بین زبانی برای هم‌ترازی اسناد دو مجموعه با زبان‌های متفاوت ارائه کرده است. در روش آنها، ساخت پیکره تطبیقی، شامل دو مرحله اکتساب[۳] و هم‌ترازی[۴] است. در مرحله اول متون دو زبان خاص (به ویژه اخبار منتشر شده در آن دو زبان) با استفاده از خزش وب جمع‌آوری می‌شوند. در مرحله دوم برای هم‌ترازی اسناد دو مجموعه، ابتدا برای هر سند مبدأ، کلید واژگان[۵] آن (که بیان‌گر محتوای کلی سند می‌باشند) مشخص شده و سپس آن کلمات به زبان مقصد ترجمه می‌شوند. در نتیجه پرس‌وجویی به زبان مقصد ساخته شده است که نمایانگر سند مبدأ است. سپس اسناد زبان مقصد توسط پرس‌وجوی ساخته شده بازیابی می‌شوند. در نهایت، هم‌ترازی بین سند مبدأ و اسناد بازیابی شده از زبان مقصد بر اساس تاریخ اسناد و میزان شباهت آنها، که توسط الگوریتم‌های بازیابی محاسبه می‌شود، صورت می‌گیرد[۸۶] .

---

[۱] header

[۲] hyper links

[۳] acquisition

[۴] alignment

[۵] key words

مرجع [۸۷]، مطالعه جدیدتر دکتر تلونساری در زمینه ساخت خودکار پیکره تطبیقی برای یک حوزه خاص می‌باشد. در آن پژوهش، نگارنده آزمایش‌های خود را توسط خزش متمرکز بر روی حوزه ژنومیک انجام داده است، بدین صورت که ابتدا تعدادی از کلمات خاص آن حوزه به موتور جستجو داده می‌شوند و اسنادی بازیابی می‌شوند. سپس پاراگراف‌های آنها استخراج شده و اگر شباهت کلمات پرس‌وجوی آن حوزه خاص با پاراگراف‌ها بیشتر از حدآستانه‌ای باشد آنگاه، آن پاراگراف انتخاب شده و در پیکره قرار می‌گیرد تا با پارگرافی در زبان دیگر هم‌تراز شود. تفاوت پژوهش صورت گرفته در حوزه خاص بودن آن است و دیگر اینکه در حوزه‌هایی به غیر از اخبار نمی‌توان برای هم‌ترازی اسناد از تاریخ استفاده کرد. همچنین هم‌ترازی در مرجع [۸۷] در سطح اسناد نیست بلکه در سطح پاراگراف صورت گرفته است. دلایل استفاده از پاراگراف به جای کل صفحه این بوده است که ممکن است تنها قسمتی از صفحات مربوط باشد، همچنین ممکن است صفحات دارای مطالبی با چندین زبان مختلف باشند و یا ممکن است هر پاراگراف یک مضمون متفاوت را داشته باشد.

در پژوهش دیگری که اخیراً در مرجع [۴۳] صورت گرفته، نگارندگان به بررسی بهبود کیفیت پیکره‌های تطبیقی به منظور استخراج دانش ترجمه پرداخته‌اند. بدین منظور معیاری برای تطبیق‌پذیری تعریف کرده و توسط فرآیند ساخت تکراری کیفیت پیکره تطبیقی ساخته شده را بهبود داده‌اند.

۲.۲.۲ استخراج دانش ترجمه از پیکره‌های تطبیقی

مطالعات بسیاری بر روی کاوش پیکره‌های تطبیقی برای پیدا کردن منابع مفید به منظور استفاده در پژوهش‌های پردازش‌های زبان‌های طبیعی صورت گرفته است. ما در این پژوهش، تحقیقات انجام شده بر روی پیکره‌های تطبیقی را به صورت چارچوب سلسه مراتبی مدل می‌کنیم. سطوح مختلف شامل کلمات، عبارات، جملات و اسناد بوده و هر مطالعه انجام شده در یک سطح به بررسی و استخراج داده‌های موازی از پیکره تطبیقی می‌پردازد. در ادامه این بخش به بررسی کارهای انجام شده بر روی استخراج دانش ترجمه کلمات، عبارات و استخراج جملات موازی از پیکره‌های تطبیقی می‌پردازیم. لازم به ذکر است که استخراج اسناد موازی را در غالب هم‌ترازی اسناد در دو زبان به طور مختصر در بخش گذشته (۲.۲.۱) بررسی کرده‌ایم و از آنجایی که برای ترجمه پرس‌وجوها در بازیابی اطلاعات بین زبانی نیاز است که ترجمه کلمات موجود باشند، بیشتر تأکید در این پایان‌نامه بر روی روش‌های مختلف استخراج ترجمه کلمات از پیکره‌های تطبیقی است. همچنین لازم به ذکر است که بیشتر پژوهش‌های انجام شده در این زمینه بر روی زبان‌هایی (همانند انگلیسی) که دارای منابع زبان‌شناسی غنی (همانند واژه‌نامه‌های چند زبانه، ماشین‌های ترجمه و پیکره‌های موازی) می‌باشد، تمرکز داشته‌اند و کوشش کمتری بر روی زبان‌هایی با منابع کمتر (همانند فارسی) صورت گرفته است.

استخراج دانش ترجمه کلمات

اکثر مطالعات اولیه بر روی یادگیری ترجمه کلمات دوزبانه بر روی منابع تک‌زبانه تمرکز داشته‌اند [۶۶، ۲۵، ۲۳، ۶۷]. فرض کلی این پژوهش‌ها بر این اساس است که یک کلمه و ترجمه‌اش در یک موضوع مشابه در زبان‌های مختلف ظاهر می‌شوند، بنابراین از آمارهای رخداد همزمان می‌توان برای تشخیص ترجمه آنها استفاده کرد. نگارندگان در مرجع [۶۶] از جملات آلمانی و انگلیسی که به صورت بردارهای رخداد همزمان نشان داده شده‌اند، استفاده کرده و نشان داده‌اند که هنگامی که ترتیب کلمات در دو زبان یکسان باشند، آنگاه بردارها بیشترین شباهت را دارند. در مرجع [۲۳] نگارندگان زمینه[۱] را بر اساس تعداد کلمات متفاوت ظاهر شده در هر دو طرف یک کلمه معین، تعریف کرده‌اند. معیارهای شباهت گوناگون دیگری، همانند بسامد اصطلاح [۲] و عکس بسامد سند [۳] در مرجع [۲۵] مورد مطالعه قرار گرفته‌اند. همچنین استفاده از پیکره‌های تطبیقی به عنوان منبعی برای استخراج دانش ترجمه به طور گسترده در منابع [۶۱، ۲۲، ۷۳، ۲۳، ۲۵، ۶۶، ۸۸، ۸۹، ۳۱، ۹۱، ۹۴، ۴۱، ۵۳، ۵۴، ۶۸، ۵۴، ۵۹، ۸۷، ۸۶] مطالعه شده است. در حوزه‌ی استخراج ترجمه کلمات، کارهای انجام شده بر روی پیکره‌های تطبیقی عموماً بر مبنای این فرض هستند که کلماتی که ترجمه یکسانی در زبان‌های متفاوت دارند، توزیع‌های مشابهی هم باید داشته باشند. بر این اساس Picchi و Peters [۶۱]، Fung [۲۳] ، Franz [۲۲] و Sadat [۷۳] از نوعی واژه‌نامه دوزبانه و یا پایگاه داده واژگان[۴] دوزبانه علاوه بر پیکره‌های تطبیقی استفاده کرده‌اند.

پژوهش‌های محدودی بر روی استخراج دانش ترجمه تنها با استفاده از پیکره‌های تطبیقی انجام شده است. ما در اینجا به طور خاص دو روش ارائه شده در [۸۷] و [۸۹] را بررسی می‌کنیم. در این دو روش، هر کلمه به صورت برداری از بسامدها در اسناد هم‌تراز شده و یا نسبتی از وزن بسامدها نمایش داده می‌شود. سپس به منظور به دست آوردن مقدار همبستگی هر دو کلمه، شباهت بردارهایشان محاسبه می‌شود. این دو روش در محاسبه شباهت‌ها متفاوت می‌باشند. به طور کلی هر دو روش بر اساس این شهود استوار هستند که کلماتی که ترجمه یکدیگرند و یا تا حدودی با یکدیگر مرتبط هستند، تمایل دارند که در یک بازه زمانی در پیکره تطبیقی با هم رخ دهند.

پژوهش انجام شده در مرجع [۸۹]، از جمله کارهای نادر، در حوزه‌ی استخراج ترجمه کلمات است که از منابع

---

[۱] context

[۲] Term Frequency (TF)

[۳] Inverse Document Frequency (IDF)

[۴] lexical

زبان‌شناسی استفاده نشده و نگارندگان تنها از پیکره تطبیقی استفاده کرده‌اند. ایده اصلی این کار استخراج همبستگی‌های تکرار برای بدست آوردن ترجمه کلمات در زبان‌های متفاوت توسط پیکره‌های تطبیقی و استنتاج نگاشت‌های بین کلمات می‌باشد. دانش ترجمه در این مرجع بر اساس ضریب همبستگی پیرسون[۱] محاسبه شده است و سپس از ارتباطات کلمات دو زبان استفاده شده تا شباهت اسناد در دو زبان انگلیسی و چینی را محاسبه کنند. به طور کلی یک سند در زبان A با سندی در زبان B در صورتی منطبق می‌شوند که کلمات سند اول ارتباط قوی با کلمات زبان دوم داشته باشند. نگارندگان در آن مرجع گزارش داده‌اند که می‌توان از ترجمه‌های استخراج شده استفاده کرد و با درصد قابل قبولی اسناد مختلف را با آنها هم‌تراز کرد. به طور مشابه در مرجع [۷۷] ضریب همبستگی پیرسون برای محاسبه شباهت کلمات از پیکره تطبیقی در دو زبان انگلیسی و عربی استفاده شده است و سپس روش مدل زبانی پرس‌وجو به منظور استفاده از دانش ترجمه استخراج شده در بازیابی اطلاعات بین زبانی بکار گرفته شده است. اگر چه ایده عمومی استخراج همبستگی بسامدها برای بدست آوردن ترجمه کلمات از پیکره‌های تطبیقی در این پژوهش‌ها [۸۸، ۶۶، ۲۳] نیز آزموده شده است، ولی هیچ یک از آنها مقایسه مستقیمی از توزیع‌های تکرار کلمات کاندید انجام نداده‌اند، بلکه آنها گرایش به محاسبه ارتباطات بین کلمات در یک زبان یکسان داشته و سپس الگوهای وابستگی را بین زبان‌های مختلف مقایسه کرده‌اند.

اخیراً تلونساری در [۸۷، ۸۶] روشی برای استخراج ترجمه کلمات ارائه کرده است. از این روش (COCOT) به عنوان خط بنیادی[۲] در این پایان‌نامه استفاده کرده‌ایم و به طور کامل آن را در بخش ۱.۱.۴ توضیح داده‌ایم.

در مرجع [۳۱] نگارندگان ترجمه اسامی خاص را از پیکره‌های تطبیقی و موازی استخراج کرده‌اند. در آن پژوهش، نگارندگان از روش پیچیده‌ای بر اساس خوشه‌بندی کلمات برای هم‌ترازی اسناد پیکره تطبیقی استفاده کرده‌اند. اگرچه آنها به مساله چگونگی به دست آوردن پیکره تطبیقی اشاره‌ای نکرده‌اند.

در پژوهش جدیدتری [۴۲] نگارندگان یک مدل ترکیبی بدون سرپرست ارائه داده‌اند که از ویژگی‌های آماری، واژه‌ای، زبان شناختی و متنی در یک مدل مبنی بر EM[۳] برای کشف واژه‌های دوزبانه از پیکره‌های تطبیقی هم ترازشده استفاده می‌کند. همچنین در این پژوهش نشان داده شده است که مدل ارائه شده، تنها با استفاده از واژه‌نامه محدودی توانسته واژه‌های دوزبانه جدیدی را کشف کند و ترجمه‌های کشف شده کیفیت خوبی دارند.

---

[۱] Pearson correlation

[۲] baseline

[۳] Expectation Maximization

در مطالعات ارزشمند دیگری، دکتر سادات[1] مدل ترجمه دو مرحله‌ای مبنی بر پیکره ارائه کرده است [۷۳، ۷۵، ۷۴] که از ترجمه‌های کلمات مبدأ به مقصد و ترجمه کلمات مقصد به مبدأ استفاده می‌کند تا ترجمه‌ای دقیق‌تر بدست آورد. در پژوهش‌های جدیدتری که اخیراً این محقق انجام داده است [۷۲، ۷۱] همان مدل دو مرحله‌ای را تا حدودی گسترش داده و بهبود قابل توجهی کسب کرده است. به طور کلی مراحل این روش بدین صورت می‌باشد که در مرحله اول هدف این است که ترجمه کلمات در زبان مبدأ به زبان مقصد و همچنین از زبان مقصد به زبان مبدأ استخراج شوند. نتایج این دو لیست از ترجمه‌های استخراج شده به منظور ابهام زدایی با یکدیگر ترکیب می‌شوند. در مرحله دوم، ترجمه‌های استخراج شده بر اساس دانش ریخت‌شناسی[2] کلمات، تصفیه[3] می‌شوند. روش هرسی[4] نیز مبنی بر اطلاعات زبان شناختی به منظور مقایسه کلمات مبدأ و ترجمه‌های معادلشان در زبان مقصد بر اساس مقوله دستوری[5] کلمات اعمال کرده‌اند. شکل ۲.۲ مدلی که در این پژوهش‌های ارائه کرده‌اند را نشان می‌دهد. نگارندگان در آزمایش‌هایشان از پیکره ژاپنی-انگلیسی استفاده کرده و در نهایت مدل ترجمه‌ای ترکیبی بر اساس ترجمه‌های استخراج شده از پیکره‌های تطبیقی و واژه‌نامه‌های دو زبانه[6] برای بازیابی اطلاعات بین زبانی ارائه داده‌اند.

استخراج دانش ترجمه عبارات

یک عبارت، ساختار نحوی است که متشکل از چندین کلمه می‌باشد و ممکن است معنایی متفاوت از معنای تک تک اعضایش داشته باشد. معمولاً عبارات، ابهام کمتری نسبت به کلمات دارند، بنابراین استفاده از ترجمه عبارات در سیستم‌های بازیابی اطلاعات بین زبانی می‌تواند دقت بازیابی را افزایش دهد [۸]. ولی در پژوهش‌های انجام شده، ترجمه عبارات کمتر از ترجمه کلمات مورد توجه قرار گرفته است. این موضوع می‌تواند نشان دهنده‌ی این باشد که ترجمه عبارات به احتمال زیاد موفقیت‌آمیز نبوده‌اند [۳۳].

در بازیابی اطلاعات بین زبانی، اگر بتوان پرس‌وجوها را به صورت عباراتی در نظر گرفت و ترجمه‌های هر کلمه‌ی یکتا از آنها را به ترجمه‌هایی محدود کرد که از لحاظ آماری و زبان شناختی بیشتر به یکدیگر مرتبط می‌باشند، آنگاه ابهام

[1] Sadat

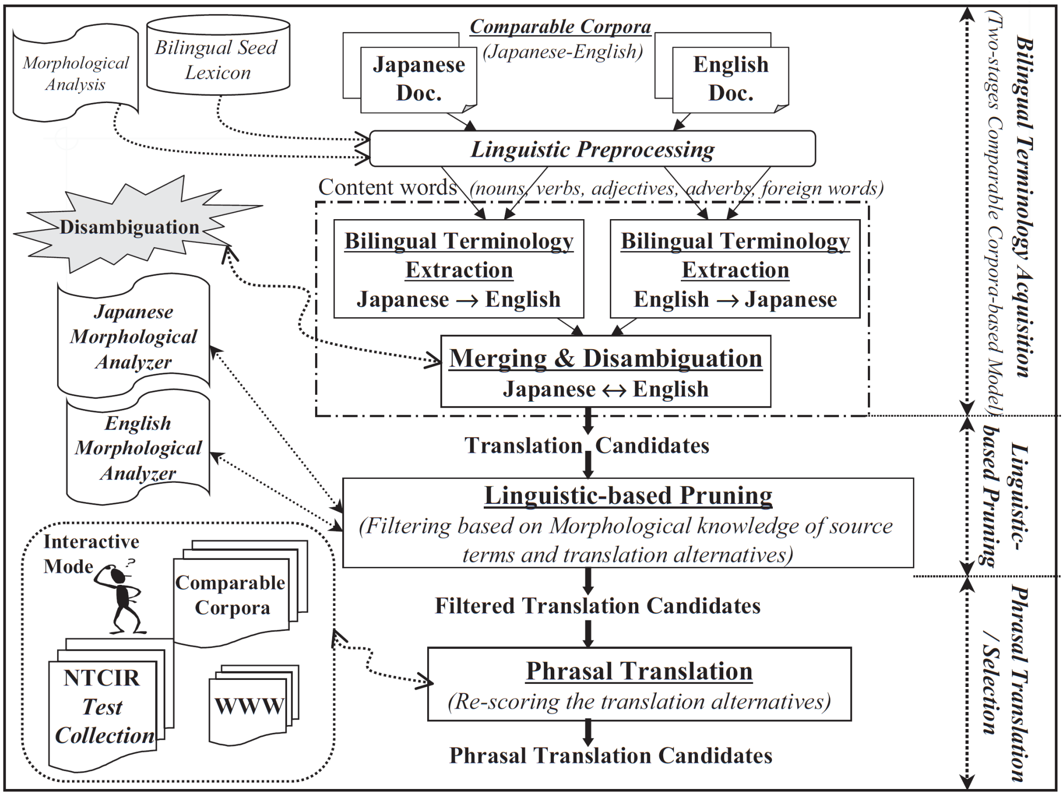
[2] morphological knowledge

[3] filter

[4] pruning

[5] part of speech

[6] bilingual



شکل ۲.۲: مدل پیشنهادی در مرجع [۷۳] به منظور استخراج ترجمه کلمات و عبارت

ترجمه بسیار کاهش می‌یابد [۷۴]. اولین مرحله برای استخراج ترجمه عبارات، تشخیص آنها می‌باشد. سه روشِ مبنی بر پیکره برای تشخیص عبارات وجود دارد: شناخت آماری، برچسب‌گذاری[۱] مقوله دستوری و تجزیه نحوی. در این بخش به بررسی مطالعات انجام شده بر اساس روش‌های آماری تشخیص و استخراج ترجمه عبارات از پیکره‌ها می‌پردازیم. به صورت کلی، روش‌های آماری از اطلاعات رخداد همزمان اصطلاحات استفاده می‌کند. در مرجع [۲۶]، نگارندگان برای تشخیص عبارات دو کلمه‌ای، ابتدا پیکره را به بایگرام‌هایی[۲] دارای اشتراک که قطعاتی شامل دو کلمه هستند، تقسیم کرده‌اند. سپس بر اساس بسامد اصطلاح و عکس بسامد سندشان، آنها را به صورت نزولی مرتب کرده‌اند. در نهایت، بایگرام‌هایی که بالاترین رتبه‌ها را کسب کرده‌اند به عنوان عبارات تشخیص داده شده‌اند.

در مرجع [۷۳]، سادات روشی را ارائه داده است که پس از تولید لیست مرتبی از ترجمه‌های محتمل یک کلمه با استفاده از پیکره تطبیقی، از روش‌های دوباره رتبه‌بندی برای تخمین انسجام پرس‌وجوی ترجمه شده استفاده می‌کند و بهترین ترجمه عبارات را انتخاب می‌نماید. این روش‌های دوباره رتبه‌بندی شامل رتبه‌بندی بر اساس وب جهان‌گستر،

---

[۱] tagging

[۲] bigram

مجموعه‌های اسناد، پیکره‌های تطبیقی و حالت تعاملی هستند. به طور کلی شهود سه روش اول بر این مبنا است که هرچه ترکیبی از ترجمه‌های کلمات به تعداد بیشتری در آن مجموعه (و یا وب) آمده باشند، آنگاه می‌توان آن ترکیب را به عنوان ترجمه‌ی مناسب‌تری از آن عبارت در نظر گرفت. در روش تعاملی نیز کاربر می‌تواند بهترین ترجمه را از لیستی که به او نشان داده می‌شود انتخاب کند. روش‌های ارائه شده توسط سادات به منظور استخراج دانش ترجمه عبارات، در مرحله آخرِ شکل ۲.۲ نمایش داده شده‌اند.

در مرجع [۴۸]، میانگاه روشی را برای تشخیص بهترین ترجمه کلمات بر مبنای همبستگی‌شان با کلمات دیگر ارائه داده است. در روش ارائه شده از اطلاعات متقابل[۱] کلمات در پیکره‌ای بزرگ به زبان فارسی استفاده می‌کنند. بدین صورت که هر چه اطلاعات متقابل دو کلمه بیشتر باشد، ارتباط کلمات قوی‌تر است و برای رفع ابهام ترجمه می‌توان بیشتر به آنها اعتماد کرد. در نتیجه بهترین ترجمه هر کلمه را بر اساس همبستگی با ترجمه‌های کلمات اطرافش در پیکره انتخاب کرده‌اند.

استخراج جملات و زیر جملات موازی

از جمله پژوهش‌های دیگری که بر روی پیکره‌های تطبیقی صورت گرفته، جستجو و استخراج جملات موازی از آنها است. پس از استخراج جملات موازی می‌توان به صورت دقیق‌تری ترجمه کلمات و عبارات را از آن جملات استخراج کرد. ولی مسأله چالش‌برانگیز پیدا کردن جملاتی از پیکره تطبیقی است که کاملاً ترجمه یکدیگر باشند. روش‌های مختلفی برای جستجوی جملات موازی از پیکره‌های تطبیقی ارائه شده است. در مراجع [۹۵، ۹۲] ابتدا اسناد موازی را پیدا کرده و سپس جملات موازی آنها را هم‌تراز می کنند. در مراجع [۵۵، ۲۴، ۲] بر روی اسناد غیر موازی کار کرده و فرض بر این است که تمام جملات سند مبدأ ممکن است ترجمه مشابهی در سند مقصد نداشته باشند. هر جمله مبدأ با تعدادی جمله‌ی مقصد منطبق شده و زوج جملاتی که بیشتر از حدآستانه شباهت دارند به عنوان جملات موازی در نظر گرفته شده‌اند. منتانو[۲] در مرجع [۵۵] از یک طبقه‌بندی مبنی بر بیشینه بی‌نظمی[۳] استفاده کرده و در مرجع [۲۴] از معیار شباهت کسینوسی استفاده شده و زوج جملات تشخیص داده شده به صورت تکراری به سیستم اضافه شده‌اند تا کیفیت جملات هم‌تراز شده بهبود یابد. همچنین در مرجع [۵۵] نگارندگان نشان داده‌اند که این روش می‌تواند برای

[۱] mutual information

[۲] Munteanu

[۳] maximum entropy

بهبود کارآیی سیستم‌های ترجمه ماشینی مفید باشد.

به نظر می‌رسد که بیشتر داده‌های موازی در پیکره‌های تطبیقی در سطح زیر جملات[۱] هستند، ولی تنها دو پژوهش در گذشته در این زمینه انجام شده است [۵۷، ۶۵]. منتانو در مرجع [۵۷] برای اولین بار روشی را برای تشخیص زیر جملات از جملات تطبیقی ارائه داده است. پس از تشخیص زوج جملاتی که حداقل تشابه را دارند توسط روشی الهام گرفته از پردازش سیگنال‌ها، به جستجوی قطعه‌های موازی در آنها می‌پردازند. هر کلمه با امتیازی که نشان دهنده حضور ترجمه معادلش در جمله دیگر است، علامت گذاری می‌شود. یک فیلتر میانگین متحرک بر روی مقادیر اعمال شده و با آن به صورت سیگنال رفتار می‌کند. قطعات موازی بر مبنای بخش‌هایی از سیگنال که مقادیر مثبت دارند، شناسایی می‌شوند. کیورک[۲] در مرجع [۶۵] این کار را با معرفی مدل‌های تولیدی[۳] که صریحاً زوج جملات تطبیقی را هم‌تراز می‌کنند گسترش داده است. آنها با اعمال این روش بر روی داده‌های آزمایشی بین حوزه‌ها، نشان داده‌اند که توانسته‌اند بهبود چشم‌گیری را در کارآیی ترجمه‌ها کسب کنند.

۳.۲.۲ استفاده از دانش ترجمه استخراج شده

از ترجمه‌های استخراج شده از پیکره‌های تطبیقی به عنوان منبع زبان‌شناختی به طور گسترده در زمینه بهبود ماشین‌های ترجمه [۱۸، ۵۵، ۱] و بازیابی اطلاعات بین زبانی [۸۶، ۷۵، ۷۲، ۷۱، ۶۱] استفاده شده است. ما در این پایان نامه بر روی استفاده از دانش ترجمه استخراج شده برای بازیابی اطلاعات بین زبانی تمرکز کرده و کارهای پیشین در این زمینه را بررسی می‌کنیم. به طور کلی در کارهای پیشین به منظور ترجمه پرس‌وجوها، بالاترین ترجمه‌های استخراج شده برای هر کلمه پرس‌وجو در زبان مبدأ را انتخاب کرده و پرس‌وجو را در زبان مقصد می‌سازند. در مرجع [۸۶] نگارندگان برای ساخت پرس‌وجوها از پرس‌جوی ساخت‌یافته ابزار Lemur[۴] بر اساس نحو پرس‌وجوی InQuery [۷] استفاده کرده‌اند. در این پایان‌نامه برای ساخت پرس‌وجوهای ترجمه شده به منظور بازیابی اطلاعات بین زبانی از روش مدل زبانی پرس‌وجوها استفاده کرده‌ایم. لازم به ذکر است که تنها روشی که کمی به روش ارائه داده شده‌ی ما شبیه است روش مطرح شده در مرجع [۹۰] می‌باشد که در آن برای آموزش مدل ترجمه در بازیابی اطلاعات پزشکی توسط پیکره استفاده

---

[۱] sub-sentence

[۲] Quirk

[۳] generative

[۴] www.lemurproject.org

کرده است و روش محاسبه امتیازهای کلمات پرس‌وجو بر اساس تخمین بیشینه‌ی درست‌نمایی[۱] می‌باشد که کاملاً با روش استخراج ترجمه ما متفاوت است. برای ساخت مدل زبانی پرس‌وجوها از روش ارائه شده در پایان‌نامه دکتر شاکری [۷۷] ایده گرفته و حالت‌های مختلف ترکیب این روش با واژه‌نامه را ارائه داده‌ایم.

مسأله دیگری که در مطالعات بازیابی اطلاعات مطرح شده است، بهبود دامنه لغات واژه‌نامه توسط پیکره‌های تطبیقی است. در پژوهش‌های [۸۶] و [۷۲]، نگارندگان روش‌هایی مبنی بر ترکیب ترجمه‌های استخراج شده با واژه‌نامه به منظور ترجمه پرس‌وجوها ارائه داده‌اند. ما هم در این پایان‌نامه آزمایش‌هایی بر روی روش‌های ترکیبی ترجمه‌های پیکره‌های تطبیقی با واژه‌نامه انجام داده‌ایم، اگرچه روش‌های ما هم در استخراج دانش ترجمه و هم در ساخت پرس‌وجوهای ترجمه شده از دیگر پژوهش‌های انجام شده متفاوت است. تلونساری در مرجع [۸۶]، ترجمه‌های استخراج شده از پیکره‌های تطبیقی را با نتایج ابزار ترجمه پرس‌وجوی مبنی بر واژه‌نامه (UTACLIR) که در دانشگاه تمپر ساخته شده است ترکیب کرده و پرس‌جوهای ترجمه شده را با فرمت InQuery ساخته است. در حالی که ما در روش‌های ترکیب با واژه‌نامه از یک واژه‌نامه ساده استفاده کرده و مدل زبانی پرس‌وجوها را ساخته‌ایم. در پژوهش دیگری (مرجع [۷۲]) نیز نگارندگان نتایج ترجمه‌های استخراج شده از پیکره تطبیقی را به صورت خطی با ترجمه‌های واژه‌نامه ترکیب کرده و پرس‌وجوهای ترجمه شده را ساخته‌اند. هر دو پژوهش گزارش داده‌اند که استفاده از پیکره‌های تطبیقی برای بازیابی اطلاعات بین زبانی تأثیرات نسبتاً خوبی داشته‌اند و به ویژه ترکیب ترجمه‌های بدست آمده از آنها باعث بهبود واژه‌نامه‌ها شده است.

۳.۲ پیکره‌های زبان فارسی

در حال حاضر پیکره‌هایی تک‌زبانه و به منظور خاص در زبان فارسی وجود دارند. به عنوان مثال پیکره‌ی همشهری [۵] یک پیکره‌ی تک زبانه است که برای ارزیابی سیستم‌های بازیابی اطلاعات فارسی استفاده می‌شود. پیکره بیژن‌خان [۱۱] نیز یک پیکره‌ی تک زبانه است که شامل برچسب[۲] دستور زبان کلمات به منظور استفاده در پردازش‌های زبان طبیعی می‌باشد. همچنین پیکره‌های موازی فارسی-انگلیسی بزرگی نیز برای استخراج دانش ترجمه در سال‌های اخیر ایجاد شده‌اند. پیکره‌های فارسی-انگلیسی موجود شامل پیکره‌ی موازی میانگاه [۵۰] با ۴۸۶۰۰۰۰ کلمه، پیکره‌ی موازی انگلیسی-فارسی تهران [۶۲] با ۶۱۲۰۸۶ جمله موازی محاوره‌ای که از زیر نویس‌های فیلم استخراج شده است

[۱] maximum likelihood

[۲] tag

و همچنین مجموعه ۷۰۷۳ سند فارسی-انگلیسی جمع‌آوری شده عابدین‌زاده از ویکی‌پدیا[۱] با حجم ۹۰ مگابایت [۳] می‌باشند. دیگر پیکره موجود، پیکره‌ی غیر موازی[۲] فارسی-انگلیسی کریمی[۳۷] می‌باشد که شامل ۱۱۰۰ سند اخبار بی‌بی‌سی فارسی و انگلیسی است. با این وجود از آنجایی که دانش ترجمه معمولاً به صورت آماری از پیکره‌های چند زبانه استخراج می‌شود، پیکره کریمی در زبان فارسی به اندازه کافی بزرگ نیست، لذا بر اساس مطالعات نگارنده به نظر می‌رسد هیچ‌گونه پیکره تطبیقی مناسبی در زبان فارسی وجود ندارد.

۴.۲ بازیابی اطلاعات بین زبانی فارسی و انگلیسی

مطالعات انجام گرفته روی بازیابی اطلاعات زبان فارسی در مقایسه با دیگر زبان‌ها بسیار کم بوده است و این در مورد بازیابی اطلاعات بین زبانی نمود بیشتری نیز دارد. یکی از دلایل اصلی که مطالعه روی زبان‌ها را محدود می‌کند، عدم وجود منابع کافی (همانند پیکره‌های استاندارد) مربوط به آن زبان‌ها است. البته بازیابی اطلاعات بین زبانی، نیاز به منابع زبانی دیگری (همانند پرس‌وجوها و اسناد قضاوت شده) هم دارد. مثلاً برای ارزیابی ترجمه‌ی پرس‌وجوها، از پیکره‌هایی با پرس‌وجوهای دو زبانه و اسناد قضاوت شده، استفاده می‌شود. پیکره همشهری، اولین پیکره‌ی استاندارد برای استفاده در پژوهش‌های بازیابی اطلاعات فارسی است. البته لازم به ذکر است که محققانی به طور مستقل پیکره‌هایی ساخته‌اند و در ادامه به بررسی این مطالعات انجام گرفته برای بازیابی اطلاعات بین زبانی می‌پردازیم.

پیکره همشهری توسط آزمایشگاه پژوهشی پایگاه داده دانشگاه تهران[۳] با کمک کنفرانس CLEF و پشتیبانی مرکز تحقیقات مخابرات ایران برای ایجاد و قضاوت پرس‌وجوهای دوزبانه ساخته شده است. این پیکره برای اولین بار در دو سال پیاپی در مسابقه بازیابی اطلاعات فارسی کنفرانس CLEF-2008, 2009 استفاده شده و گروه‌های مختلف، الگوریتم‌های بازیابی اطلاعات خود را بر روی زبان فارسی آزموده‌اند. در این بخش به بررسی کارهای منتشر شده در CLEF برای بازیابی اطلاعات بین زبانی بر روی پیکره همشهری می‌پردازیم. در CLEF-2008 دو گروه بر روی بازیابی اطلاعات بین زبانی فارسی-انگلیسی آزمایش‌هایی انجام داده‌اند. گروه اول از دانشگاه جان هاپکینز بوده و از ماشین ترجمه[۴] برای ترجمه پرس‌وجوها استفاده کرده‌اند [۴۶]. گروه دوم از دانشگاه تهران بوده و روشی احتمالی برای

[۱] www.wikipedia.org

[۲] non-parallel

[۳] UT DBRG

[۴] http://www.parstranslator.net/eng/translate.htm

استخراج ترجمه از واژه‌نامه ارائه کرده‌اند [۶]. در CLEF-2009 تنها یک گروه از دانشگاه آزاد قزوین آزمایش‌هایی بر روی بازیابی اطلاعات بین زبانی انجام داده است که از ویکی‌پدیا استفاده کرده و کارآیی بسیار کمی بدست آورده‌اند که دلیل آن را کم بودن اسناد فارسی در ویکی‌پدیا نسبت به دیگر زبان‌ها دانسته‌اند [۲۰].

پژوهش‌های مستقلی هم بر روی بازیابی اطلاعات بین زبانی صورت گرفته که نگارندگان خود پیکره‌ای برای ارزیابی الگوریتم‌های پیشنهادی ساخته‌اند. از مطالعاتی که اخیراً انجام شده می‌توان به پژوهش دکتر علیزاده [۱۹] اشاره کرد. در این پژوهش میزان تأثیر فنون پردازش‌های زبان طبیعی بر روی بازیابی اطلاعات بین زبانی سنجیده شده است. نگارندگان با استفاده از پرس‌وجوهای فارسی، اسناد انگلیسی مرتبط را بازیابی کرده‌اند. برای ترجمه پرس‌وجوهای فارسی از واژه‌نامه دو زبانه Farsidic[۱] استفاده شده است. همچنین از فنون پردازش زبان طبیعی همانند تحلیل ریخت‌شناسی[۲] و برچسب‌زنی مقوله دستوری در فازهای قبل و بعد از ترجمه استفاده کرده‌اند. به طور خاص نگارندگان از ریخت‌شناسی کلمات برای ترجمه کلماتی که در واژه‌نامه یافت نشده و از برچسب‌زنی نقش دستوری کلمات برای تشخیص عبارات درون پرس‌وجوها استفاده کرده‌اند. نتایج آزمایش‌های آنها نشان داده است که اعمال روش‌های پردازش زبان طبیعی در افزایش کارآیی بازیابی اطلاعات بین زبانی مؤثر است. لازم به ذکر است در پژوهش انجام شده، نگارندگان پس از ساخت چندین پرس‌وجو، از موتور جستجوی گوگل[۳] برای بازیابی اسناد مرتبط انگلیسی استفاده کرده‌اند و تنها تعداد محدودی (۱۰۰ سند) از این اسناد را بر اساس روش ایجاد مخزن[۴] قضاوت کرده‌اند [۱۹].

در پژوهش دیگری، نگارندگان از یک رویکرد معنایی[۵] برای بازیابی اطلاعات بین زبانی استفاده کرده‌اند [۶۴]. آنها با تاکید بر روی روش‌های مبنی بر واژه‌نامه (برای ترجمه پرس‌وجوها) حالت‌هایی را در نظر گرفته‌اند که ابهام ترجمه وجود دارد، یعنی کلمه‌ای در زبانی بیش از یک ترجمه در زبان دیگر دارد. همچنین برای بهبود ترجمه پرس‌وجو توسط واژه‌نامه از گسترش پرس‌وجو قبل و بعد از ترجمه استفاده کرده و با ارائه الگوریتمی احتمالی بهترین ترجمه را برای کلمات و عبارات انتخاب کرده‌اند. آزمایش‌هایی که انجام شده است بر مبنای بازیابی اسناد انگلیسی برای پرس‌وجوهای فارسی می‌باشد و نشان داده شده است که با استفاده از روش ارائه شده می‌توان استفاده از واژه‌نامه را برای بازیابی اطلاعات بین زبانی بهبود داد.

---

[۱] www.farsidic.com

[۲] morphological analysis

[۳] Google

[۴] pooling

[۵] semantic

مطالعاتی بر روی استخراج دانش ترجمه از پیکره‌های موازی برای بازیابی اطلاعات بین زبانی فارسی-انگلیسی توسط خانم میانگاه انجام شده است [۴۹، ۵۱]. در این پژوهش‌ها، نگارندگان معیارهایی آماری برای محاسبه وابستگی کلمات و عبارات بین دو زبان توسط پیکره‌های موازی فارسی-انگلیسی ارائه کرده‌اند. آنها تنها نمونه‌ای از ترجمه‌های استخراج شده را ارزیابی کرده‌اند و بازیابی اطلاعات بین زبانی را بر روی پیکره‌ی اسناد نیازموده‌اند.

ما در این پایان‌نامه بازیابی اطلاعات بین زبانی را با فرض داشتن پیکره‌های تطبیقی، بین دو زبان فارسی و انگلیسی بررسی می‌کنیم. لازم به ذکر است که تا کنون تحقیقاتی در زمینه پیکره‌های تطبیقی به منظور بازیابی اطلاعات بین زبانی در زبان فارسی صورت نگرفته است. همچنین در ادامه پژوهش نشان می‌دهیم که می‌توان واژه‌نامه‌ها را توسط ترجمه‌های استخراج شده از پیکره‌های تطبیقی بهبود داد.

فصل ۳

ساخت پیکره تطبیقی فارسی-انگلیسی

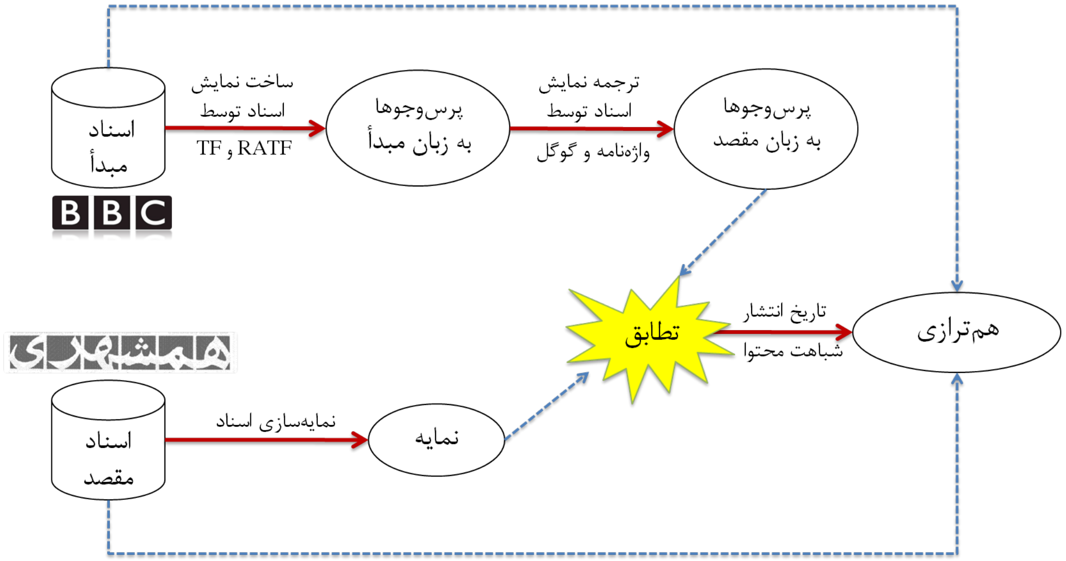
در بسیاری از مطالعاتی که بر روی پیکره‌های تطبیقی برای استخراج وابستگی کلمات بین زبانی صورت گرفته‌اند فرض شده است که پیکره‌های تطبیقی را می‌توان به راحتی با استفاده از هم‌ترازی تاریخ انتشار اخبار در زبان‌های مختلف بدست آورد [۸۹، ۹۳، ۹]. اگرچه ممکن است این فرض برای بعضی از زبان‌ها درست باشد، ولی با توجه به منابع موجود در حال حاضر این امکان برای زبان فارسی وجود ندارد. همان گونه که در فصل مقدمه نیز گفتیم، اخبار خبرگزاری‌هایی که روزانه اخبار را به هر دو زبان فارسی و انگلیسی منتشر می‌کنند بدین منظور مناسب نمی‌باشند، زیرا که مقالات انگلیسی آنها معمولاً بسیار کوتاهند و در اکثر مواقع ترجمه خلاصه شده‌ای از مقالات فارسی هستند و علاوه بر این، بایگانی طولانی مدتی به صورت برخط[۱] ندارند که به راحتی قابل دسترسی باشد. از طرف دیگر، از آنجایی که اخبار مرتبط به یک واقعه‌ی مشترک در تاریخ‌های مختلفی در خبرگزاری‌های متفاوت منتشر می‌شوند، نمی‌توان به راحتی آنها را تنها توسط تاریخ انتشار هم‌تراز کرد. بنابراین برای ساخت پیکره تطبیقی فارسی-انگلیسی از دو مجموعه اخبار متفاوت از دو مبدأ مجزا، اخبار فارسی خبرگزاری همشهری و اخبار انگلیسی خبرگزاری بی‌بی‌سی[۲] استفاده می‌کنیم. همچنین برای هم‌ترازی اسناد این دو زبان، علاوه بر تاریخ انتشار اخبار، شباهت محتوای اسناد را نیز در نظر می‌گیریم.

در پیکره‌ی تطبیقی ساخته شده، هدف ما این است که اسناد هم‌تراز شده تا حد ممکن به یکدیگر شبیه باشند. به

---

[۱] online

[۲] BBC



شکل ۱.۳: مراحل ساخت پیکره‌ی تطبیقی

طور شهودی، دو سند هنگامی به یکدیگر شبیه هستند که کلید واژگانشان[۱] (کلماتی که بهترین توصیف را از موضوع سند دارند) به یکدیگر شبیه باشند. همچنین معیاری دیگر برای پیدا کردن هم‌ترازی‌های خوب، تاریخ انتشار اخبار است. به طور کلی، اسنادی که محتوای مشابهی دارند اگر در یک تاریخ منتشر شده باشند به احتمال زیاد درباره‌ی یک رویداد صحبت می‌کنند. بر اساس این شهود و روش ارائه شده در مرجع [۸۶]، مراحل زیر را برای هم‌ترازی اسناد طی می‌کنیم: ابتدا کلید واژگان اسناد زبان مبدأ (انگلیسی) را استخراج کرده و آنها را به زبان مقصد (فارسی) ترجمه می‌کنیم. سپس کلمات ترجمه شده نماینده اسناد انگلیسی را به عنوان پرس‌وجوهایی بر روی مجموعه مقصد اعمال می‌کنیم. در نهایت اسناد را بر اساس امتیاز شباهت محتوایی و تاریخ انتشارشان هم‌تراز می‌کنیم. مراحل ساخت پیکره تطبیقی به طور کلی در شکل ۱.۳ آمده است. در ادامه این فصل، روش ساخت و ارزیابی پیکره‌ی تطبیقی را با جزئیات بیشتری توضیح می‌دهیم.

۱.۳ ساخت و ترجمه نمایش اسناد

به منظور انتخاب کلمات نمایش‌دهنده هر سند مبدأ از معیار میانگین نسبی بسامد اصطلاح (RATF[۲]) [۶۳] استفاده کرده‌ایم. این معیار بر مبنای بسامد تکرار کلمات در مجموعه اسناد، مقداری عددی به هر کلمه اختصاص می‌دهد و به

[۱] key words

[۲] Relative Average Term Frequency

صورت زیر تعریف شده است:

$$RATF(k) = (cf_k/df_k) \times 10^3/ln(df_k + SP)^p, \qquad (۱.۳)$$

در این رابطه cf_k بسامد مجموعه[۱] کلمه k یعنی تعداد دفعاتی که آن کلمه در کل مجموعه اسناد ظاهر شده است و df_k بسامد سند[۲] کلمه k یعنی تعداد اسنادی که آن کلمه در آنها ظاهر شده است، می‌باشند. SP و p ثابت‌هایی وابسته به مجموعه می‌باشند که برای کم کردن وزن کلمات نادر تعریف شده‌اند. برای تعیین مقادیر این پارامترها همانند مرجع [۸۶] چند نمونه آزمایش برای سنجش دقت مقادیر RATF کلمات انجام داده‌ایم و به طور شهودی مقادیرِ پارامترهای SP و p را به صورت $SP = 1800$ و $p = 3$ در نظر گرفته‌ایم. به منظور ساخت پرس‌وجوهایی که نماینده‌ی اسناد مبدأ هستند، ابتدا کلمات هر سند را بر اساس بسامد تکرارشان در هر سند به صورت نزولی مرتب می‌کنیم. سپس کلید واژگانی که دارای بسامد یکسان هستند بر اساس مقدار RATF هر یک مرتب می‌شوند. در حالت کلی m کلید واژه با بالاترین رتبه به عنوان پرس‌وجوی نمایش دهنده آن سند انتخاب می‌شوند. برای ایجاد پیکره تطبیقی در آزمایش‌های انجام شده ۳۰ کلید واژه را در نظر گرفته‌ایم. از آنجایی که ممکن است به کمک منابع موجود نتوان تمام کلمات انتخاب شده‌ی زبان مبدأ را به زبان مقصد ترجمه کرد، تعداد نسبتاً زیادی کلید واژه را به عنوان نماینده هر سند انتخاب کرده‌ایم. به منظور ترجمه‌ی کلید واژگان از یک واژه‌نامه انگلیسی-فارسی که بیش از ۵۰،۰۰۰ لغت دارد استفاده کرده‌ایم. از آنجایی که تعداد کلمات خارج از واژگان[۳] همانند اسامی خاص، بسیار زیاد بود، از ماشین ترجمه‌ی گوگل[۴] برای ترجمه کلماتی که در واژه‌نامه یافت نشده‌اند استفاده کرده‌ایم. سپس از سیستم نویسه گردانی[۵] گوگل[۶] به منظور ترجمه کلماتی که هنوز ترجمه نشده بودند همانند تعدادی اسامی خاص و کلمات ریشه‌یابی شده، استفاده می‌کنیم. پس از ساخت پرس‌وجوی نمایش دهنده هر سند به زبان مقصد، از یک مدل بازیابی برای رتبه‌بندی اسناد زبان مقصد بر اساس شباهتشان به پرس‌وجوی ساخته شده بهره می‌جوییم. مدل‌های بازیابی استفاده شده در بخش ارزیابی پیکره‌های تطبیقی، در ادامه این فصل بررسی شده‌اند. در مرحله‌ی آخر نیز، از امتیازات شباهت بدست آمده به همراه تاریخ انتشار اسناد برای هم‌ترازی اسناد استفاده می‌کنیم.

---

[۱] collection frequency

[۲] document frequency

[۳] Out Of Vocabulary (OOV)

[۴] http://translate.google.com/

[۵] transliteration

[۶] www.google.com/transliterate/persian

۲.۳ هم‌ترازی اسناد

به منظور ساخت زوج‌های هم‌ترازی[۱] اسناد فارسی و انگلیسی، از دو معیار پایه استفاده می‌کنیم: (۱) امتیازهای شباهت محتوایی اسناد و (۲) تاریخ انتشار آنها. به طور شهودی، اگر یک سند انگلیسی و یک سند فارسی امتیاز شباهت محتوایی بالایی داشته باشند و در یک تاریخ نیز منتشر شده باشند، آنگاه احتمال اینکه درباره یک رخداد مرتبط باشند بسیار زیاد است. از طرف دیگر، اگر امتیاز شباهت محتوایی آنها پایین باشد یا با فاصله زمانی زیادی منتشر شده باشند، آنگاه آن دو سند احتمالاً زوج خوبی نیستند. برای پیدا کردن هم‌ترازی‌های مناسب اسناد، ترکیبی از سه حد آستانه متفاوت ($\theta_1 < \theta_2 < \theta_3$) را بر روی امتیازهای شباهت محتوایی اعمال می‌کنیم. مقادیر حد آستانه‌های انتخاب شده به صورت در صد هستند، به عنوان مثال $\theta = 60$ به این معنی است که امتیاز باید بیشتر از ۶۰٪ از تمام امتیازات شباهت محتوایی در آن اجرای بازیابی اطلاعات باشد. اگر n تعداد اسناد مبدأ باشد و برای هر سند مبدأ به تعداد r سند مقصد بررسی شده باشند، آنگاه برای محاسبه حد آستانه‌ها $n \times r$ امتیاز باید در نظر گرفته شوند [۸۶]. در آزمایش‌های انجام شده در این پژوهش، r را برابر با ۵۰ در نظر گرفته‌ایم.

به طور کلی، مراحل هم‌ترازی اسناد به این صورت است. یک سند انگلیسی را در نظر می‌گیریم، ابتدا در مجموعه شبیه‌ترین اسناد مقصد به آن که شامل بالاترین r سند فارسی می‌باشند، جستجویی را انجام می‌دهیم، بدین صورت که اسناد فارسی که در همان روز سند انگلیسی منتشر شده‌اند و همچنین امتیاز شباهت محتوایشان بیشتر از θ_1 است را پیدا کرده و به عنوان زوج هم‌ترازی در نظر می‌گیریم. اگر هیچ زوج هم‌ترازی پیدا نکردیم، آنگاه حد آستانه را به θ_2 افزایش داده و به دنبال اسناد مقصد در بازه زمانی چهار روز می‌گردیم که امتیاز شباهت محتوایشان بیشتر از این حد آستانه باشد. در نهایت، اگر هنوز هیچ سندی که با معیارهایمان مطابقت کند پیدا نکرده باشیم، حدآستانه امتیاز به θ_3 افزایش پیدا کرده و بازه زمانی چهارده روز جستجو می‌شود. حدآستانه‌ها را به طور شهودی بر اساس آزمایش‌های انجام شده بر روی ایجاد پیکره‌های تطبیقی و ارزیابی آنها، تعیین کرده‌ایم.

۳.۳ ارزیابی پیکره تطبیقی ساخته شده

در این پژوهش، آزمایش‌های مختلفی برای ساخت پیکره‌های تطبیقی انجام داده‌ایم. در ادامه به ارزیابی پیکره‌های تطبیقی ساخته شده می‌پردازیم.

---

[۱] alignment pair

۱.۳.۳ معیار ارزیابی

معیار اصلی ارزیابی پیکره تطبیقی، کیفیت هم‌ترازی‌ها می‌باشد. در آزمایش‌های انجام شده، کیفیت هم‌ترازی‌ها را با استفاده از مقیاس ارتباطی پنج سطحی، که از مرجع [۱۲] اقتباس شده است، ارزیابی کرده‌ایم. پنج سطح ارتباطی به صورت زیر تعریف شده است:

(۱) موضوع یکسان. دو سند دقیقاً درباره یک رویداد هستند.

(۲) موضوع مرتبط. دو سند درباره‌ی یک رویداد هستند اما از نقطه نظرهای متفاوت. (به عنوان مثال یک سند ممکن است جزئی از سند دیگر باشد)

(۳) ظاهر مشترک. دو سند درباره‌ی رویدادهای مرتبط هستند، به عنوان مثال رویدادهای مربوط به یک مکان یا افراد مشترک.

(۴) اصطلاحات مشترک. شباهت بین رویدادها ناچیز است ولی تعداد قابل توجهی واژگان مشترک دارند.

(۵) نامرتبط. هیچ گونه شباهت آشکاری بین اسناد وجود ندارد.

اینکه کدام کلاس‌ها به عنوان هم‌ترازی‌های خوب در نظر گرفته شوند بستگی به مورد استفاده در نظر گرفته شده دارد. برای مثال، نگارندگان در مرجع [۱۲] کلاس‌های (۱) تا (۴) را برای استخراج ترجمه‌های خوب کلمات در سیستم‌های بازیابی اطلاعات بین زبانی مفید دانسته‌اند. ما در آزمایش‌هایمان، کلاس‌های (۱) تا (۳) را به عنوان هم‌ترازی‌های مناسب در نظر می‌گیریم. بنابراین در یک پیکره تطبیقی با کیفیت بالا انتظار می‌رود که بیشتر هم‌ترازی‌ها در سطوح (۱)، (۲) و (۳) باشد و تعداد زیادی هم‌ترازی در سطوح (۴) و (۵) وجود نداشته باشد. همچنین تعداد هم‌ترازی‌های پیدا شده با کیفیت بالا را به عنوان معیاری دیگر برای ارزیابی پیکره ساخته شده در نظر می‌گیریم.

۲.۳.۳ مشخصات مجموعه اسناد فارسی و انگلیسی

همان طور که در بخش‌های قبل ذکر کردیم، در این پژوهش به منظور ساخت پیکره تطبیقی از اخبار منتشر شده فارسی و انگلیسی استفاده شده است. مجموعه انگلیسی که استفاده کرده‌ایم شامل اخبار منتشر شده توسط خبرگزاری بی‌بی‌سی و مجموعه فارسی استفاده شده شامل اخبار منتشر شده در روزنامه همشهری می‌باشد. برای این منظور پنج سال از اخبار منتشر شده توسط این دو خبرگزاری را از ژانویه ۲۰۰۲ تا دسامبر ۲۰۰۶ در نظر گرفته‌ایم. اخبار بی‌بی‌سی از روی

جدول ۱.۳: مشخصات مجموعه اسناد انگلیسی (بی‌بی‌سی) و فارسی (همشهری)

مجموعه	تعداد اسناد	مدت	میانگین طول اسناد	کلمات یکتا
بی‌بی‌سی	۵۳،۶۹۷	ژانویه۲۰۰۲-دسامبر۲۰۰۶	۴۶۱	۱۴۱،۸۱۹
همشهری	۱۹۱،۴۴۰	ژانویه۲۰۰۲-دسامبر۲۰۰۶	۵۲۷	۵۲۸،۸۶۴

جدول ۲.۳: نمونه‌ای از هم‌ترازی اسناد فارسی و انگلیسی

سند مبدأ	سند مقصد	درجه ارتباط
01/04/2002-South Asia economy battles for attention, The summit is in the Himalayan state of Nepal The real threat of war between India and Pakistan is likely to dominate the agenda at a summit of seven South Asian nations this weekend. But the South Asian Association for Regional Cooperation (SAARC) has been battling an equally serious war against poverty for the past 15 years. Member countries India Pakistan Sri Lanka Maldives Bangladesh Bhutan Nepal	۲۰۰۲/۰۴/۰۱-نشست سارک امروز آغاز شد. نشست وزرای امور خارجه کشورهای عضو مجمع همکاری‌های منطقه‌ای جنوب آسیا سارک چهارشنبه در کاتماندو کار خود را آغاز کرد. به گزارش خبرگزاری فرانسه از کاتماندو وزرای امور خارجه هفت کشور هند، پاکستان، سریلانکا، نپال، بنگلادش، بوتان و مالدیو مقدمات نشست سران را که قرار است از امروز به مدت سه روز تحت تدابیر شدید امنیتی در پایتخت نپال آغاز شود آماده می کنند.	۱۵۴.۲۶
02/28/2002-US economy shows surprise strength, The US economy grew at a much faster rate than previously thought during the last three months of 2001. The latest figures show that exceptionally strong consumer spending propelled the economy to grow at an annual rate of 1.4%.	۲۰۰۲/۰۱/۰۳-رئیس خزانه داری فدرال در جلسه کنگره ، نشانه‌های رونق در اقتصاد آمریکا ظاهر می‌شود. پس از اظهارات رئیس خزانه داری فدرال آمریکا شاخص وال استریت یک درصد بالا رفت.	۱۲۹.۷۱۲
07/25/2006-Yahoo unveils security initiative, Internet company Yahoo and security provider Symantec have unveiled a joint internet security service, say reports.	۲۰۰۶/۲۹/۰۷-خدمات امنیتی مشترک یاهو و سیمانتک، گروه علمی فرهنگی- یاهو و سیمانتک با همکاری یکدیگر خدمات امنیت اینترنتی جدیدی را به نام نورتون اینترنت سکیوریتی راه اندازی کردند.	۱۹۰.۰۹۸

وب سایت خبرگزاری بی‌بی‌سی خزش[۱] شده‌اند و پس از پیش پردازش، صفحات به فرمت استاندارد تبدیل شده‌اند. همچنین اخبار محلی انگلستان که به احتمال زیاد با هیچ یک از اسناد فارسی هم‌تراز نخواهند شد را حذف کرده‌ایم. اسناد همشهری از پیکره‌ی همشهری [۵] که شامل اخبار منتشر شده در سال‌های ۱۹۹۶ تا ۲۰۰۷ است، استخراج شده‌اند. جزئیات مجموعه‌های استفاده شده برای ساخت پیکره تطبیقی در جدول ۱.۳ آمده است. همچنین نمونه‌ای از هم‌ترازی اسناد فارسی و انگلیسی که توسط روش مطرح شده در بخش ۲.۳ ایجاد شده است، در جدول ۲.۳ آورده شده است.

[۱] crawl

۳.۳.۳ نتایج ارزیابی پیکره‌های تطبیقی

در این بخش نتایج آزمایش‌هایمان را برای ساخت پیکره‌های تطبیقی فارسی-انگلیسی و تحلیل کیفیت آنها گزارش می‌دهیم. در آزمایش‌های انجام شده از ابزار Lemur[۱] به عنوان سیستم بازیابی اطلاعات استفاده کرده‌ایم. به عنوان پیش پردازش مجموعه انگلیسی از ریشه‌یاب پُرتر[۲] برای ریشه‌یابی کلمات انگلیسی و از لیست ایست واژه[۳] Inquery (۴۱۸ کلمه) استفاده کرده‌ایم. آزمایش‌هایی نیز با استفاده از چند نوع ریشه‌یاب زبان فارسی (از جمله ریشه‌یاب نوشته شده توسط دانشگاه نوشاتل [۴]) برای ریشه‌یابی کلمات فارسی انجام داده‌ایم، ولی از آنجایی که هیچ کدام از آنها کیفیت لازم را نداشتند در نهایت بر روی مجموعه فارسی هیچ گونه پیش‌پردازشی انجام نداده‌ایم و تنها از لیست ایست واژگان فارسی دانشگاه نوشاتل (۳۵۹ کلمه)[۵] استفاده کرده‌ایم.

در این پژوهش، آزمایش‌های گوناگونی را بر اساس انواع مختلف (۱) روش ترجمه کلید واژگان زبان مبدأ و (۲) مدل‌های بازیابی اطلاعات انجام شده است. به منظور مقایسه کیفیت هم‌ترازی‌های مختلف، که متناظر با پیکره‌های تطبیقی مختلف هستند، به صورت دستی کیفیت هم‌ترازی‌های پیدا شده در یک بازه زمانی یک ماهه، ژانویه ۲۰۰۲، را بر اساس مقیاس ارتباطی سطوح پنج‌گانه (معرفی شده در بخش ۱.۳.۳) ارزیابی کرده‌ایم. نتایج ارزیابی نشان می‌دهند که اعمال انواع مختلف روش‌های ترجمه و مدل‌های بازیابی برای ساخت پیکره‌های تطبیقی، پیکره‌هایی با کیفیت‌های متفاوت را نتیجه می‌دهد.

در اولین مجموعه آزمایش‌ها، از یک واژه‌نامه ساده برای ترجمه کلید واژگان زبان مبدأ و از ماشین ترجمه گوگل برای ترجمه کلماتی که در واژه‌نامه نیستند، استفاده می‌کنیم. همچنین از مدل بازیابی KL-divergence با شبه بازخورد ارتباطی[۶] [۴۰] به منظور محاسبه شباهت محتوایی اسناد انگلیسی به فارسی استفاده شده است. در این دسته از آزمایش‌ها مقادیر حدود آستانه را به صورت زیر در نظر گرفته‌ایم: $\theta_1 = 60, \theta_2 = 80, \theta_3 = 95$. جدول ۳.۳(آ) ارزیابی کیفیت هم‌ترازی‌ها را با استفاده از تمام ترجمه‌های موجود یک کلمه در واژه‌نامه نشان می‌دهد. همان گونه که جدول نشان

[۱] http://www.lemurproject.org/

[۲] Porter

[۳] stop word

[۴] http://members.unine.ch/jacques.savoy/clef/persianStemmerUnicode.txt

[۵] www.unine.ch/info/clef/

[۶] pseudo relevance feedback

جدول ۳.۳: ارزیابی کیفیت هم‌ترازی‌ها در یک ماه، ژانویه ۲۰۰۲. بر اساس مدل بازیابی KL-divergence. (آ) استفاده از تمام ترجمه‌های هر کلمه. (ب) استفاده از ۳ بالاترین ترجمه هر کلمه.

(آ)

	تعداد هم‌ترازی‌ها	درصد هم‌ترازی‌ها
کلاس ۱	۴	۱۱.۷۶
کلاس ۲	۴	۱۱.۷۶
کلاس ۳	۷	۲۰.۵۸
کلاس ۴	۱۱	۳۲.۳۵
کلاس ۵	۸	۲۳.۵۲
کل	۳۴	۱۰۰

(ب)

	تعداد هم‌ترازی‌ها	درصد هم‌ترازی‌ها
کلاس ۱	۳	۶.۹۷
کلاس ۲	۱۷	۳۹.۵۳
کلاس ۳	۱۴	۳۲.۵۵
کلاس ۴	۸	۱۸.۶
کلاس ۵	۱	۲.۳۲
کل	۴۳	۱۰۰

می‌دهد، به سختی ۴۵٪ از هم‌ترازی‌های ارزیابی شده درباره رویدادهای مرتبطی هستند و بیش از نیمی از هم‌ترازی‌ها در کلاس‌های (۴) و (۵) قرار دارند، بدین معنی که این اسناد بسیار کم و یا اصلاً شباهتی با یکدیگر ندارند.

در واژه‌نامه استفاده شده در این پژوهش، هر کلمه تعداد متفاوتی ترجمه دارد که این ممکن است ترجمه پرس‌وجوهای نمایش دهنده اسناد را تحت تاثیر قرار دهد. در مجموعه آزمایش‌های دومی که انجام داده‌ایم، تنها از سه تا از بالاترین ترجمه‌های هر کلمه در واژه‌نامه برای ساخت پرس‌وجوهای زبان مقصد استفاده کرده‌ایم. جدول ۳.۳ (ب) نتایج هم‌ترازی‌ها را در مجموعه دوم آزمایش‌ها نشان می‌دهد. همان طور که دیده می‌شود، تقریباً ۷۹٪ از هم‌ترازی‌ها درباره رویدادهای مرتبط هستند. این نتایج نشان می‌دهند که با استفاده از سه تا بالاترین ترجمه هر کلمه می‌توان بهبود نسبتاً زیادی در کیفیت پیکره تطبیقی به دست آورد.

پس از استفاده از واژه‌نامه به همراه ماشین ترجمه، هنوز کلماتی در زبان مبدأ وجود دارند که ترجمه نمی‌شوند. این کلمات یا از نوع اسامی خاص هستند (همانند Euroland، Brinkema، Moussaoui، Toiba و Belkheir) و یا کلمات ریشه‌یابی شده می‌باشند (همانند Sydnei، Melbourn و Athen). به نظر می‌رسد که این کلمات تأثیر بسزایی در هم‌ترازی‌هایمان داشته باشند. یعنی اگر بتوان به گونه‌ای ترجمه آنها را وارد پرس‌وجوهای نمایش دهنده هر سند کرد، به احتمال زیاد می‌توان دقت هم‌ترازی‌های پیدا شده را افزایش داد. به همین دلیل در مجموعه بعدی آزمایش‌هایمان سعی کرده‌ایم که کلماتی را که هنوز ترجمه نشده‌اند را نویسه گردانی کنیم، یعنی حرف به حرف کلمات از زبان انگلیسی به زبان فارسی برگردانده شوند. جدول ۴.۳ نتایج اضافه کردن نویسه‌گردانی گوگل بر روی کلمات ترجمه نشده را نشان می‌دهد. به منظور مقایسه راحت‌تر نتایج بدون استفاده از نویسه‌گردانی را دوباره تکرار کرده‌ایم. همان گونه که از جدول می‌توان

جدول ۴.۳: نتایج ۳ بالاترین ترجمه‌ها با و بدون نویسه گردانی و مدل بازیابی KL-divergence

	بدون نویسه‌گردانی		با نویسه‌گردانی	
	تعداد هم‌ترازی‌ها	درصد هم‌ترازی‌ها	تعداد هم‌ترازی‌ها	درصد هم‌ترازی‌ها
کلاس ۱	۳	۶.۹۷	۵	۹.۴۳
کلاس ۲	۱۷	۳۹.۵۳	۲۴	۴۵.۲۸
کلاس ۳	۱۴	۳۲.۵۵	۱۴	۲۶.۴۱
کلاس ۴	۸	۱۸.۶	۸	۱۵.۰۹
کلاس ۵	۱	۲.۳۲	۲	۳.۷۷
کل	۴۳	۱۰۰	۵۳	۱۰۰

جدول ۵.۳: نتایج ۳ بالاترین ترجمه‌ها با و بدون نویسه گردانی و مدل بازیابی Okapi

	بدون نویسه‌گردانی		با نویسه‌گردانی	
	تعداد هم‌ترازی‌ها	درصد هم‌ترازی‌ها	تعداد هم‌ترازی‌ها	درصد هم‌ترازی‌ها
کلاس ۱	۱۱	۱۳.۵۸	۱۳	۱۴.۹۴
کلاس ۲	۴۶	۵۶.۷۹	۵۱	۵۸.۶۲
کلاس ۳	۲۰	۲۴.۶۹	۱۹	۲۱.۸۳
کلاس ۴	۴	۴.۹۳	۴	۴.۵۹
کلاس ۵	۰	۰	۰	۰
کل	۸۱	۱۰۰	۸۷	۱۰۰

دید، استفاده از نویسه‌گردانی زوج‌های هم‌ترازی بهتری را نتیجه داده است.

در مجموعه آزمایش‌های بعدی، از مدل بازیابی Okapi با شبه بازخورد ارتباطی [۷۰] و همچنین از سه تا از بالاترین ترجمه‌ی کلمات واژه‌نامه استفاده کرده‌ایم. جدول ۵.۳ نتایج بدست آمده را نشان می‌دهد. از آنجایی که هدف اصلی ما این است که تا آنجایی که ممکن است جفت اسنادی که کیفیت بالایی را دارند پیدا کنیم، بنابراین اندازه پیکره‌های هم‌تراز شده را علاوه بر کیفیتشان با یکدیگر مقایسه می‌کنیم. همان گونه که از جداول ۴.۳ و ۵.۳ مشاهده می‌شود، استفاده از نویسه‌گردانی باعث افزایش تعداد هم‌ترازی می‌شود. نکته جالب توجه اینجا است که هم‌ترازی‌های پیدا شده‌ی جدید بیشتر در کلاس‌های (۱) و (۲) پخش شده‌اند. این مشاهده نشان می‌دهد که جفت اسناد هم‌تراز شده‌ی جدید بسیار و یا تا حدود زیادی به یکدیگر مربوط هستند.

جدول ۶.۳: توزیع کیفیت هم‌ترازی‌ها برای دو مجموعه متفاوت از حدود آستانه

	$\theta_1 = 60, \theta_2 = 80, \theta_3 = 95$		$\theta_1 = 0, \theta_2 = 60, \theta_3 = 95$	
	تعداد هم‌ترازی‌ها	درصد هم‌ترازی‌ها	تعداد هم‌ترازی‌ها	درصد هم‌ترازی‌ها
کلاس ۱	۱۳	۱۴.۹۴	۱۶	۱۲.۸
کلاس ۲	۵۱	۵۸.۶۲	۶۲	۴۹.۶
کلاس ۳	۱۹	۲۱.۸۳	۳۰	۲۴
کلاس ۴	۴	۴.۵۹	۱۴	۱۱.۲
کلاس ۵	۰	۰	۳	۲.۴
کل	۸۷	۱۰۰	۱۲۵	۱۰۰

این مجموعه از آزمایش‌ها بهترین نتایج را از میان آزمایش‌های هم‌ترازی ما دارد، بنابراین از این مجموعه از اسناد هم‌تراز شده به عنوان پیکره تطبیقی ساخته شده در این پژوهش استفاده می‌کنیم. این پیکره تطبیقی شامل ۱۰،۳۶۵ زوج هم‌ترازی اسناد است که از این میان ۱۰٪ از ۵۵،۶۹۷ سند مبدأ هم‌تراز شده‌اند. از آنجایی که اسناد مبدأ و مقصد بسیار با هم متفاوت هستند، پیدا کردن تعداد نسبتاً کمی هم ترازی قابل پیش‌بینی بود. از طرف دیگر، تعداد هم‌ترازی‌ها با کم کردن حدود آستانه می‌تواند افزایش یابد، البته این تغییر ممکن است بر روی کیفیت پیکره‌ی تطبیقی تاثیر داشته باشد. جدول ۶.۳ نتایج ارزیابی پیکره تطبیقی با حدود آستانه $\theta_1 = 0, \theta_2 = 60, \theta_3 = 95$ را نشان می‌دهد. همان‌گونه که جدول نشان می‌دهد، کم کردن حدود آستانه باعث می‌شود ۴۳.۷٪ تعداد هم‌ترازی اسناد بیشتر شود، ولی درصد هم‌ترازی‌های با کیفیت بالا از ۹۵.۴٪ به ۸۶.۴٪ افت می‌کند. پیکره‌ی تطبیقی ما شامل ۷۶٪ تمام هم‌ترازی‌های با کیفیت بالا و تعداد بسیار کمی هم‌ترازی‌های با کیفیت پایین می‌باشد. لازم به ذکر است که کیفیت هم‌ترازی‌ها هنگام استخراج دانش ترجمه از پیکره امری ضروری است و به همین دلیل ما بیشتر بر روی آن تأکید داریم.

به طور کلی، نتایج ارزیابی روش‌های مختلف بر روی آزمایش‌های انجام شده نشان می‌دهند که (۱) استفاده از تعداد محدود ترجمه‌های هر کلمه در واژه‌نامه (به عنوان مثال سه ترجمه‌ی اول هر کلمه) نسبت به استفاده از کل ترجمه‌های آن کلمه بهتر است و می‌تواند کیفیت پیکره تطبیقی را بهبود دهد. (۲) استفاده از نویسه گردانی علاوه بر واژه‌نامه و ماشین ترجمه، دقت پیکره را افزایش می‌دهد. (۳) استفاده از مدل بازیابی بازخورد[۱] نیز می‌تواند در انجام آزمایش‌ها مفید واقع شود.

[۱] feedback

۴.۳ مقایسه با کارهای پیشین

در این پژوهش به طور کلی از روش‌های ارائه شده در دو مرجع [۱۲] و [۸۶] برای ساخت پیکره تطبیقی فارسی-انگلیسی استفاده شده است. تالونساری در [۸۶] روشی را برای ساخت پیکره تطبیقی از دو مجموعه با زبان‌های متفاوت و مبدأهای مختلف ارائه کرده است. در آن تحقیق، نویسندگان بهترین کلید واژگان نماینده اسناد یک مجموعه را توسط فرمول RATF استخراج می‌کنند. کلید واژگان توسط یک برنامه ترجمه پرس‌وجوی مبنی بر واژه‌نامه (UTACLIR [۳۴]) به زبان مجموعه دیگر ترجمه می‌شوند. پرس‌وجوهای ترجمه شده بر روی مجموعه مقصد اعمال شده و اسناد بر اساس امتیاز شباهت محتوا و تاریخ انتشارشان هم‌تراز می‌شوند. روش آنها بر روی مجموعه اخبار خبرگزاری سوئدی TT[۱] و مجموعه اخبار روزنامه آمریکایی لس آنجلس تایمز[۲] آزموده شده است. روشی که آنها در مرجع [۸۶] استفاده کرده‌اند بسیار شبیه به روش ارائه شده در مرجع [۱۲] است، که در آن برای هم‌ترازی اسناد در زبان‌های مختلف از تاریخ‌ها، موضوع‌ها، اسامی خاص مشترک، اعداد درون اسناد و یک واژه‌نامه کوچک استفاده کرده‌اند.

روشی که ما در این پژوهش برای ساخت پیکره تطبیقی استفاده کرده‌ایم، در پاره‌ای از جزئیات با روش‌های ارائه شده در [۱۲] و [۸۶] متفاوت است. بیشتر این تفاوت‌ها به دلیل تفاوت‌های زبان و منابع زبان‌شناسی موجود برای آن زبان‌ها است. تالونساری در [۸۶] از ابزار بن‌واژه سازی[۳] TWOL [۳۹] بر روی کلمات اسناد سوئدی استفاده کرده است، در حالی که ما هیچ گونه پیش پردازشی بر روی مجموعه‌ی فارسی انجام نداده‌ایم. تفاوت دیگر در ترجمه کلید واژگان است. نگارندگان در مرجع [۸۶] از ابزار UTACLIR [۳۴] که یک برنامه ترجمه‌ی پرس‌وجوی مبنی بر واژه‌نامه است، استفاده کرده‌اند. این برنامه از ساختار پرس‌وجو برای ابهام‌زدایی ترجمه‌های مختلف یک کلمه استفاده می‌کند و همچنین یک روش تطابق رشته فازی[۴] برای تبدیل کلماتی که در واژه‌نامه نیستند در آن در نظر گرفته شده است. این در حالی است که ما از یک واژه‌نامه ساده و ماشین ترجمه گوگل برای ترجمه کلماتی که در واژه‌نامه نیستند، استفاده کرده‌ایم. با این حال کلمات بسیاری وجود داشتند که توسط ماشین ترجمه نیز قابل ترجمه نبودند. وجود این گونه کلمات ما را به سوی استفاده از نویسه گردانی کلمات ترجمه نشده سوق داد که بدین منظور از ابزار نویسه گردانی گوگل استفاده کرده‌ایم. آزمایش‌های انجام شده نشان داده‌اند که استفاده از نویسه گردانی علاوه بر ترجمه کلمات، بسیار سودمند است و باعث

[۱] Tidningarnas Telegrambyra

[۲] Los Angeles Times

[۳] lemmatization

[۴] fuzzy string matching

افزایش کیفیت پیکره تطبیقی ساخته شده می‌شود. علاوه بر این، تالونساری فرض‌های زیادی را هم در مرحله هم‌ترازی و هم برای تعیین پارامترها و آستانه‌ها کرده است. در حالی که ما سعی کرده‌ایم به صورت کلی آزمایش‌هایی را انجام دهیم و حتی الامکان از پیش فرض‌های اضافه در پژوهشمان استفاده نکنیم.

مهم‌ترین دستاورد این قسمت از پژوهش انجام شده، ایجاد پیکره‌ی تطبیقی فارسی-انگلیسی است، که بر طبق مطالعات انجام شده، نخستین پیکره از این نوع در زبان فارسی می‌باشد. از آنجایی که برای ساخت این پیکره تطبیقی از روش‌های ارائه شده در پژوهش‌های پیشین استفاده شده است در روش ایجاد پیکره نوآوری وجود ندارد. اما همان طور که ذکر شد، جزئیات روش ساخت پیکره تطبیقی فارسی-انگلیسی ایجاد شده، تفاوت‌هایی با کارهای پیشین دارد.

فصل ۴

بازیابی اطلاعات بین زبانی با استفاده از پیکره‌های تطبیقی

در بازیابی اطلاعات بین زبانی مبنی بر ترجمه، سه مسأله اصلی مطرح است، مسأله اول اینکه چه چیزی را باید ترجمه کرد، دوم این که دانش ترجمه را چگونه باید بدست آورد و در نهایت اینکه چگونه می‌توان از دانش ترجمه‌ی استخراج شده استفاده کرد [۳۳]. در این پژوهش، ترجمه‌ی کلمات[۱] و عبارات[۲] (به طور کلی اصطلاحات[۳]) را از پیکره تطبیقی ساخته شده استخراج می‌کنیم. همچنین از آنجایی که بعضی کلمات ممکن است ترجمه‌شان از پیکره تطبیقی بدست نیاید، بر اساس اطلاعات همسایگی‌شان، آنها را به عنوان برون‌هشته[۴] شناسایی کرده و از منبع غنی دیگری همانند واژه‌نامه برای ترجمه آنها استفاده می‌کنیم. سپس مدل زبانی پرس‌وجوها را بر اساس ترجمه‌های استخراج شده می‌سازیم.

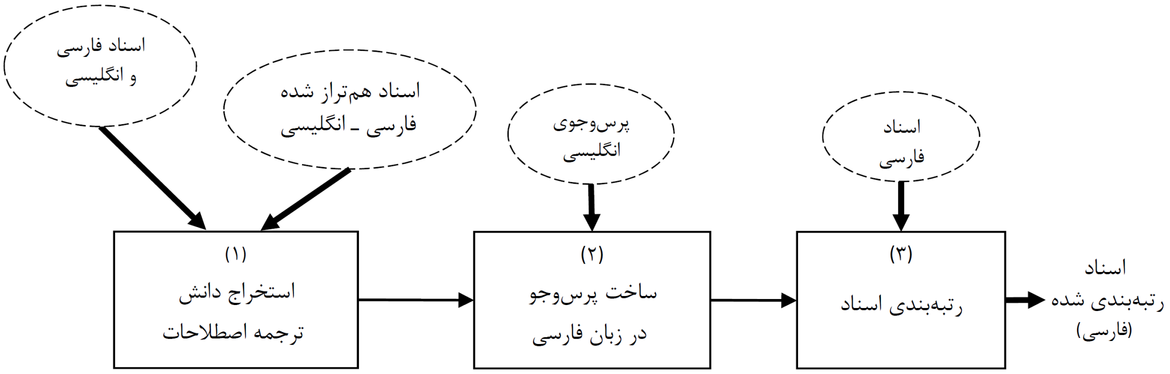
در این فصل روش‌هایی برای استخراج دانش ترجمه از پیکره‌های تطبیقی و ساخت مدل زبانی پرس‌وجوها برای انجام بازیابی اطلاعات بین زبانی ارائه می‌دهیم. به صورت کلی مراحل بازیابی بدین صورت است که پس از استخراج دانش ترجمه بر اساس رخداد همزمان اصطلاحات در هم‌ترازی‌های پیکره‌ی تطبیقی، پرس‌وجو را در زبان مقصد ساخته و اسناد

[۱] word

[۲] phrase

[۳] term

[۴] outlier



شکل ۱.۴: مراحل بازیابی اطلاعات بین زبانی

مقصد را با بازیابی آن پرس‌وجو رتبه‌بندی می‌کنیم. شکل ۱.۴ مراحل بازیابی اطلاعات بین زبانی را که در این فصل بررسی می‌کنیم، نشان می‌دهد. در بخش ۱.۴ روش‌های ارائه شده برای استخراج دانش ترجمه اصطلاحات را معرفی کرده و در بخش ۲.۴ روش‌های پیشنهادی مختلف ساخت مدل زبانی پرس‌وجو در زبان مقصد را بررسی می‌کنیم.

۱.۴ استخراج دانش ترجمه از پیکره‌های تطبیقی

از آنجایی که اسناد پیکره‌های تطبیقی بر خلاف اسناد موجود در پیکره‌های موازی کاملاً ترجمه یکدیگر نیستند و تنها تا حدودی به یکدیگر مرتبط می‌باشند، استخراج دانش ترجمه از پیکره‌های تطبیقی به سادگی انجام پذیر نیست [۸۸]. در این بخش به معرفی روش‌های مختلف یادگیری ارتباط اصطلاحات دو زبانه با استفاده از پیکره‌های تطبیقی خواهیم پرداخت. در این روش‌ها از ویژگی‌های آماری استخراج شده برای اصطلاحات موجود در پیکره استفاده شده است. به عنوان روش پایه، ترجمه اصطلاحات موجود در پیکره را با استفاده از معیار رخداد همزمان اصطلاح‌ها در اسناد انگلیسی و فارسی هم‌تراز شده در پیکره‌ی تطبیقی استخراج می‌کنیم. در گام بعدی این قسمت از پژوهش، روشی مبنی بر کاوش شبکه اصطلاحات ارائه می‌دهیم که به منظور استخراج دانش ترجمه بهتر، از اطلاعات داده‌های تک‌زبانی علاوه بر ارتباطات اصطلاحات دو زبان استفاده می‌کند. در ادامه یک روش ممیزی بر اساس تشخیص برون‌هشته‌ها برای بررسی صحّت ترجمه‌های استخراج شده معرفی می‌کنیم. همچنین روشی را به منظور استخراج ترجمه عبارات برای بهبود کارآیی بازیابی اطلاعات بین زبانی ارائه کرده‌ایم. جزئیات بیشتر هر یک از این روش‌ها در ادامه‌ی این بخش آورده شده است.

۱.۱.۴ COCOT: روش پایه استخراج دانش ترجمه

ما در این پژوهش به منظور استخراج وابستگی اصطلاحات موجود در پیکره‌ی تطبیقی، از روش ارائه شده توسط تالونساری در مرجع [۸۷] به عنوان روش پایه استخراج دانش ترجمه استفاده کرده‌ایم. به طور شهودی می‌توان گفت که این روش، مبنی بر استفاده از رخداد همزمان اصطلاحات در هم‌ترازی‌ها، برای استخراج وابستگی اصطلاحات می‌باشد. در الگوریتم ارائه شده توسط تالونساری ابتدا یک وزن m_{ik} برای هر اصطلاح s_i در سند d_k براساس رابطه ۱.۴ محاسبه می‌شود:

$$m_{ik} = \begin{cases} 0 & \text{if}\, tf_{ik} = 0 \\ (0.5 + 0.5 \times \frac{tf_{ik}}{Maxtf_k}) \times ln(\frac{NT}{dl_k}) & \text{otherwise} \end{cases} \quad (۱.۴)$$

در اینجا متغیرهای tf_{ik} بسامد تکرار s_i در سند d_k، $Maxtf_k$ بیشترین بسامد واژه در d_k و dl_k تعداد اصطلاحات یکتا در آن سند هستند. NT را می‌توان برابر با تعداد اصطلاحات یکتا و یا تقریبی از آن قرار داد. این رابطه به نوعی $tf.idf$ تغییر یافته می‌باشد و از مرجع [۸۰] اقتباس شده است که در آن مرجع از روشی مشابه برای محاسبه شباهت فرهنگ جامع[۱] استفاده کرده‌اند. در این روش نهایتاً وزن اصطلاح مقصد t_j در مجموعه‌ی رتبه‌بندی شده‌ی اسناد مقصد D به صورت زیر محاسبه می‌شود:

$$M_j = \sum_{r=1}^{|D|} \frac{m_{jr}}{ln(r+1)}, \quad (۲.۴)$$

که در این رابطه D لیست مرتبی از اسناد مقصدی است که با سند مبدأ که شامل واژه t_j است هم‌تراز شده است. لازم به ذکر است که اسناد درون D بر اساس امتیاز هم‌ترازیشان با سند مبدأ مرتب شده‌اند. به این ترتیب که اسنادی که کمتر شبیه سند مبدأ هستند در لیست با رتبه پایین‌تری ظاهر شده‌اند و برای استخراج ترجمه، اطمینان کمتری به آنها است. این جریمه توسط $ln(r+1)$ در مخرج کسر نائل شده است.

در نهایت امتیاز شباهت بین دو اصطلاح مبدأ s_i و اصطلاح مقصد t_j به صورت زیر محاسبه می‌شود:

$$w_c(s_i \to t_j) = \frac{\sum_{\langle d_k, D\rangle \in A} m_{ik} \times M_j}{\| s_i \| \times ((1-\alpha) + \alpha \times \frac{\|t_j\|}{\|T\|})}, \quad (۳.۴)$$

که در این رابطه m_{ik} وزن اصطلاح مبدأ s_i در سند مبدأ d_k است، M_j وزن اصطلاح مقصد t_j در مجموعه اسناد مقصد D که با سند مبدأ d_k هم‌تراز شده‌اند می‌باشد، A مجموعه‌ای از تمام هم‌ترازی‌ها است، $\| s_i \|$ اندازه بردار s_i است و

[۱] thesaurus

$\|T\|$ میانگین طول بردار واژه زبان مقصد است. لازم به ذکر است که در این رابطه، طرحِ نرمال‌سازی بردار محور[۱] [۸۲] بکار گرفته شده است. مقدار α پارامتری برای این طرح است و عددی ثابت در بازه پیوسته [۰،۱] می‌باشد که ما در آزمایش‌هایمان همانند مرجع [۸۷]، α را برابر ۰.۲ در نظر گرفته‌ایم. از آنجایی که روش نرمال‌سازی کسینوسی استاندارد[۲]، به کلمات کمیاب که دارای بردارهای ویژگی کوچکی هستند امتیاز بالایی می‌دهد، این طرح برای جریمه این موارد بکار گرفته شده است.

۲.۱.۴ TAN: مدل شبکه‌ای وابستگی اصطلاحات

همانطور که در بخش قبل به تفصیل بیان کردیم، در روش پایه‌ی استخراج دانش اصطلاحات، تنها رخداد همزمان اصطلاحات در هم‌ترازی‌های زبان مبدأ و مقصد در نظر گرفته می‌شوند. اما مشاهده کرده‌ایم که می‌توان از اطلاعات همبستگی اصطلاحات در مجموعه داده تک‌زبانی نیز به عنوان منبع اطلاعاتی مفیدی برای بهبود نتایج ترجمه استفاده کرد. به طور کلی می‌توان گفت که اصطلاحاتی در هر زبان وجود دارند که در یک حیطه خاص به کار می‌روند و احتمال اینکه این اصطلاحات با یکدیگر در یک سند ظاهر شوند بسیار زیاد است. به عنوان مثال کلمات "ویمبلدون" و "تنیس" رخداد همزمان بالایی با یکدیگر دارند. از این نوع اطلاعات که از داده‌های یک زبان می‌توان به دست آورد می‌توان برای بهبود کیفیت ترجمه‌های استخراج شده استفاده کرد. به طور شهودی می‌توان گفت که، دو اصطلاح در دو زبان متفاوت به احتمال زیاد ترجمه یکدیگر هستند اگر با یکدیگر همبستگی داشته باشند و علاوه بر این، اصطلاحات همسایگی اطرافشان در زبان بومی‌شان نیز به یکدیگر وابستگی زیادی داشته باشند. اصطلاح A در زبان مبدأ و اصطلاح B در زبان مقصد را در نظر بگیرید. اگر اصطلاحات واقع در همسایگی‌های A و B در زبان بومی‌شان با یکدیگر همبستگی زیادی نداشته باشند، آنگاه می‌توان گفت که دو اصطلاح A و B به احتمال زیاد نمی‌توانند ترجمه یکدیگر باشند، حتی اگر در روش COCOT به عنوان ترجمه یکدیگر استخراج شده‌اند.

به منظور دستیابی به چنین هدفی، در گام اول نیاز است که شبکه‌ای از تمام اصطلاحات زبان مبدأ و مقصد و ارتباط میان آنها بسازیم. شکل ۲.۴ ساختار کلی چنین شبکه‌ای را نشان می‌دهد. در این شکل خطوط بین اصطلاحات ارتباط آنها را نشان می‌دهند و هر چه این خطوط پر رنگ‌تر باشند نشان دهنده‌ی این است که آن دو اصطلاح با وزن بیشتری با یکدیگر ارتباط دارند. یال‌های بین دو اصطلاح در یک زبان نشان دهنده رخداد همزمان آن دو اصطلاح در آن زبان هستند.

---

[۱] Pivoted vector normalization scheme

[۲] standard cosine normalization

ما در آزمایش‌هایمان از اطلاعات متقابل[1] دو اصطلاح به عنوان پیوند بین آن دو در یک زبان استفاده کرده‌ایم. یال‌های بین دو اصطلاح در دو زبان متفاوت، شباهت کاوش شده بین آن دو باشد، که ما در آزمایش‌هایمان از مقادیر نرمالسازی شده ترجمه‌های استخراج شده در روش پایه COCOT به این منظور استفاده کرده‌ایم.

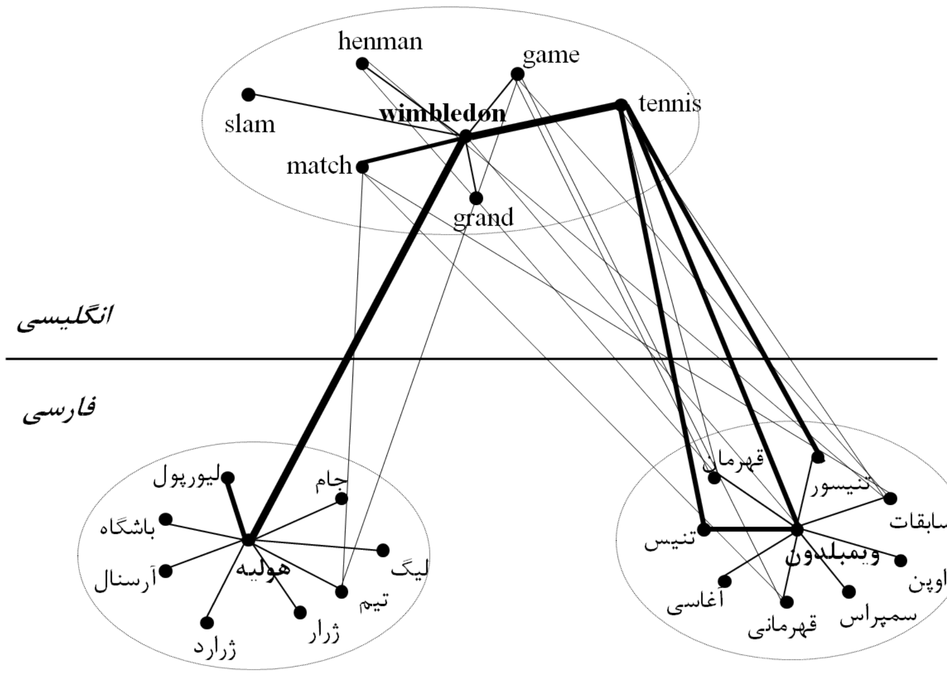
به طور شهودی، اگر دو اصطلاح ترجمه یکدیگر باشد، همسایگی آنها ارتباط قوی با یکدیگر دارند و اگر ترجمه یکدیگر نباشند، همسایگی آن دو مرتبط نیستند. شکل ۲.۴ نمونه‌ای از وابستگی اصطلاحات و همسایگی‌هایشان را نشان می‌دهد. برای کلمه "wimbledon"، کلمه با بیشترین همبستگی که توسط روش COCOT از پیکره تطبیقی استخراج شده، کلمه "هولیه" است که یک اسم خاص می‌باشد (اسم فامیل سر مربی یک تیم ورزشی است). واضح است که این کلمه ترجمه درست "wimbledon" نیست. این در حالی است که ترجمه‌ی درست آن یعنی کلمه فارسی "ویمبلدون" در رده پایین‌تری از ترجمه‌ها استخراج شده است، بدان معنی که امتیاز همبستگی آن نسبت به دیگر کلمات نادرست کمتر بوده است. لازم به ذکر است که این کلمات نامرتبط می‌تواند به دلیل این باشد که کلمه "wimbledon" در هم‌ترازی‌های کمی در پیکره‌ی تطبیقی آمده‌است و درنتیجه نمی‌توان با دقت کافی ترجمه‌ی درست آن را با روش‌های آماری استخراج کرد. به عبارت دیگر، هدف ما در این مرحله این است که با استفاده از شبکه‌ی ارتباطی کلمات و ارتباط همسایگی‌های آنها با یکدیگر کیفیت ترجمه‌ها را بهبود دهیم. همان گونه که در شکل نیز مشاهده می‌شود همسایگی "wimbledon" در زبان انگلیسی بسیار با همسایگی "ویمبلدون" در زبان فارسی مرتبط هستند. این در حالی است که اگر چه درجه‌ی ارتباط اولیه "wimbledon" با "هولیه" بیشتر است، ولی همسایگی آن دو ارتباط کمتری دارند و احتمالاً "هولیه" نمی‌تواند ترجمه‌ی درستی برای "wimbledon" باشد. بنابراین هدف ما از ارائه یک روش مبنی بر شبکه اصطلاحات این است که بتوان ترجمه‌های نادرست را جریمه کرده و به ترجمه‌های درست که همسایگی‌شان با همسایگی کلمه مبدأ ارتباط بیشتری دارد، اهمیت بیشتری داده شود.

به این منظور ما در این پژوهش با استفاده از شبکه TAN (Term Association Network)، تابع عمومی f را برای به روزرسانی یال‌های ارتباط بین اصطلاحات زبان مبدأ و مقصد s و t به صورت زیر ارائه می‌کنیم:

$$w(s,t) = f(w_0(s,t), sim(N_S(s), N_T(t))) \quad (۴.۴)$$

که در آن $w_0(s,t)$ وزن ابتدایی محاسبه شده بین s و t توسط روش پایه COCOT است، $N_S(s)$ مجموعه‌ای از اصطلاحات زبان مبدأ در همسایگی s و $N_T(t)$ مجموعه اصطلاحات مقصد در همسایگی t هستند و $sim(N_S(s), N_T(t))$ شباهت بین همسایگی آن دو را نشان می‌دهد که امتیاز ارتباط اصطلاحات مبدأ در $N_S(s)$ با اصطلاحات مقصد در $N_T(t)$ است. در این رابطه به منظور محاسبه شباهت همسایگی‌ها، مسیرهای بین آن دو اصطلاح (s و t) را که از

[1] mutual information



شکل ۲.۴: ساختار شبکه وابستگی اصطلاحات

همسایگی‌هایشان می‌گذرد، در نظر گرفته‌ایم. به صورت شهودی می‌توان گفت که هر چه تعداد پیوندهای با وزن بالا بین همسایگی‌ها بیشتر باشد، دو همسایگی شباهت بیشتری دارند. در آزمایش‌هایی که در این پژوهش انجام داده‌ایم، یک حالت خاص از این مدل را در نظر گرفته‌ایم که از وزن تمام مسیرهای به طول یک، دو و سه بین دو اصطلاح مشخص مبدأ و مقصد برای محاسبه شباهت همسایگی‌ها استفاده می‌کند. وزن ترجمه‌ی به روزرسانی شده توسط حالت خاص استفاده شده بین دو اصطلاح مبدأ و مقصد s و t به صورت زیر محاسبه می‌شود:

$$\begin{aligned} w(s,t) &= f(w_0(s,t), sim(N_S(s), N_T(t))) \\ &= \alpha w_0(s,t) \\ &+ \beta(\sum_{s_i \in N_S(s)} w_{MI}(s,s_i).w_0(s_i,t) + \sum_{t_i \in N_T(t)} w_0(s,t_i).w_{MI}(t_i,t)) \\ &+ \gamma \sum_{s_i \in N_S(s), t_i \in N_T(t)} w_{MI}(s,s_i).w_0(s_i,t_i).w_{MI}(t_i,t) \end{aligned} \qquad (۵.۴)$$

در این رابطه w_{MI} وزن نرمال‌شده اطلاعات متقابل بر اساس رخداد همزمان اصطلاحات یک زبان است و α، β و γ به ترتیب ضریب مسیرهای به طول یک، دو و سه می‌باشند. همچنین w_0، وزن نرمال‌شده اصطلاحات در زبان‌های مختلف است که با استفاده از روش پایه استخراج ترجمه COCOT محاسبه می‌شود. همان طور که در بخش ۱.۱.۴ ذکر شد، در روش COCOT شباهت بین اصطلاح مبدأ s به اصطلاح زبان مقصد t به صورت $w_c(s \rightarrow t)$ نشان داده شده است.

از طرف دیگر مشاهده کرده‌ایم که می‌توان شباهت بین اصطلاحات زبان مقصد را نیز با اصطلاحات زبان مبدأ در نظر گرفت (یعنی $(w_c(t \rightarrow s)$). بنابراین وزن جدید $w_0(s,t)$ را تعریف کرده‌ایم که ترکیبی از شباهت دو طرفه اصطلاحات زبان مبدأ و مقصد است:

$$w_0(s,t) = g(w_c(s \rightarrow t), w_c(t \rightarrow s)) \quad (۶.۴)$$

در آزمایش‌هایمان، مجموع وزن‌های نرمال‌شده شباهت‌ها را به عنوان تابع g استفاده کرده‌ایم.

اطلاعات متقابل

در شبکه وابستگی اصطلاحات، به منظور محاسبه وزن یال‌های بین دو اصطلاح در یک زبان، از امتیاز اطلاعات متقابل آن دو اصطلاح در مجموعه اسناد استفاده می‌کنیم. دو اصطلاح s_i و s_k را در یک زبان در نظر بگیرید. اطلاعات متقابل هموار شده[۱] آنها به صورت زیر محاسبه می‌شود [۱۳]:

$$p_{11} = \frac{0.5+xy}{2+n} \quad , \quad p_{10} = \frac{x-xy+0.5}{2+n}$$

$$p_{01} = \frac{y-xy+0.5}{2+n} \quad , \quad p_{00} = 1-(p_{11}+p_{10}+p_{01}) \quad (۷.۴)$$

$$p_x = \frac{0.5+x}{1+n} \quad , \quad p_y = \frac{0.5+y}{1+n} \quad (۸.۴)$$

$$\begin{aligned} MI(s_i,s_k) &= p_{11}\log\frac{p_{11}}{p_xp_y} + p_{10}\log\frac{p_{10}}{p_x(1-p_y)} \\ &+ p_{01}\log\frac{p_{01}}{(1-p_xp_y)} + p_{00}\log\frac{p_{00}}{(1-p_x)(1-p_y)} \end{aligned} \quad (۹.۴)$$

در این روابط x و y بسامد اسنادی هستند که به ترتیب شامل اصطلاحات s_i و s_k می‌باشند، xy نیز بسامد اسنادی است که هر دو اصطلاح در آنها آمده و n تعداد کل اسناد مجموعه در آن زبان خاص می‌باشد.

نرمال‌سازی وزن یال‌های شبکه

از آنجایی که شبکه ساخته شده شامل یال‌هایی با وزن‌های به دست آمده در بازه‌های متفاوتی می‌باشد، همانند وزن ارتباط متقابل بین اصطلاحات یک زبان و وزن ارتباط ترجمه استخراج شده بین اصطلاحات در زبان‌های مختلف، به منظور

[۱] smoothed

ترکیب آنها نیاز است که به گونه‌ای وزن‌ها نرمال شوند. یک روش نرمالسازی پایه، استفاده از تخمین احتمال[۱] امتیازات به صورت ساده است. ولی یکی از معایب این روش این است که به وزن‌های پایین بهای زیادی می‌دهد. به طور شهودی در ارتباطات استخراج شده بین اصطلاحات، هر چه وزنِ ارتباطی بیشتر باشد، آن یال بیشتر قابل اطمینان است. بنابراین مقدار نرمال شده باید با کم شدن وزن شدیدتر کم شود. به منظور جریمه وزن‌های کم، از تبدیل نمایی استفاده می‌کنیم. فرض کنید s یک اصطلاح مبدأ است و $t_1, ..., t_N$ بالاترین N اصطلاح مرتبط به آن در زبان مقصد به ترتیب با وزن‌های $r_1, ..., r_N$ باشند (یعنی $r_i = r(s, t_i), 1 < i < N$). به منظور نرمال‌سازی، تخمین احتمالات را به صورت زیر محاسبه می‌کنیم:

$$w_c^N(s \to t_i) = \frac{(\frac{e}{b})^{r_i}}{\sum_{j=1}^{N}(\frac{e}{b})^{r_j}} \quad (۱۰.۴)$$

که b پارامتری برای کنترل جریمه وزن‌های پایین است و هر چه مقدار b بیشتر باشد، وزن‌های پایین را بیشتر جریمه می‌کند. ما در آزمایش‌هایمان b را برابر ۲ در نظر گرفته‌ایم. لازم به ذکر است که به منظور نرمال‌سازی اطلاعات متقابل نیز از همین روش استفاده می‌کنیم.

۳.۱.۴ بررسی صحّت ترجمه

در ترجمه‌هایی که از پیکره‌ی تطبیقی به دست آورده‌ایم، بعضی کلمات هستند که ترجمه‌های استخراج شده‌ی آنها درست نمی‌باشد. در نتیجه اگر از آنها برای ترجمه کلمات پرس‌وجو استفاده شود، وجود ترجمه‌های نادرست باعث می‌شود پرس‌وجو ترجمه‌ی متفاوتی با عبارت اولیه‌اش شده و مفهوم اصلی خود را از دست بدهد و یا حتی در مواردی کاملاً تغییر یابد. در این صورت، استفاده از این اطلاعات اشتباه تأثیر بسزایی در کارآیی بازیابی اطلاعات بین زبانی دارد و از دقت آن تا حد زیادی می‌کاهد. مطلوب این است که بتوان کلماتِ استثنائی که کشف ترجمه‌شان از پیکره‌ی تطبیقی مقدور نبوده است را تشخیص داد.

به طور شهودی، به این نتیجه رسیده‌ایم که این کلمات استثنا در برخی اوقات ممکن است کلماتی باشند که در پیکره خیلی زیاد تکرار شده‌اند و لذا دارای بی‌نظمی[۲] بالایی هستند، همانند کلمات انگلیسی make، new و year، یا در برخی اوقات کلماتی هستند که در هم‌ترازی‌ها به تعداد کمی ظاهر شده‌اند همانند macao، jemal و alumni. در هر صورت منطقی است که بتوان این گونه کلمات استثنا را کشف کرد و نگذاشت ترجمه‌های استخراج شده‌ی نادرستشان

[۱] probability estimation

[۲] entropy

وارد ترجمه‌ی پرس‌وجوها شوند و دقت بازیابی اطلاعات بین زبانی را کاهش دهند. به منظور کشف چنین کلماتی از روش‌های کشف برون‌هشته[۱] استفاده می‌کنیم. کافی است که این کلمات استثنا یعنی همان برون‌هشته‌ها را توسط روش‌های کشف برون‌هشته از دیگر کلمات به طور نسبی متمایز کنیم. به طور شهودی اگر بتوان برون‌هشته‌ها را به خوبی شناسایی کرد، در این صورت حذف ترجمه‌های نادرست آنها از پرس‌وجوی ترجمه شده به طوری که نتوانند ترجمه‌های گمراه کننده را وارد کنند باید بتواند دقت ترجمه‌ها را بهبود دهد. در همین راستا اگر بتوان ترجمه‌ی آنها را از منبعی غنی همانند واژه‌نامه استخراج کرد، آنگاه به مراتب می‌توان کارآیی بسیار بهتری را برای سیستم بازیابی اطلاعات بین زبانی کسب کرد.

در این بخش، روشی را معرفی می‌کنیم که بر اساس توزیع فاصله بین وزن ارتباطات کلمات در یک زبان و ترجمه‌هایشان در زبان دیگر کلماتی که درست ترجمه نشده‌اند (برون‌هشته‌ها) را تشخیص می‌دهد. بدین منظور از نسخه تغییر یافته‌ی روش ارائه شده توسط پاپادمیتریو[۲] در مرجع [۶۰] استفاده کرده‌ایم. در این پژوهش [۶۰]، نگارندگان روشی را به عنوان کشف برون‌هشته‌های متقاطع [۳] با استفاده از معیارهای احتمالی برای تشخیص خودکار برون‌هشته‌ها ارائه داده‌اند. تا پیش از پژوهش انجام شده در مرجع [۶۰]، روش‌های کشف برون‌هشته بر این فرض بودند که تنها یک مجموعه از مشاهدات وجود دارد و هدف این است که بتوان آنهایی که از بقیه متفاوت هستند و به نظر مشکوک می‌آیند، پیدا شوند.

در این پژوهش، از روش ارائه شده در [۶۰] به منظور کشف برون‌هشته‌ها، کلماتی که ترجمه درستی برایشان پیدا نشده است، استفاده می‌کنیم. بنابراین از روش ارائه شده برای کشف برون‌هشته‌ها برای بررسی صحّت ترجمه[۴] استفاده شده است. مساله مطرح شده بدین صورت است که می‌خواهیم کلمات برون‌هشته در یک زبان (L_1) را با در نظر گرفتن ترجمه‌های استخراج شده برای آنها در زبان دیگر (L_2) پیدا کنیم. بنابراین می‌خواهیم کلماتی در L_1 که بر اساس ارتباطشان با کلمات L_2 مظنون هستند را مشخص کنیم. به طور شهودی، اگر توزیع وزن‌های ترجمه‌ی یک کلمه از وزن‌های ترجمه‌های همسایگانش در زبان خودش متفاوت باشد، به عنوان برون‌هشته در نظر گرفته می‌شود. به عنوان مثال، توزیع‌های وزن‌ها در همسایگی‌های کلمات ”persian“ و ”tehran“ در شکل ۴.۳ را در نظر بگیرید. در این شکل کلمات زبان L_1 با ”$\times$“ و کلمات زبان L_2 با ”$\square$“ نشان داده شده‌اند. همچنین ارتباط ترجمه‌های کلمات با خطوط بین آنها نشان داده شده‌اند و هر چه وزن ترجمه‌های استخراج شده بیشتر باشد، این خطوط پررنگ‌تر نمایش داده شده‌اند. همان

---

[۱] outlier detection

[۲] Papadimitriou

[۳] cross-outlier detection

[۴] Translation Validity Check

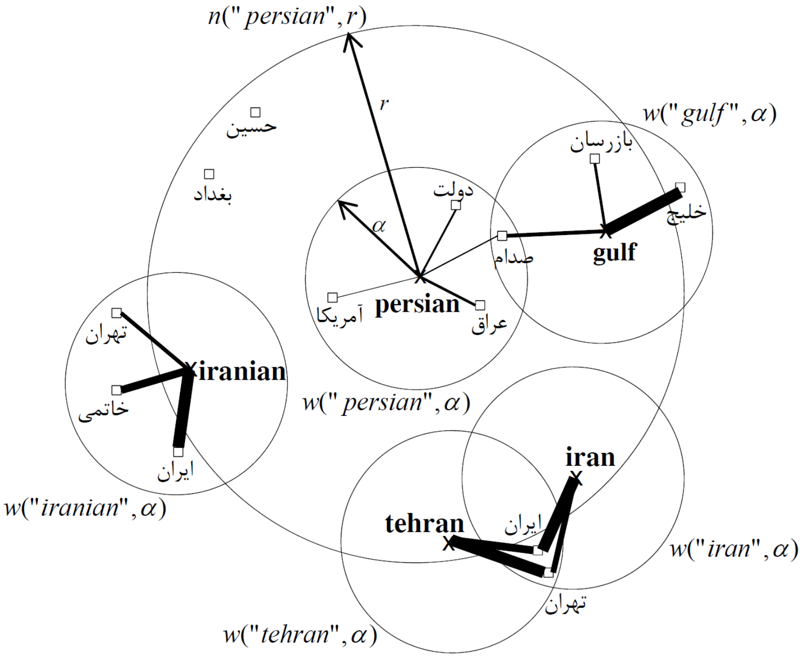
طور که در شکل ۴.۴(آ) می‌توان دید، مرتبط‌ترین کلمات فارسی استخراج شده توسط روش شبکه وابستگی اصطلاحات برای کلمه“persian” کلمات ”عراق“، ”دولت“، ”صدام“ و ”آمریکا“ می‌باشند که با وزن‌های کمی با کلمه ”persian“ ارتباط دارند. این درحالیست که کلمات فارسی مرتبط با نزدیکترین همسایگان کلمه ”persian“ در بخش انگلیسی پیکره یعنی کلمات ،”gulf“ ،”iran“ و”tehran“ ”iranian“ با وزن بالایی با این کلمات ارتباط دارند. به عنوان مثال، مرتبط‌ترین کلمات فارسی یافت شده برای ”iranian“ کلمات ”ایران“، ”خاتمی“ و ”تهران“ می‌باشند که با وزن بالایی به کلمه ”iranian“ متصل شده‌اند. بصورت شهودی می‌توان گفت از آنجایی که که الگوی رفتاری کلمه ”persian“ در شبکه وابستگی اصطلاحات با کلمات مرتبط با آن در فارسی، متفاوت از الگوی رفتاری نزدیکترین همسایگانش در انگلیسی با کلمات مرتبط آنها در فارسی است، لذا کلمه ”persian“ بعنوان برون هشته شناسایی می‌شود. بنابراین می‌توان نتیجه گرفت که به احتمال زیاد برای کلمه ”persian“ با استفاده از روش شبکه کلمات، ترجمه‌های درستی پیدا نشده است. از طرف دیگر در شکل ۴.۴(ب)، برای کلمه ”tehran“ ترجمه‌های خوبی با وزن‌های بالایی پیدا شده است. همچنین رفتار همسایگانش نیز به همین منوال است. با مقایسه این دو مثال، می‌توان به این نتیجه رسید که کلماتی که رفتارشان با رفتار همسایگانشان متفاوت است، کلمات استثنا هستند و به احتمال زیاد ترجمه‌ی درستی برایشان استخراج نشده است.

جزئیات روش کشف برون‌هشته استفاده شده در ادامه آورده شده است. فرض کنید کلمه t متعلق به زبان L_1 است. ابتدا همسایگان t را در زبان L_1 (یعنی $N_{L_1}(t,r)$) که در شعاع همسایگی r با آن قرار دارد جمع‌آوری می‌کنیم. برای هر کلمه در همسایگی t، یک همسایگی محلی با شعاع α تعریف می‌کنیم که شامل ترجمه‌های استخراج شده برای آن در زبان L_2 می‌باشد. شعاع‌های همسایگی در شکل ۳.۴ مشخص شده‌اند. سپس میانگین وزن‌های محلی بر روی تمام کلماتی که در همسایگی r هستند را به صورت زیر حساب می‌کنیم:

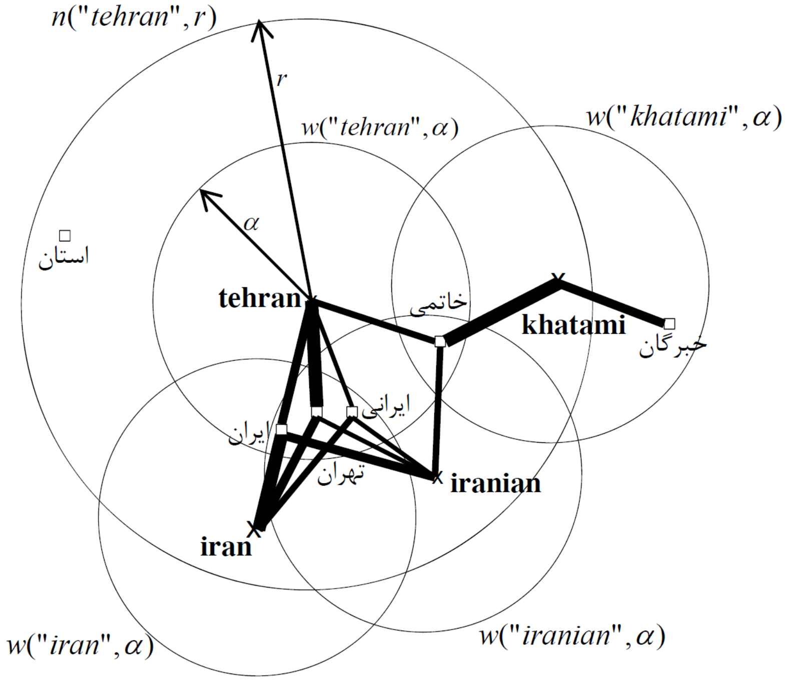
$$\hat{w}_{L_1,L_2}(t,r,\alpha) := \frac{\Sigma_{q\in N_{L_1}(t,r)} w_{L_2}(q,\alpha)}{n_{L_1}(t,r)} \qquad (۱۱.۴)$$

که $w_{L_2}(q,\alpha)$ مجموع وزن‌های ترجمه‌های استخراج شده برای کلمه q در زبان L_2 در شعاع α است و $n_{L_1}(t,r)$ تعداد همسایه‌های کلمه t در شعاع همسایگی r می‌باشد ($N_{L_1}(t,r)$). به عنوان مثال در شکل ۳.۴، $n_{L1}(``persian", r)$ تعداد همسایگی کلمه ”persian“ به علاوه خودش برابر ۵ است و $w_{L_2}(q,\alpha)$ برابر مجموع وزن ترجمه‌های q می‌باشد. روش کشف برون‌هشته استفاده شده بر اساس انحراف معیار وزن محاسبه شده کلمات زبان L_2 در همسایگی α است. از این رو کمیت $\hat{\sigma}_{L_1,L_2}(t,r,\alpha)$ برای کلمه t در زبان L_1 با شعاع r و α به صورت زیر تعریف شده است:

$$\hat{\sigma}_{L_1,L_2}(t,r,\alpha) := \sqrt{\frac{\Sigma_{q\in N_{L_1}(t,r)} (w_{L_2}(q,\alpha) - \hat{w}_{L_1,L_2}(t,r,\alpha))^2}{n_{L_1}(t,r)}} \qquad (۱۲.۴)$$



(آ) “persian” برون‌هشته است



(ب) “tehran” برون‌هشته نیست

شکل ۳.۴: دو مثال برای بررسی صحّت ترجمه‌ها بر اساس کشف برون‌هشته‌ها.

در نهایت کلمه t در زبان L_1 با شعاع r در صورتی بنابر ترجمه‌های استخراج شده برایش در زبان L_2 برون‌هشته است

که رابطه زیر بر قرار باشد:

$$|\hat{w}_{L_1,L_2}(t,r,\alpha) - w_{L_2}(t,\alpha)| > k_\sigma \hat{\sigma}_{L_1,L_2}(t,r,\alpha) \quad (۱۳.۴)$$

در این رابطه مقادیر $\hat{w}_{L_1,L_2}(t,r,\alpha)$ و $\hat{\sigma}_{L_1,L_2}(t,r,\alpha)$ به ترتیب مقادیر میانگین وزن‌ها و انحراف معیار کلمه t با شعاع‌های r و α هستند که توسط روابط ۱۱.۴ و ۱۲.۴ محاسبه می‌شوند.

ما در آزمایش‌هایمان از رابطه ۱۳.۴ به طور نسبی استفاده کرده‌ایم، یعنی هر چه نسبت تفاضل وزن همسایه‌های یک کلمه از میانگین همسایگانش به انحراف معیار بیشتر باشد، آن کلمه امتیاز برون‌هشته بیشتری دارد. در نهایت امتیاز برون‌هشته تمام کلمات را به طور نزولی مرتب کرده‌ایم. در آزمایش‌های انجام شده در این پژوهش نشان می‌دهیم که میانگین و انحراف معیار بر اساس شعاع r به ما اطلاعات مفیدی را درباره همسایگی کلمات می‌دهند. بنابراین می‌توانیم بسیاری از کلماتی که به درستی ترجمه نشده‌اند را از این طریق مشخص کنیم.

روشی که ما به منظور بررسی صحّت ترجمه‌ها استفاده کرده‌ایم با روش ارائه شده در مرجع [۶۰] تفاوت‌هایی دارد. تفاوت اصلی روش ما با روش ارائه شده در مرجع [۶۰] در این است که در آن پژوهش نگارندگان از **تعداد** همسایگی‌های هر مشاهده برای تشخیص برون‌هشته بودن استفاده کرده‌اند، در حالی که ما **وزن ارتباطی** مشاهدات را در نظر گرفته‌ایم. همچنین در آن پژوهش، نگارندگان یک مقدار ثابت را برای شعاع r فرض کرده و شعاع α را برابر یک چهارم r قرار داده‌اند در حالی که ما به صورت پویا این مقادیر را برای هر مشاهده محاسبه می‌کنیم. در نهایت اینکه در آن پژوهش نگارندگان برای تشخیص نهایی برون‌هشته بودن یک مشاهده از رابطه ۱۳.۴ با مقدار ۳ برای k_σ استفاده کرده‌اند، ولی ما به طور نسبی مقادیر محاسبه شده را در نظر گرفته‌ایم. به طور کلی ما سعی کرده‌ایم هیچ‌گونه فرض ثابتی را برای مقدار متغیرها در نظر نگیریم.

۴.۱.۴ استخراج ترجمه عبارات

در پرس‌وجوهای بازیابی اطلاعات مواردی مشاهده می‌شود که از ترجمه‌ی تک تک کلمات پرس‌وجو نمی‌توان به دقت کافی بازیابی اطلاعات بین زبانی را انجام داد. به عنوان مثال پرس‌وجوی ”world cup 2002“ را در نظر بگیرید. اگر بخواهیم به صورت مجزا هر کلمه آن را ترجمه کنیم برای کلمه‌ی ”cup“ هر دو ترجمه‌ی آن، ”فنجان“ و ”جام“ را می‌توان در نظر گرفت. ولی در این پرس‌وجوی خاص مسلماً منظور ترجمه‌ی ”فنجان“ آن نبوده است. می‌توان به گونه‌ای از روی محتوای پرس‌وجو تشخیص داد که کدام ترجمه را می‌توان استفاده کرد. بنابراین از دیگر کلمات پرس‌وجو برای استخراج ترجمه درست از پیکره تطبیقی استفاده می‌کنیم. به عنوان مثال ”world cup“ را به صورت یک عبارت در نظر گرفته

و مستقیماً ترجمه‌ی آن را استخراج می‌کنیم. از آنجایی که پرس‌وجوهایی که شامل چنین عباراتی باشند کم نیستند، در این مرحله از پژوهش، روش نسبتاً متفاوتی در زمینه استخراج ترجمه‌ها ارائه می‌دهیم که بر مبنای استخراج ترجمه‌های عبارات است.

به طور کلی، در هر زبان جفت کلماتی وجود دارند که در بیشتر مواقع به همراه یکدیگر ظاهر می‌شوند همانند "جام جهانی" و "world cup". ما این جفت کلمات را به عنوان عبارات در نظر می‌گیریم و سعی داریم که بتوانیم پس از تشخیص آنها، ترجمه‌شان را از پیکره‌ی تطبیقی استخراج کنیم. در این مرحله برای استخراج ترجمه‌ی عبارات با دو چالش اصلی روبه‌رو هستیم: (۱) چگونگی تشخیص عبارت و اینکه آیا می‌توان هر دو کلمه کنار هم را به عنوان یک عبارت در نظر گرفت؟ (۲) چگونگی استخراج ترجمه عبارات از پیکره‌های تطبیقی.

در اولین گام به منظور تشخیص عبارات، تمام جفت کلماتی در یک زبان که در کنار یکدیگر با تکرار نسبتاً زیادی ظاهر می‌شوند و دارای اطلاعات متقابل بالاتری می‌باشند را به عنوان یک عبارت در نظر می‌گیریم. در ادامه در این مرحله، پیش پردازشی بر روی اسناد پیکره‌ی تطبیقی انجام می‌دهیم که در آن تمام عبارات تشخیص داده شده و به صورت یک واحد مجزا در نظر گرفته می‌شوند. لازم به ذکر است که دو نگرش را در مرحله پیش‌پردازش اسناد پیش گرفته‌ایم:

- اول اینکه سعی کرده‌ایم تمام دو کلمه‌ای‌های کنار یکدیگر را به صورت عبارت در نظر بگیریم یعنی هر سند تنها متشکل از عبارات شده است.
- در نگرش دوم تنها هنگامی دو کلمه مجاور هم را عبارت در نظر می‌گیریم که دارای اطلاعات متقابل بالایی باشند در ضمن در اسناد پیش پردازش شده هم عبارات و هم تک تک کلماتی که عبارات را تشکیل می‌دهند قرار گرفته‌اند. بدین صورت به هر سند یک‌سری عبارات اضافه شده است. هدف از ارائه نگرش دوم این است که بتوان ترجمه عباراتی از دو زبان را بدست آورد که در زبان دیگر تنها یک کلمه هستند، مثلاً دو کلمه *رئیس* و *جمهور* را در نظر بگیرید که اگر به صورت یک عبارت بیان شوند، تنها یک ترجمه در زبان انگلیسی یعنی *president* دارند. از طرف دیگری کلمه *آمریکا* را در نظر بگیرید که در زبان انگلیسی به دو کلمه *United States* ترجمه می‌شود.

پس از اینکه اسناد فارسی و انگلیسی به صورت مجموعه‌ای از عبارات ساخته شدند، می‌توانیم روش‌های استخراج ترجمه از پیکره تطبیقی را اعمال کنیم. تنها تفاوت در اجزای اسناد است که می‌توانند کلمه و یا عبارت باشند، یعنی به عبارتی روش‌های استخراج ترجمه اصطلاحات را در این مرحله بکار می‌بریم. نتایج این مرحله نشان می‌دهد که می‌توان بر روی استخراج ترجمه‌های عبارات بیشتر تکیه کرد زیرا که پرس‌وجوهایی که شامل این گونه ترکیب کلمات باشند کم نیستند و استخراج این گونه از اطلاعات می‌تواند روشی ارزشمند در بازیابی اطلاعات بین زبانی باشد.

۲.۴ ساخت مدل زبانی پرس‌وجو

در این بخش، سه روش برای ترجمه‌ی پرس‌وجوها با استفاده از ترجمه‌های استخراج شده و بر اساس ساخت مدل زبانی پرس‌وجو[۱] ارائه می‌دهیم. در دو روش اول به طور مستقیم هر کلمه‌ی پرس‌وجو را با ترجمه‌های استخراج شده مربوط به آن جایگزین می‌کنیم. ولی در روش سوم، به منظور ساخت مدل‌های زبانی پرس‌وجو، روشی را ارائه داده‌ایم که از ترجمه‌های واژه‌نامه علاوه بر دانش ترجمه‌ی استخراج شده از پیکره‌ی تطبیقی، استفاده می‌کند. در ادامه به ارائه‌ی جزئیات بیشتری درباره هر روش می‌پردازیم.

۱.۲.۴ انتخاب تعداد همسایه ثابت

فرض کنید پرس‌جوی Q در زبان L_1 داده شده است. برای هر کلمه‌ی پرس‌وجو، به تعداد ثابتی (k) از بالاترین ترجمه‌های استخراج شده برای آن در زبان L_2 را انتخاب کرده و مدل زبانی کل پرس‌وجو را در زبان L_2 می‌سازیم. در مدل ارائه شده فرض کرده‌ایم که تمام کلمات پرس‌وجو به یک اندازه اهمیت دارند، بنابراین وزن‌های مساوی در مدل زبانی پرس‌وجو خواهند داشت. به عنوان مثال اگر یک پرس‌وجو دارای سه کلمه است، هر یک از آنها به طور مساوی وزن $1/3$ خواهد داشت. علاوه بر این، وزن تأثیر هر کلمه‌ی ترجمه بر اساس احتمال تخمین‌زده شده ترجمه‌ی آن کلمه است که مقدار آن از امتیازهای ترجمه‌های استخراج شده محاسبه می‌شود.

در این پژوهش، از روش ارائه شده در مرجع [۷۷] برای ساخت ترجمه‌ی پرس‌وجو در زبان L_2 استفاده می‌کنیم. با فرض $Q = q_1...q_n$ در زبان L_1 و $t_1, ..., t_k$ بالاترین k کلمه مرتبط به کلمه پرس‌وجوی q_i (که $q_i \in Q$) در زبان L_2، مدل زبانی پرس‌وجو در L_2 به صورت زیر ساخته می‌شود:

$$P(t|Q) = \sum_{i=1}^{n} \frac{1}{n} \frac{p(t|q_i)}{\sum_{j=1}^{k} p(t_j|q_i)} \tag{۱۴.۴}$$

که در آن $p(t_l|q_j) > 0$، اگر t_l بین k بالاترین ترجمه‌ی کلمه q_j باشد و در غیر این صورت $p(t_l|q_j) = 0$. همچنین می‌توان این روش را برای ساخت مدل زبانی پرس‌وجوها توسط عبارات نیز به کار برد. به علاوه، از این روش در موارد حذف برون‌هشته‌ها نیز استفاده می‌کنیم. تفاوت ساخت مدل زبانی پرس‌وجو در روش حذف برون‌هشته‌ها در این است که کلماتی از پرس‌وجو که برون‌هشته تشخیص داده شده‌اند، هنگام ساخت مدل زبانی پرس‌وجو در نظر گرفته نمی‌شود. به عنوان مثال، اگر پرس‌وجو شامل سه کلمه باشد و یکی از آنها برون‌هشته شناسایی شود، ما مدل زبانی را تنها برای دو

---

[۱] query language model

کلمه باقیمانده می‌سازیم بدون اینکه کلمه برون‌هشته را در نظر بگیریم.

۲.۲.۴ انتخاب تعداد همسایه پویا

در روش پایه ساخت ترجمه‌ی پرس‌وجو در بخش قبل، برای تمام کلمات پرس‌وجو به طور برابر به تعداد ثابت (k) از بالاترین ترجمه‌های استخراج شده در زبان L_2 انتخاب شده‌اند. اما محدوده‌ی کیفیت ترجمه برای هر کلمه متفاوت است. برای مثال، همان طور که در شکل ۴.۴ دیده می‌شود، در لیست ترجمه‌های استخراج شده برای کلمه "sanction"، تنها کلمه‌ی اول لیست یعنی "تحریم" ترجمه‌ی درست می‌باشد، اما برای کلمه "atomic" تمام هفت ترجمه‌ی استخراج شده‌ی اول نسبتاً مربوط می‌باشند و می‌توانند به عنوان ترجمه‌ی پرس‌وجو و برای گسترش آن استفاده شوند. بنابراین اگر به تعداد برابر ترجمه برای هر کلمه پرس‌وجو انتخاب کنیم، یک سری اطلاعات مفید از دست خواهیم داد. به عنوان مثال اگر k را برابر ۳ در نظر بگیریم، آنگاه برای بعضی کلمات همانند "sanction" تمام کلمات ترجمه‌ی انتخاب شده ممکن است درست نباشند. مثلا "عراق" و "سازمان" نمی‌توانند ترجمه‌های خوبی برای "sanction" باشند. از طرف دیگر برای بعضی کلمات دیگر همانند "atomic" ممکن است آن مقدار ثابت کافی نباشد و ما اطلاعات مفیدی را اشتباهاً نادیده بگیریم. مثلاً تمام هفت کلمه‌ی اول ترجمه‌های استخراج شده برای "atomic" (یعنی اتمی، آژانس، البرادعی، هسته، انرژی، اورانیوم و غنی) تا حدودی به آن مربوط هستند و بهتر است که در پرس‌وجوی گسترش‌یافته ترجمه شده آورده شوند. در نتیجه از توزیع امتیازهای ترجمه‌های استخراج شده، استفاده کرده و تعداد همسایه هر کلمه را به طور پویا محاسبه کنیم.

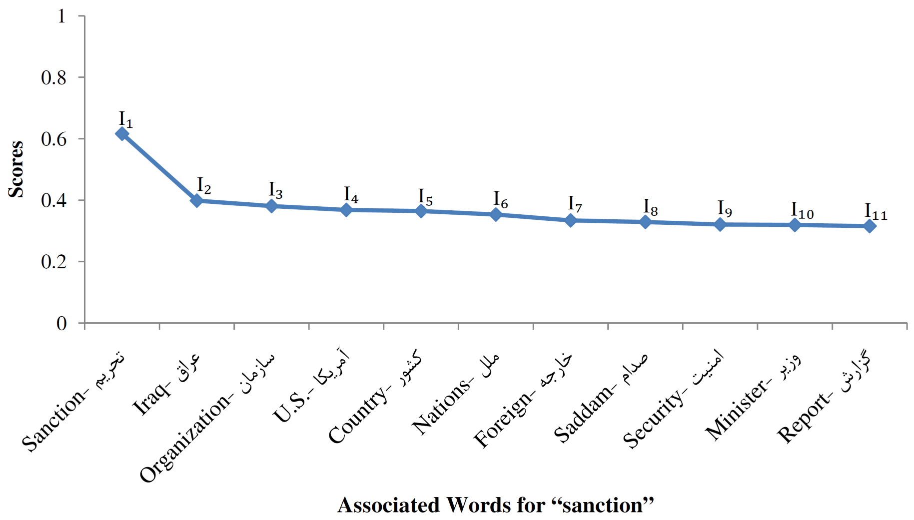
بر اساس توزیع امتیازهای ترجمه‌های استخراج شده برای هر کلمه، روشی را ارائه می‌دهیم که بیشترین شیب نمودار را پیدا کرده و تعداد نقاطی که قبل از نقطه شکست وجود دارند را به عنوان بهترین تعداد برای انتخاب کلمات مرتبط انتخاب می‌کنیم. در روش مذکور ابتدا تفاضل دو امتیاز متوالی را بدست می‌آوریم:

$$\triangle_i \quad = \quad I_i - I_{i+1} \tag{۱۵.۴}$$

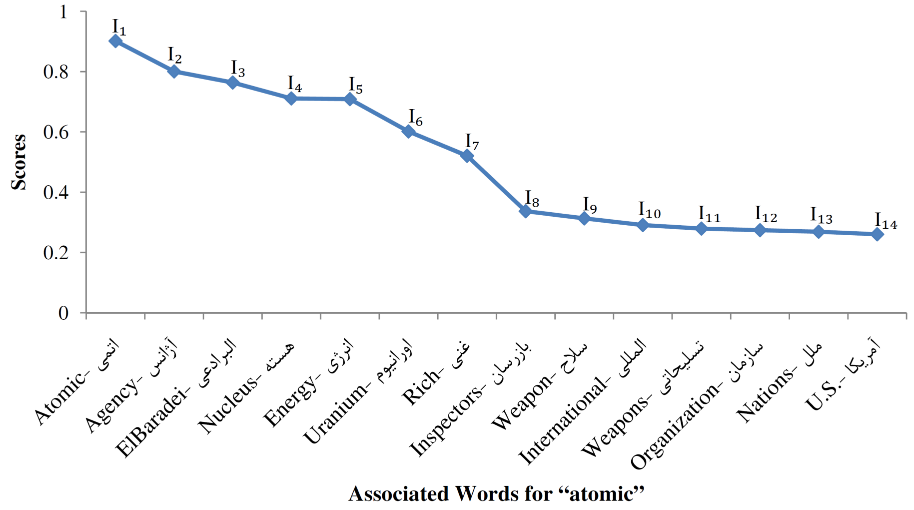
در این رابطه I_i امتیاز ارتباط ترجمه‌ی دو کلمه و $\triangle_i$ تفاضل دو امتیاز متوالی است. سپس نقطه‌ای را که دارای بیشترین شیب است به صورت زیر مشخص می‌کنیم:

$$\hat{k} = argmax_{i=1}^{n}(\frac{\Sigma\triangle_i}{i+1}) \tag{۱۶.۴}$$

که n برابر تعداد ترجمه‌های استخراج شده است و $\hat{k}$ تعداد بهترین ترجمه‌های آن کلمه می‌باشد. پس از انتخاب همسایه‌های هر کلمه به تعداد متغیر $\hat{k}$ برای هر کلمه‌ی پرس‌وجو، مدل زبانی پرس‌وجو را بر اساس روش توضیح داده



(آ)



(ب)

شکل ۴.۴: انتخاب تعداد همسایه پویا برای کلمات انگلیسی ”sanction“ و ”atomic“

شده در بخش ۱.۲.۴ می‌سازیم.

لازم به ذکر است که از روش انتخاب همسایه پویای ارائه شده نه تنها برای انتخاب تعداد همسایه‌ها برای ترجمه‌های یک کلمه استفاده می‌شود بلکه برای انتخاب تعداد همسایه‌های کلمات در یک زبان نیز استفاده می‌شود. در آزمایش‌های انجام شده در این پژوهش، از این روش برای انتخاب تعداد متغیر همسایه‌های هر کلمه در آن زبان که با اطلاعات متقابل تعیین شده‌اند، استفاده می‌کنیم.

۴.۲.۳ استفاده از واژه‌نامه

از جمله مزایای پیکره‌های تطبیقی این است که می‌توان توسط آنها ترجمه‌ی کلمات خارج از واژگان [1] را بدست آورد. همچنین علاوه بر کلمات خارج از واژگان، برای سایر کلمات نیز می‌توان کلمات مربوط (نه تنها ترجمه) را استخراج کرد. در این راستا، به منظور غنی‌تر کردن منبعی همانند واژه‌نامه، ترجمه‌های بدست آمده از پیکره‌ی تطبیقی را با ترجمه‌های واژه‌نامه با چندین روش مختلف ترکیب کرده و مدل زبانی پرس‌وجو را به منظور انجام بازیابی اطلاعات بین زبانی می‌سازیم. در روش‌های ترکیبی به ترجمه‌های بدست آمده از واژه‌نامه نیز وزنی می‌دهیم که بتوان در مدل زبانی پرس‌وجو از آنها استفاده کرد. در ادامه‌ی این بخش به ارائه این روش‌های ترکیبی می‌پردازیم.

واژه‌نامه به همراه ترجمه‌های استخراج شده از پیکره تطبیقی (Dic&CC) در این روش ترکیبی، به منظور ترجمه‌ی هر کلمه پرس‌وجو، هم از ترجمه‌های واژه‌نامه‌ی کلمه مربوطه و هم از ترجمه‌های استخراج شده توسط پیکره تطبیقی برای آن استفاده می‌کنیم. از آنجایی که واژه‌نامه منبعی قابل اطمینان است، ما به ترجمه‌های آن وزن بیشتری نسبت به ترجمه‌های استخراج شده از پیکره تطبیقی اختصاص می‌دهیم.

فرض کنید s کلمه‌ای در زبان مبدأ است و $t_{cc_1},...,t_{cc_k}$ کلماتی در زبان مقصد می‌باشند که بیشترین همبستگی را با s دارند. کلمات $t_{cc_1},...,t_{cc_k}$ بر اساس ارتباطشان با s از پیکره تطبیقی استخراج شده و هر یک دارای وزنی نرمال‌شده می‌باشند که به صورت $r_{cc_1},...,r_{cc_k}$ نمایش می‌دهیم. (یعنی $r_{cc_i}=r(s,t_{cc_i}), 1\leq i\leq k$). همچنین فرض کنید برای s به تعداد k' ترجمه در زبان مقصد در واژه‌نامه وجود دارد ($t_{dic_1},...,t_{dic_{k'}}$). هدف این است که تمام ترجمه‌های کلمه s را (به تعداد $k+k'$) در مدل زبانی پرس‌وجو در نظر بگیریم. از آنجایی که ترجمه‌های واژه‌نامه وزن ندارند، به هر یک از آنها وزنی برابر اختصاص می‌دهیم ($r_{dic_i}, 1\leq i\leq k'$) و این وزن را نیز بیشتر از بیشترین وزن ترجمه استخراج شده از پیکره تطبیقی در نظر می‌گیریم ($r_{dic_i}>r_{cc_1}, 1\leq i\leq k'$). به طور کلی، وزن‌هایی که به ترجمه‌های واژه‌نامه می‌دهیم به صورت تابعی از امتیازهای ترجمه‌های پیکره تطبیقی هستند و از آنجایی که می‌خواهیم وزن ترجمه‌های واژه‌نامه بیشتر از ترجمه‌های پیکره تطبیقی باشد، وزن ترجمه‌های استخراج شده از واژه‌نامه یعنی r_{dic_i} را به صورت زیر محاسبه می‌کنیم:

$$r_{dic_i}=r_{cc_1}+c.(r_{cc_1}-r_{cc_2}) \quad , \quad r_{cc_1}\geq r_{cc_2} \tag{۴.۱۷}$$

که r_{cc_1} و r_{cc_2} امتیاز اولین و دومین ترجمه‌های استخراج شده از پیکره تطبیقی برای کلمه s هستند و c یک ضریب ثابت

[1] Out Of Vocabulary

عددی است که ما در آزمایش‌هایمان آن را برابر دو در نظر گرفته‌ایم. پس از اینکه به هر یک از ترجمه‌های واژه‌نامه برای یک کلمه پرس‌وجو به نسبت امتیاز ترجمه‌های استخراج شده آن از پیکره تطبیقی، وزنی تعلق گرفت، می‌توان بر اساس روش ارائه شده در بخش ۱.۲.۴ مدل زبانی پرس‌وجو را برای تمام ترجمه‌های کلمات پرس‌وجو ساخت.

استفاده از پیکره تطبیقی برای کلمات خارج از واژه‌نامه (Dic-CC) در این روش تمام کلمات پرس‌وجو را توسط واژه‌نامه ترجمه می‌کنیم و سپس اگر کلمه پرس‌وجویی در واژه‌نامه نبود، آن را توسط پیکره تطبیقی ترجمه می‌کنیم. هدف از این روش این است که بتوان کیفیت واژه‌نامه‌های دو زبانه را بهبود داد. فرض کنید پرس‌وجوی Q شامل n کلمه به زبان L_1 است: $Q = q_1...q_n$ و $t_1, ..., t_k$ بالاترین k کلمه مرتبط به کلمه پرس‌وجوی q_i (که $q_i \in Q$) در زبان L_2 می‌باشند. هر کلمه پرس‌وجو یا از طریق واژه‌نامه ترجمه می‌شود و یا توسط پیکره تطبیقی، در نتیجه مدل زبانی آن به صورت زیر ساخته می‌شود:

$$P(t|Q) = \begin{cases} \frac{1}{n\times k} & \text{if}\, t \in Dic \\ \\ \sum_{i=1}^{n} \frac{1}{n} \frac{p(t|q_i)}{\sum_{j=1}^{k} p(t_j|q_i)} & \text{otherwise} \end{cases} \qquad (۱۸.۴)$$

که در این رابطه k برای کلماتی که ترجمه‌شان از پیکره تطبیقی استخراج می‌شود بر اساس روش انتخاب همسایه پویا (ارائه شده در بخش ۲.۲.۴) محاسبه می‌شود. برای کلماتی که ترجمه‌شان در واژه‌نامه وجود دارد، k می‌تواند بر اساس تعداد ترجمه‌های هر کلمه متغیر باشد.

استفاده از واژه‌نامه برای ترجمه برون‌هشته‌ها (CC-Dic) در این روش، ابتدا کلمات پرس‌وجو توسط پیکره تطبیقی ترجمه شده سپس کلمات پرس‌وجویی که به عنوان برون‌هشته تشخیص داده شده‌اند، توسط واژه‌نامه ترجمه می‌شوند. مدل زبانی پرس‌وجو نیز همانند بخش قبل ساخته می‌شود.

فصل ۵

ارزیابی نتایج آزمایش‌ها

در این بخش به ارائه نتایج آزمایش‌های بازیابی اطلاعات بین زبانی توسط اعمال روش‌های ارائه شده در فصل ۴ می‌پردازیم. لازم به ذکر است که روش‌های ارائه شده در این پژوهش مستقل از زبان‌های پیکره تطبیقی بوده و می‌توان آنها را بر روی دیگر زبان‌ها نیز اعمال کرد. در این پژوهش آزمایش‌هایمان را بر روی پیکره همشهری و پرس‌وجوهای دو زبانه فارسی و انگلیسی آن (استفاده شده در CLEF 2008, 2009[۱]) انجام داده‌ایم. از آنجایی که پیکره تطبیقی ساخته‌شده ما نیز از اخبار روزنامه همشهری ایجاد شده است (اگر چه تا حدودی سال‌های متفاوتی را پوشش می‌دهند)، روش‌های ارائه شده را بر روی پیکره‌ای دیگر و با مبدأ متفاوتی نیز اعمال کرده‌ایم، تا بتوان از کارآیی روش‌ها اطمینان حاصل کرده و تحلیلی درست از نتایج ارائه داد. پیکره INFILE[۲] که متشکل از اخبار خبرگزاری فرانسه به سه زبان فرانسوی، انگلیسی و عربی می‌باشد، بدین منظور استفاده شده است. پیکره‌های استفاده شده و نتایج آزمایش‌ها در ادامه این بخش با جزئیات بیشتری آورده شده‌اند.

[۱] www.clef-campaign.org

[۲] http://www.infile.org/

جدول ۱.۵: خصوصیات پیکره تطبیقی ساخته شده

تعداد هم‌ترازی‌ها	۱۰۳۶۵
تعداد اسناد انگلیسی (بی‌بی‌سی) یکتا	۵۵۷۲
تعداد اسناد فارسی (همشهری) یکتا	۴۵۹۸
تعداد هم‌ترازی‌ها در یک روز	۲۲۷۶
تعداد هم‌ترازی‌ها در بازه‌ی چهار روز	۳۶۰۵
تعداد هم‌ترازی‌ها در بازه‌ی چهارده روز	۴۴۸۴

۱.۵ مجموعه‌های اسناد

۱.۱.۵ پیکره تطبیقی UTPECC

مشخصات پیکره تطبیقی ایجاد شده UTPECC در بخش ۲.۳.۳ و جدول ۱.۳ آورده شده است. جدول ۱.۵ خصوصیات دیگری از این پیکره را نشان می‌دهد. در آزمایش‌های انجام شده از پیکره تطبیقی ساخته شده برای استخراج دانش ترجمه استفاده می‌کنیم.

۲.۱.۵ پیکره همشهری

در آزمایش‌های انجام شده از داده‌های مسابقه بازیابی اطلاعات بین زبانی CLEF-2008 و CLEF-2009، بازیابی اسناد فارسی بر اساس پرس‌وجوهای انگلیسی، استفاده کرده‌ایم. مجموعه اسناد این مسابقه شامل ۱۶۶،۷۷۴ خبر به حجم ۵۷۸ مگا بایت است که در روزنامه همشهری بین سال‌های ۱۹۹۶ و ۲۰۰۲ انتشار یافته‌اند [۵]. این پیکره شامل ۱۰۰ پرس‌وجو به فارسی و ترجمه‌ی انگلیسی آنها است. ما در این پژوهش از ۸۵ پرس‌وجوی آن استفاده کرده‌ایم. از آنجایی که ۱۵ پرس‌وجوی دیگر به طور خاص مربوط به زبان فارسی بوده و کلمات ترجمه‌ی آنها به زبان انگلیسی در پیکره تطبیقی ایجاد شده وجود نداشت، همانند "کنسرت شجریان" و "بازسازی تونل کندوان"، آنها را در نظر نگرفته‌ایم. پرس‌وجوهای فارسی برای بازیابی اطلاعات تک‌زبانی استفاده شده‌اند.

۳.۱.۵ پیکره INFILE

پیکره INFILE (INformation, Filtering, Evaluation) در مسابقه‌های CLEF-2008، CLEF-2009 و کارگاه LREC2010[1] برای ارزیابی سیستم‌های تصفیه توافقی بین زبانی[2] استفاده شده است [۱۰]. این مجموعه شامل ۳۰۰،۰۰۰ خبر از خبرگزاری فرانسه[3] بین سال‌های ۲۰۰۴ و ۲۰۰۶ به سه زبان عربی، انگلیسی و فرانسوی است. همچنین مجموعه‌ای متشکل از ۵۰ موضوع عمومی و خاص (همانند اطلاعات علمی و فنی) بر روی آن وجود دارد. به منظور انجام بازیابی اطلاعات بین زبانی برای بازیابی اسناد انگلیسی با استفاده از پرس‌وجوهای فارسی، ما پرس‌وجوهای انگلیسی این پیکره را به طور دستی به فارسی ترجمه کرده‌ایم. بعضی از این پرس‌وجوها بسیار خاص بودند، همانند 'The slam: poetry democratized' و نمی‌توانستند به صورت یک پرس‌وجو به فارسی ترجمه شوند، بنابراین آنها را از آزمایش‌هایمان حذف کرده و در کل از ۴۵ پرس‌وجوی این مجموعه استفاده کرده‌ایم. ترجمه‌های فارسی انجام شده برای پرس‌وجوهای پیکرهINFILE در پیوست آ آمده است.

۲.۵ معیارهای ارزیابی

معیارهای کارآیی گوناگونی که معمولاً بر مبنای فراخوانی[4] و دقت[5] می‌باشند، برای ارزیابی سیستم‌های بازیابی اطلاعات استفاده می‌شوند.

فراخوانی و دقت

فراخوانی، کسری از اسناد مرتبط به پرس‌وجو است که بازیابی شده‌اند:

---

[1] http://www.infile.org

[2] cross-language adaptive filtering

[3] Agence France Presse (AFP)

[4] recall

[5] precision

$$\text{فراخوانی} = \frac{|\{\text{اسناد مربوط}\} \cap \{\text{اسناد بازیابی شده}\}|}{|\{\text{اسناد مربوط}\}|}$$

با بازیابی تمام اسناد مربوط به پرس‌وجو می‌توان به فراخوانی ۱۰۰% دست یافت. بنابراین فراخوانی به تنهایی معیاری کافی نیست بلکه تعداد اسناد نامربوط نیز مطلوب است که در نظر گرفته شود.

معیار دقت به صورت کسری از اسناد بازیابی شده مربوط، تعریف می‌شود:

$$\text{دقت} = \frac{|\{\text{اسناد مربوط}\} \cap \{\text{اسناد بازیابی شده}\}|}{|\{\text{اسناد بازیابی شده}\}|}$$

دقت را می‌توان در تعداد معینی از رتبه‌های بالای اسناد بازیابی‌شده نیز محاسبه کرد. این معیار دقت در n سند بالا، $Prec@n$، نامیده می‌شود. از آنجایی که کاربران معمولاً به ندرت اسناد بازیابی شده پایین‌تر را نگاه می‌کنند، دقت در میان اسناد بالای رتبه‌بندی شده اهمیت بسزایی دارد. در این پژوهش ما نیز دقت در رتبه ۵ و ۱۰ را برای هر پرس‌وجو محاسبه کرده و میانگین دقت در این رتبه‌ها را برای تمام پرس‌وجوها به صورت $Prec@5$ و $Prec@10$ گزارش داده‌ایم. همچنین لازم به ذکر است که معنا و استفاده‌ای که از کلمه ”دقت“ در بازیابی اطلاعات می‌شود، با تعریف آن در شاخه‌های دیگر علوم متفاوت است.

متوسط دقت

دقت و فراخوانی، معیارهایی تک مقداره[۱] می‌باشند که بر اساس لیست کلی اسناد بازیابی شده توسط سیستم محاسبه می‌گردند. برای سیستم‌هایی که دنباله‌ای از اسناد رتبه‌بندی شده برمی‌گردانند، مطلوب این است که ترتیب اسناد بازیابی شده نیز در نظر گرفته شود. متوسط دقت[۲] بر روی رتبه اسناد مربوط بازیابی شده تأکید بیشتری دارد و به صورت میانگینی از دقت‌های محاسبه شده بر اساس مکان قرارگیری هر سند مربوط در دنباله اسناد رتبه‌بندی شده تعریف می‌شود:

$$\text{AP} = \frac{\sum_{r=1}^{N}(Prec@r \times rel(r))}{\text{تعداد اسناد مربوط}}$$

[۱] single-value

[۲] Average Precision (AP)

در این رابطه r رتبه است، N تعداد کل اسناد بازیابی شده می‌باشد، $Prec@r$ دقت در رتبه داده شده و $rel()$ تابعی دودویی بر روی مرتبط بودن سندِ رتبه‌ی r است. لازم به ذکر است که هر سند مجموعه برای یک پرس‌وجوی خاص به صورت دودوییِ مربوط (۱) و یا نامربوط (۰) مقداردهی شده است و تابع $rel(r)$ این مقدار را به ازای سندی که در رتبه‌ی r قرار گرفته، بر می‌گرداند. همچنین مخرج کسر یعنی "تعداد اسناد مربوط"، تعداد اسناد مربوط در کل مجموعه است، بنابراین معیار میانگین دقت کارآیی سیستم را بر روی کل اسناد مربوط در نظر می‌گیرد.

میانگین متوسط دقت

میانگین متوسط دقت[۱] برای مجموعه‌ای از پرس‌وجوها، به صورت میانگین بر روی امتیازهای متوسط دقت برای هر پرس‌وجو است:

$$\text{MAP} = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

که در این رابطه Q تعداد پرس‌وجوها و $AP(q)$ متوسط دقت پرس‌وجوی q می‌باشد.

در این پژوهش، از معیار میانگین متوسط دقت به عنوان معیاری اصلی برای ارزیابی روش‌های مختلف بازیابی اطلاعات بین زبانی استفاده می‌کنیم.

۳.۵ نتایج بازیابی اطلاعات بین زبانی

در این بخش تأثیر اعمال روش‌های مختلف تخمین مدل زبانی پرس‌وجو را بر اساس پیکره تطبیقی فارسی-انگلیسی ساخته شده بررسی می‌کنیم. طبیعاً هر چه کیفیت پیکره‌های تطبیقی بیشتر باشد، روابط دقیق‌تری برای کلمات بین زبان‌ها می‌توان بدست آورد و در نتیجه کارآیی سیستم‌های بازیابی اطلاعات بهتر خواهد بود. در آزمایش‌هایمان ابتدا دانش ترجمه را از پیکره تطبیقی استخراج کرده و سپس کیفیت ترجمه‌های بدست آمده را با انجام بازیابی اطلاعات بین زبانی می‌سنجیم. آزمایش‌هایمان را با مدل‌های بازیابی متفاوتی از جمله Okapi با شبه بازخورد ارتباطی[۲] [۷۰] و KL-divergence [۴۰] انجام داده‌ایم، اما از آنجایی که نتایج مدل بازیابی KL-divergence نسبتاً بهتر بود، تنها نتایج بدست آمده از آن را گزارش می‌کنیم.

---

[۱] Mean Average Precision (MAP)

[۲] pseudo relevance feedback

در آزمایش‌های انجام شده، از ابزار Lemur به عنوان سیستم بازیابی، از ریشه‌یاب پُرتر برای ریشه‌یابی کلمات انگلیسی و از لیست ایست واژه انگلیسی Inquery (۴۱۸ کلمه) استفاده کرده‌ایم. مجموعه فارسی تنها بر اساس حذف لیست ایست واژگان دانشگاه نوشاتل[۱] (۳۳۲ کلمه)[۲] پیش پردازش شده است. اگر چه بر روی مجموعه زبان فارسی چندین ریشه‌یاب را اعمال کرده‌ایم ولی هیچ کدام بهبود قابل توجهی نداشتند. در نتیجه تصمیم گرفتیم که از ریشه‌یابی برای زبان فارسی استفاده نکنیم.

همچنین، در آزمایش‌هایمان به عنوان منبع غنی ترجمه، از واژه‌نامه بر خط Farsidic[۳] برای ترجمه کلمات فارسی به انگلیسی و انگلیسی به فارسی، استفاده کرده‌ایم. لازم به ذکر است که در تمامی اجراهای بازیابی اطلاعات انجام شده در این پژوهش، تنها از بخش عنوان پرس‌وجوها استفاده می‌کنیم.

۱.۳.۵ نتایج بازیابی اطلاعات تک زبانی

به عنوان یکی از معیارهای ارزیابی کیفیت روش‌های استخراج ترجمه، از نتایج بازیابی تک زبانی فارسی-فارسی پیکره همشهری استفاده می‌کنیم. در این مجموعه آزمایش‌ها، دو نوع اجرای تک زبانی انجام شده است، یکی با استفاده از مدل بازیابی KL-divergence بدون گسترش پرس‌وجو و دیگری گسترش پرس‌وجو توسط شبه بازخورد ارتباطی. برای اجرای شبه بازخورد ارتباطی، از روش مدل ترکیبی[۴] پیاده سازی شده در ابزار Lemur با استفاده از ۱۰ سند با بالاترین رتبه برای انجام بازخورد استفاده کرده‌ایم. به عنوان پارامترهای این مدل، از ۱۰۰ کلمه برای گسترش مدل پرس‌وجو، ۰.۸ به عنوان ضریب بازخورد و از مدل بازیابی KL-divergence استفاده شده است. جدول ۲.۵، میانگین متوسط دقت، دقت در ۵ سند و دقت در ۱۰ سند را برای نتایج اجراهای تک زبانی نشان می‌دهد. همان طور که در جدول دیده می‌شود اجرای همراه با گسترش پرس‌وجو، میانگین متوسط دقتِ بازیابی تک‌زبانی را نسبت به اجرای بدون گسترش پرس‌وجو، ۳.۵% بهبود داده است.

---

[۱] Neuchatel

[۲] www.unine.ch/info/clef/

[۳] www.farsidic.com

[۴] mixture model

جدول ۲.۵: کارآیی بازیابی تک زبانی فارسی-فارسی

اجرا	MAP	Prec@۵	Prec@۱۰
بدون گسترش پرس‌وجو	۰.۴۰۹	۰.۶۱۶	۰.۵۹
گسترش پرس‌وجو	۰.۴۲۴	۰.۶۱۸	۰.۶۰۷

جدول ۳.۵: نمونه‌ای از ترجمه‌های استخراج شده انگلیسی-فارسی با بیشترین امتیاز شباهت

کلمه انگلیسی	کلمه فارسی	امتیاز
iraqi	عراق	۱۰۴.۹۱
korea	کره	۹۳.۷۳
elect	انتخابات	۸۹.۷۳
nuclear	هسته	۸۵.۰۲
champion	مدال	۸۲.۶۳
weapon	عراق	۸۲.۱۵
cancer	سرطان	۸۰.۷۳
korea	شمالی	۷۸.۸۶
market	بازار	۷۸.۴۱
price	قیمت	۷۸.۰۷
economi	اقتصاد	۷۷.۷۷
econom	اقتصادی	۷۴.۸۰
oil	نفت	۷۴.۲۴
attack	حمله	۷۳.۰۲

۲.۳.۵ COCOT با تعداد همسایه ثابت

در اولین سری از آزمایش‌های انجام شده، سعی کرده‌ایم ترجمه کلمات را از پیکره تطبیقی توسط روش پایه استخراج دانش ترجمه COCOT که در بخش ۱.۱.۴ توضیح داده‌ایم، استخراج کنیم. جدول ۳.۵ نمونه‌ای از بهترین نتایج وابستگی کلمات انگلیسی به فارسی را نشان می‌دهد که از پیکره تطبیقی توسط این روش استخراج شده است. همان طور که از جدول دیده می‌شود، بیشتر کلمات مرتبط با هم کیفیت بالا و به نسبت امتیاز بالایی هم دارند زیرا که امتیازهای استخراج شده از این پیکره‌ی تطبیقی در بازه‌ی (۱۰۵ ، ۲) قرار دارند. لازم به یادآوری است که برای کلمات انگلیسی از ریشه‌یابی استفاده کرده‌ایم به همین دلیل ممکن است بعضی پسوندهای کلمات انگلیسی حذف شده باشد.

در جدول ۴.۵ نمونه‌ای دیگر از وابستگی‌های استخراج شده آورده شده است. این جدول بهترین کلمات فارسی استخراج شده برای چهار کلمه انگلیسی را به همراه امتیاز ارتباطشان نشان می‌دهد. همان طور که جدول نشان می‌دهد،

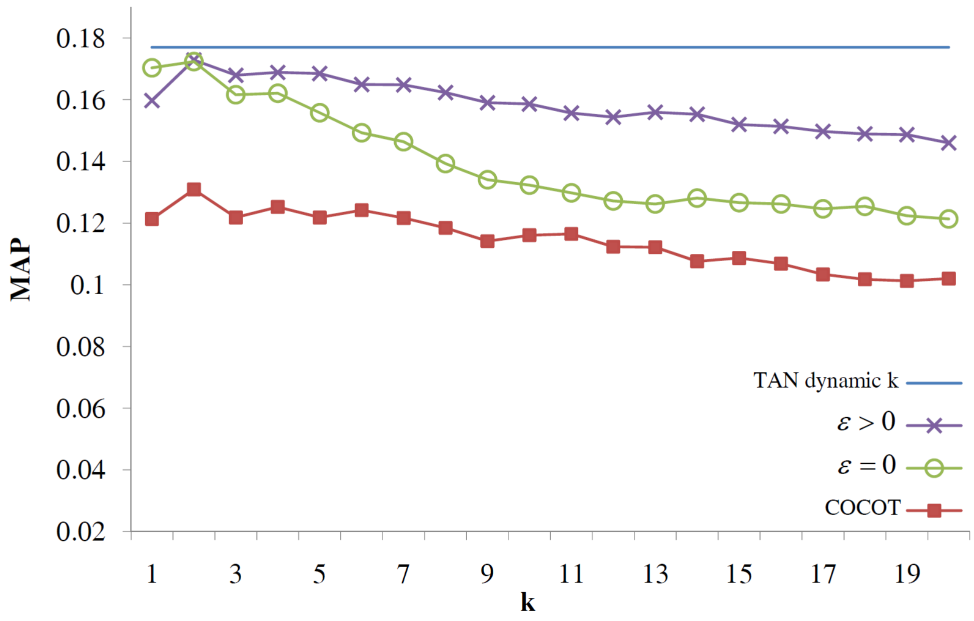
جدول ۴.۵: ترجمه‌های فارسی استخراج شده از پیکره تطبیقی برای چهار کلمه انگلیسی

کلمه انگلیسی	کلمه فارسی	امتیاز
korea	کره	۹۳.۷۳
	شمالی	۷۸.۸۶
	پیونگ	۷۱.۷۷
	ینگ	۷۱.۴۰
	جنوبی	۶۱.۳۵
cancer	سرطان	۸۰.۷۳
	بیماری	۵۲.۰۱
	بدن	۵۱.۲۶
	سلول	۴۱.۶۷
	مبتلا	۳۹.۹۷
iraqi	عراق	۱۰۴.۹۱
	صدام	۹۵.۰۵
	عراقی	۸۲.۹۷
	بغداد	۸۲.۷۸
	حسین	۷۵.۲۹
champion	مدال	۸۲.۶۳
	المپیک	۸۲.۲۰
	قهرمان	۷۳.۵۰
	قهرمانی	۷۲.۲۶
	مسابقات	۷۲.۱۷

امتیاز و یا به عبارتی اطمینان درستی ترجمه کلمات فارسی به کلمه انگلیسی با پایین رفتن در لیست کم می‌شود ولی کماکان کلمات فارسی استخراج شده به کلمه انگلیسی مربوط هستند. این مشاهدات نشان می‌دهند که نتایج بدست آمده می‌توانند برای گسترش پرس‌وجو[1] استفاده شوند.

در گام بعدی از ارتباط‌های استخراج شده بین کلمات در دو زبان استفاده کرده و بازیابی اطلاعات بین زبانی را بر اساس ترجمه‌ی پرس‌وجوهای همشهری انجام می‌دهیم. شکل ۱.۵ کارآیی بازیابی اطلاعات بین زبانی را با استفاده از نتایج استخراج ترجمه توسط روش پایه COCOT نشان می‌دهد. محور افقی این نمودار تعداد ترجمه‌های انتخاب شده

[1] query expansion



شکل ۱.۵: میانگین متوسط دقت برای روش‌های مختلف مبنی بر پیکره تطبیقی

(از ۱ تا ۲۰) برای هر کلمه پرس‌وجو را نشان می‌دهد. از این تعدادِ ثابتِ ترجمه‌های انتخاب شده، برای ساخت مدل زبانی پرس‌وجو استفاده شده است. به منظور ساخت مدل زبانی پرس‌وجو از روش تعداد همسایه‌های ثابت ارائه شده در بخش ۱.۲.۴ استفاده کرده‌ایم. کارآیی اجراهای این روش برای تعداد همسایه‌های متفاوت در جدول ۵.۵ آمده است. همان طور که شکل و جدول نشان می‌دهند، کارآیی COCOT در بهترین حالت، هنگام استفاده از دو کلمه‌ای که بیشترین ارتباط را دارند (top-2)، برابر ۰.۱۳۰ با معیار میانگین متوسط دقت، ۰.۲۲۱ با معیار دقت در ۵ سند و ۰.۲۰۵ با معیار دقت در ۱۰ سند است. این نتایج بدین معنا است که با استفاده از تنها ترجمه‌های استخراج شده از پیکره تطبیقی به عنوان منبع ترجمه پرس‌وجوها توانسته‌ایم به ۳۲٪ میانگین متوسط دقت، ۳۶.۲٪ دقت در ۵ سند و ۳۴.۸٪ دقت در ۱۰ سند در مقایسه با خط بنیادی بازیابی تک زبانی دست پیدا کنیم.

۳.۳.۵ شبکه وابستگی اصطلاحات با تعداد همسایه ثابت

در آزمایش‌های بعدی که انجام داده‌ایم، از روش شبکه وابستگی اصطلاحات (TAN) (ارائه شده در بخش ۲.۱.۴) برای بهبود ترجمه در بازیابی اطلاعات بین زبانی استفاده کرده‌ایم. برای پیاده‌سازی این ایده، ابتدا شبکه‌ای از تمام کلمات انگلیسی و فارسی می‌سازیم. در این شبکه پیوندهای بین کلمات در یک زبان اطلاعات متقابل آنها است و پیوندهای بین کلمات بین دو زبان امتیاز محاسبه شده بین آنها توسط روش COCOT می‌باشد. لازم به ذکر است به دلیل یکسان‌سازی

جدول ۵.۵: کارآیی روش پایه COCOT

روش	MAP	%Mono	Prec@۵	%Mono	Prec@۱۰	%Mono
تک زبانی	۰.۴۰۹		۰.۶۱۶		۰.۵۹	
top-۱	۰.۱۲	۲۹.۳۳	۰.۱۹۵	۳۱.۶۵	۰.۱۸۳	۳۱.۰۱
top-۲	۰.۱۳	۳۱.۷۸	۰.۲۲۱	۳۵.۸۷	۰.۲۰۵	۳۴.۷۴
top-۳	۰.۱۲۱	۲۹.۳۳	۰.۱۸۵	۳۰.۰۳	۰.۱۷۸	۳۰.۱۶
top-۴	۰.۱۲۵	۳۰.۵۶	۰.۱۹	۳۰.۸	۰.۱۸۹	۳۲
top-۶	۰.۱۲۴	۳۰.۳۱	۰.۱۷۶	۲۸.۵۷	۰.۱۸۵	۳۱.۳۵
top-۸	۰.۱۱۸	۲۸.۸۵	۰.۱۹۵	۳۱.۶۵	۰.۱۸۳	۳۱.۰۱
top-۱۰	۰.۱۱۶	۲۸.۳۶	۰.۲	۳۲.۴۶	۰.۱۸	۳۰.۵
top-۱۵	۰.۱۰۸	۲۶.۴	۰.۱۸۱	۲۹.۳۸	۰.۱۶۱	۲۷.۲۸
top-۲۰	۰.۱۰۱	۲۴.۶	۰.۱۷۶	۲۸.۵۷	۰.۱۵۲	۲۵.۷۶

وزن ارتباط کلمات، تمام پیوندها توسط روش ارائه شده در بخش ۲.۱.۴ نرمال‌سازی شده‌اند. به طور شهودی، دو کلمه به احتمال زیاد ترجمه یکدیگر می‌باشند، اگر که همسایگی‌هایشان با یکدیگر بسیار مرتبط و در نتیجه مشابه باشند. به منظور محاسبه شباهت همسایگی‌های دو کلمه، مسیرهای بین آن دو کلمه را که از همسایگی‌هایشان می‌گذرند در نظر می‌گیریم. در آزمایش‌هایمان تمام ۱۰۰ کلمه‌ی نزدیک هر کلمه به عنوان همسایگی‌های آن در نظر گرفته می‌شوند.

به عنوان اولین آزمایش بر اساس شبکه وابستگی اصطلاحات، ما تنها از پیوند مستقیم بین دو کلمه در دو زبان مختلف، بدون در نظر گرفتن همسایگی، استفاده می‌کنیم. نتایج این آزمایش به صورت $\epsilon = 0$ در شکل ۱.۵ نشان داده شده‌اند. لازم به ذکر است که با فرض کلمه F در زبان فارسی و کلمه E در زبان انگلیسی، پیوند مستقیم بین F و E به معنای مسیر به طول یک بین آنها در شبکه می‌باشد و برابرِ ترکیبِ وزن ارتباطی محاسبه شده توسط COCOT از F به E ($w_c(F \rightarrow E)$) و وزن ارتباطی E به F ($w_c(E \rightarrow F)$) است. در آزمایش‌های این پژوهش از جمع مقادیر وزن دو طرف به عنوان روش ترکیب آنها استفاده کرده‌ایم. همچنین، لازم به ذکر است که نتایج مشابهی را نیز با استفاده از ضرب آنها به عنوان تابع ترکیب بدست آورده‌ایم.

در ادامه آزمایش‌ها بر روی شبکه وابستگی اصطلاحات، با در نظر گرفتن تمام مسیرهای بین دو کلمه F و E، شباهت همسایگی‌هایشان را برای محاسبه شباهت خودشان در نظر گرفته‌ایم. به این صورت که هر چه وزن مسیرهای بین آن دو کلمه که از همسایگی‌هایشان می‌گذرد بیشتر باشد، ارتباط کلمات را قوی‌تر در نظر می‌گیریم. همچنین در آزمایش‌های انجام شده از رابطه ۵.۴ استفاده کرده و مقادیر ضریب‌های α، β و γ را برابر در نظر گرفته‌ایم. برای نمایش نتایج این

جدول ۶.۵: کارآیی مدل شبکه TAN بدون در نظر گرفتن همسایگی

روش	MAP	%Mono	Prec@۵	%Mono	Prec@۱۰	%Mono
تک زبانی	۰.۴۰۹		۰.۶۱۶		۰.۵۹	
top-۱	۰.۱۷	۴۱.۶۴	۰.۲۸۴	۴۶.۲۱	۰.۲۷۵	۴۶.۶۶
top-۲	۰.۱۷۲	۴۲.۱۴	۰.۲۷۲	۴۴.۳	۰.۲۵۱	۴۲.۶۷
top-۳	۰.۱۶۱	۳۹.۵۱	۰.۲۶۱	۴۲.۳۹	۰.۲۵	۴۲.۸۷
top-۴	۰.۱۶۲	۳۹.۶۳	۰.۲۶۱	۴۲.۳۹	۰.۲۵۵	۴۳.۲۷
top-۶	۰.۱۴۹	۳۶.۴۹	۰.۲۴۷	۴۰.۱	۰.۲۲۲	۳۷.۶۸
top-۸	۰.۱۳۹	۳۴.۰۶	۰.۲۲۸	۳۷.۰۵	۰.۲۱۸	۳۷.۰۸
top-۱۰	۰.۱۳۲	۳۲.۳۴	۰.۲۰۹	۳۳.۹۹	۰.۲۰۳	۳۴.۴۹
top-۱۵	۰.۱۲۶	۳۰.۹۵	۰.۲۱۶	۳۵.۱۴	۰.۱۹۸	۳۳.۶۹
top-۲۰	۰.۱۲۱	۲۹.۵۸	۰.۲	۳۲.۴۶	۰.۱۸۴	۳۱.۱۸

آزمایش‌ها از نماد $\epsilon > 0$ برای در نظر گرفتن همسایگی کلمات استفاده کرده‌ایم. در شکل ۱.۵، کارآیی بازیابی اطلاعات بین زبانی توسط ترجمه‌های استخراج شده از روش TAN با توجه به مسیرهای به طول یک، دو و سه بین هر دو کلمه شبکه در دو زبان متفاوت، نشان داده شده است.

نتایج کارآیی میانگین متوسط دقت سه روش COCOT، $\epsilon = 0$ و $\epsilon > 0$ بر مبنای تعداد کلمات ترجمه انتخاب‌شده برای هر کلمه پرس‌وجو در شکل ۱.۵ با یکدیگر قابل مقایسه هستند. همان طور که از شکل مشاهده می‌شود، روش مبنی بر شبکه وابستگی اصطلاحات، بهبود قابل ملاحظه‌ای نسبت به روش پایه COCOT دارد. همچنین کمتر بودن میانگین متوسط دقت $\epsilon = 0$ در مقایسه با $\epsilon > 0$ نشان می‌دهد که اطلاعات همسایگی کلمات می‌تواند به افزایش دقت ترجمه‌های استخراج شده از پیکره تطبیقی کمک کند.

علاوه بر این، نتایج پایدار $\epsilon > 0$ با افزایش تعداد k نشان می‌دهد که روش TAN توانسته به ترجمه‌های درست وزن بیشتری دهد و آنها را در ابتدای لیست ترجمه‌های کاندید آن کلمه قرار دهد. به همین ترتیب ترجمه‌های نادرست وزن کمتری گرفته و در لیست به رتبه‌های پایین‌تری نزول می‌کنند. در نتیجه با افزایش مقدار k یعنی تعداد ترجمه‌های کاندید انتخاب شده برای هر کلمه پرس‌وجو، در روش $\epsilon > 0$ نسبت به دیگر روش‌ها، نتایج بهتری بدست آمده است.

جدول‌های ۶.۵ و ۷.۵ کارآیی اجراهای مختلف روش‌های $\epsilon = 0$ و $\epsilon > 0$ را برای تعداد متفاوت کلمات ترجمه برای هر کلمه پرس‌وجو نشان می‌دهند. لازم به ذکر است که این جدول‌ها دقت اجراها در ۵ سند و ۱۰ سند را علاوه بر میانگین متوسط دقت آورده شده در شکل ۱.۵ نشان می‌دهند.

جدول ۷.۵: کارآیی مدل شبکه TAN با در نظر گرفتن همسایگی

روش	MAP	%Mono	Prec@۵	%Mono	Prec@۱۰	%Mono
تک زبانی	۰.۴۰۹		۰.۶۱۶		۰.۵۹	
top-۱	۰.۱۵۶	۳۸.۲۷	۰.۲۶۳	۴۲.۷۸	۰.۲۵۲	۴۲.۸۷
top-۲	۰.۱۷۲	۴۲.۲۶	۰.۲۷۷	۴۵.۰۷	۰.۲۵۲	۴۲.۸۷
top-۳	۰.۱۶۷	۴۱.۰۵	۰.۲۶۳	۴۲.۷۸	۰.۲۵۷	۴۳.۶۶
top-۴	۰.۱۶۸	۴۱.۲۷	۰.۲۷۷	۴۵.۰۷	۰.۲۵۵	۴۳.۲۷
top-۶	۰.۱۶۴	۴۰.۳۲	۰.۲۶۵	۴۳.۱۶	۰.۲۴۵	۴۱.۶۷
top-۸	۰.۱۶۲	۳۹.۶۸	۰.۲۶۱	۴۲.۳۹	۰.۲۴۵	۴۱.۶۷
top-۱۰	۰.۱۵۸	۳۸.۷۷	۰.۲۵۸	۴۲.۰۱	۰.۲۳۱	۳۹.۲۸
top-۱۵	۰.۱۵۱	۳۷.۱۵	۰.۲۳۰	۳۷.۴۳	۰.۲۱۷	۳۶.۸۸
top-۲۰	۰.۱۴۵	۳۵.۶۹	۰.۲۲۳	۳۶.۲۸	۰.۲۰۵	۳۴.۸۹

در ادامه پژوهش انجام شده، روش $\epsilon > 0$ به عنوان روش مبنی بر شبکه وابستگی اصطلاحات TAN در نظر گرفته شده است.

۴.۳.۵ شبکه وابستگی اصطلاحات با تعداد همسایه پویا

در این قسمت از آزمایش‌ها، از روش پیشنهاد شده در بخش ۲.۲.۴ برای محاسبه تعداد پویای ترجمه‌های هر کلمه پرس‌وجو استفاده می‌کنیم. ایده کلی این روش بر مبنای این مشاهده است که تعداد ترجمه‌های استخراج شده مرتبط به هر کلمه متفاوت است و با در نظر گرفتن توزیع وزن ترجمه‌های آن می‌توان مرز این انتخاب را به صورت پویا بدست آورد. به عبارت دیگر به جای اینکه برای تمام کلمات پرس‌وجوها یک تعداد ثابت از ترجمه‌هایشان را برای گسترش پرس‌وجو انتخاب کنیم، ترجمه‌های استخراج شده برای هر یک از کلمات پرس‌وجو را به طور مجزا مورد بررسی قرار داده و تعداد ترجمه‌ها را به طور متغیر بر اساس همسایگی‌هایش انتخاب نماییم.

نتایج انتخاب تعداد همسایه پویا در شکل ۱.۵ با نماد TAN dynamic k و به صورت یک خط افقی نشان داده شده است. همان گونه که مشاهده می‌شود میانگین متوسط دقت این روش در بالای تمام روش‌ها قرار دارد. از آنجایی که مقدار k برای هر کلمه متفاوت است، تنها یک مقدار مشخص برای کارآیی این روش محاسبه می‌شود و به همین دلیل در شکل آن مقدار را تنها به صورت یک خط افقی نشان داده‌ایم. نتیجه حاصل نشان می‌دهد که انتخاب تعداد ترجمه‌های پویا برای هر کلمه توانسته است دقت بازیابی اطلاعات بین زبانی را افزایش دهد.

جدول ۸.۵: کارآیی روش‌های مختلف ترجمه پرس‌وجو با استفاده از پیکره تطبیقی

روش	MAP	%Mono	Impr.	Prec@۵	%Mono	Impr.	Prec@۱۰	%Mono	Impr.
تک‌زبانی	۰.۴۰۹			۰.۶۱۶			۰.۵۹		
COCOT	۰.۱۳	۳۱.۷۸		۰.۲۲۳	۳۶.۲۰		۰.۲۰۵	۳۴.۷۴	
TAN تعداد ثابت k	۰.۱۷۲	۴۲.۰۵	۳۲.۳%	۰.۲۷۷	۴۴.۹۶	۲۴.۲%	۰.۲۵۲	۴۲.۷۱	۲۲.۹%
TAN تعداد پویا k	**۰.۱۷۷**	۴۳.۲۷	۳۶.۱%	**۰.۲۸**	۴۵.۴۵	۲۵.۵%	**۰.۲۶۴**	۴۴.۷۴	۲۸.۷%

جدول ۸.۵ کارآیی روش انتخاب تعداد همسایه پویا را نسبت به بازیابی تک زبانی، روش پایه COCOT و بهترین نتیجه حاصل از روش TAN با تعداد k ثابت ($top-2$) نشان می‌دهد. در این جدول کارآیی این روش‌های را نسبت به روش تک‌زبانی و در صد بهبودی که نسبت به روش پایه COCOT کسب کرده‌اند سنجیده‌ایم. همان طور که مشاهده می‌شود، انتخاب تعداد همسایه پویا بهترین نتایج را از بین روش‌های بازیابی اطلاعات بین زبانی کسب کرده است. با استفاده از این روش در مقایسه با روش COCOT توانسته‌ایم میانگین متوسط دقت را ۳۶.۱% بهبود دهیم.

در ادامه این پژوهش، از ترجمه‌های استخراج شده توسط روش شبکه وابستگی اصطلاحات TAN با استفاده از انتخاب تعداد همسایه پویا به عنوان روش استخراج ترجمه از پیکره تطبیقی برای ترجمه پرس‌وجو استفاده خواهیم کرد.

۵.۳.۵ بررسی صحّت ترجمه

بررسی نتایج کارآیی روش شبکه وابستگی اصطلاحات (TAN) نشان می‌دهد که کارآیی بعضی پرس‌وجوها نسبت به روش پایه COCOT بهبود نیافته‌اند و به عبارتی ترجمه بهتری برای آنها بدست نیامده است که بتواند بازیابی اطلاعات بین زبانی را بهبود بخشد. بنابراین، به منظور تشخیص کلماتی که وابستگی‌های مناسبی برایشان استخراج نشده است، از روشی برای کشف برون‌هشته‌ها به منظور دسته‌بندی نسبی کلمات بر اساس صحّت ترجمه‌های استخراج شده استفاده می‌کنیم.

با استفاده از روش ارائه شده در ۳.۱.۴، می‌توانیم به لیستی مرتب شده از کلمات پرس‌وجو بر اساس مقدار برون‌هشته بودنشان دست یابیم. همچنین در آزمایش‌های انجام شده دو شعاع r و α را به صورت پویا بر اساس همسایگی‌های هر کلمه انتخاب کرده‌ایم. این نوع روش انتخاب به طور کامل در بخش ۲.۲.۴ توضیح داده شده است. نتایج آزمایش‌ها نشان داده‌اند که حتی با حذف این کلمات برون‌هشته از پرس‌وجو که در بالای لیست قرار دارند و ترجمه درستی برایشان پیدا نشده، کارآیی بازیابی اطلاعات بین زبانی افزایش می‌یابد. شکل ۲.۵ میانگین متوسط دقت بازیابی اطلاعات بین

زبانی را برای روش حذف برون‌هشته‌ها نشان می‌دهد. همچنین در روشی دیگر نشان داده‌ایم که با جایگزینی ترجمه کلمات برون‌هشته با معادل درست آنها از منبعی غنی‌تر همانند واژه‌نامه، می‌توان نتایج بازیابی اطلاعات بین زبانی را تا حد زیادی بهبود داد. مقایسه این دو روش استفاده از برون‌هشته‌ها در شکل ۲.۵ نشان داده شده است.

همان طور که در شکل مشاهده می‌شود، افزایش شیب نمودار در نتایج آزمایش حذف برون‌هشته‌ها نشان می‌دهد که حتی با حذف کلمات پرس‌وجو می‌توان به کارآیی بهتری دست یافت. به عبارت دیگر، با حذف ترجمه‌های استخراج شده‌ی نادرست از پرس‌وجوهای به زبان مقصد، اسناد مرتبط‌تری بازیابی می‌شوند. این نتایج برای تشخیص صحت ترجمه‌ها بسیار رضایت‌بخش هستند، اگرچه رتبه‌بندی کلمات برون‌هشته تشخیص داده شده نیاز به بررسی بیشتر دارد. زیرا که با حذف کلمه برون‌هشته دهُم میانگین متوسط دقت افزایش یافته است و مطلوب این است که این افزایش با حذف کلمات جلوتر اتفاق افتاد.

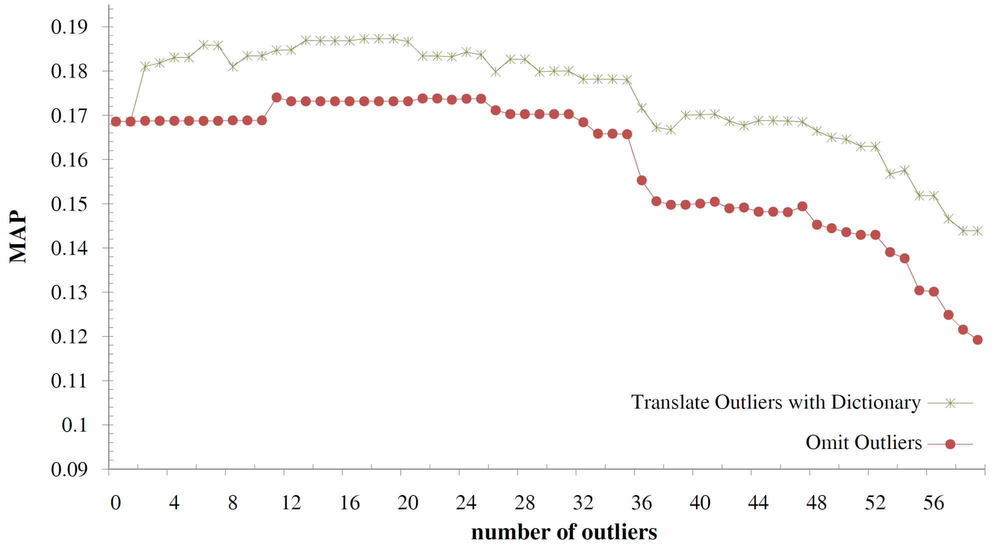
در دیگر مجموعه آزمایش‌ها، تنها کلمات پرس‌وجویی که برون‌هشته تشخیص داده شده‌اند را با ترجمه‌شان از منبع قابل اطمینان همانند واژه‌نامه جایگزین می‌کنیم و بقیه کلمات پرس‌وجو را توسط ترجمه‌های استخراج شده از پیکره تطبیقی، ترجمه می‌کنیم. همان گونه که در شکل ۲.۵ مشاهده می‌شود، تنها با ترجمه تعداد محدودی از کلمات پرس‌وجو (۵ کلمه) توسط واژه‌نامه، توانسته‌ایم به بالاترین حد میانگین متوسط دقت (۰.۱۸۷) دست یابیم که برابر ۴۵.۸% روش پایه تک‌زبانی است. نتایج این قسمت از آزمایش‌ها بسیار جالب توجه هستند زیرا نشان می‌دهند برون‌هشته‌های خوبی تشخیص داده شده‌اند و تنها با ترجمه تعداد کمی از آنها توسط واژه‌نامه و یا حذف آنها از کلمات پرس‌وجو، دقت بازیابی اطلاعات بین زبانی افزایش یافته است.

۶.۳.۵ ترجمه عبارات

در گام بعدی پژوهش سعی داریم که ترجمه‌های عبارات را از پیکره تطبیقی استخراج کرده و پرس‌وجوهایی که شامل عبارات هستند را توسط این روش ترجمه کنیم. به طور کلی همان طور که در بخش ۴.۱.۴ آورده شده است، دو نگرش اصلی را در این زمینه اتخاذ کرده‌ایم.

در **نگرش اول**، تمام اسناد پیکره تطبیقی را پیش پردازش کرده‌ایم، به گونه‌ای که هر دو کلمه کنار هم با هم ترکیب شده‌اند و تشکیل یک بایگرام[۱] را داده‌اند. بدین صورت توانسته‌ایم با اعمال روش‌های استخراج ترجمه از پیکره تطبیقی ترجمه‌ی بایگرام‌ها را از پیکره تطبیقی استخراج کنیم. در این نگرش با عبارات همانند کلمات رفتار کرده‌ایم. نمونه‌ای

[۱] bigram



شکل ۲.۵: میانگین متوسط دقت با استفاده از کشف کلمات برون‌هشته

از ترجمه‌های استخراج شده در جدول ۹.۵ آورده شده است. به منظور استفاده از این ترجمه‌ها برای بازیابی اطلاعات بین زبانی، ابتدا عبارت‌های درون پرس‌وجوها را توسط ترجمه‌های استخراج شده برای عبارات بدست آورده و سپس اگر کلمه‌ای از پرس‌وجوها ترجمه نشده بود (بدین معنا که کلمه‌ای تک بود و در عبارتی قرار نمی‌گرفت) از ترجمه‌های استخراج شده توسط روش شبکه وابستگی اصطلاحات برای کلمات استفاده می‌کنیم. جدول ۱۱.۵ کارآیی این روش را نشان می‌دهد.

در **نگرش دوم**، ابتدا کل مجموعه اسناد را پیش پردازش کرده و هر دو کلمه‌ای که احتمال وقوعشان با یکدیگر زیاد است را به عنوان عبارت در نظر گرفته‌ایم، همچنین در کنار آن عبارت هر یک از کلمات مجزای عبارت را نیز قرار داده‌ایم. یعنی اگر عبارتی به صورت "رئیس جمهور" در سندی تشخیص داده شود، هم کلمه "رئیس" و هم کلمه "جمهور" و هم ترکیب آنها یعنی "رئیس‌جمهور" (به صورت چسبیده) را در کنار یکدیگر در نظر گرفته‌ایم. در آزمایش‌های انجام شده از اطلاعات متقابل کلمات به منظور تشخیص عبارات استفاده شده است و به طور شهودی حدآستانه‌ای بر روی امتیاز اطلاعات متقابل محاسبه شده در نظر گفته‌ایم. این حد آستانه‌ها برای دو مجموعه فارسی و انگلیسی متفاوت است که برای مجموعه فارسی 5×10^5 و برای مجموعه انگلیسی 8×10^5 در نظر گرفته شده است. در کل حدود ۱۰۰۰ عبارت در زبان فارسی و ۷۵۰ عبارت در زبان انگلیسی تشخیص داده شده است.

پس از پیش پردازش اسناد در زبان‌های مبدأ و مقصد و تشخیص عباراتشان به روش ذکر شده، روش شبکه وابستگی اصطلاحات را بر روی آنها به منظور استخراج ترجمه کلمات و ترجمه عبارات اعمال کرده‌ایم. جدول ۱۰.۵ نمونه‌ای از ترجمه‌های استخراج شده توسط این نگرش را نشان می‌دهد. همان طور که از جدول دیده می‌شود ترجمه اصطلاحات

جدول ۹.۵: نمونه‌هایی از ترجمه عبارات استخراج شده از پیکره تطبیقی توسط نگرش اول

عبارت انگلیسی	عبارت فارسی	امتیاز
stock market	بازار سهام	۲۱.۴۶
	سرمایه گذاران	۱۹.۲۷
	ارزش سهام	۱۷.۳۲
	تولید ناخالص	۱۷.۱۳
	ناخالص داخلی	۱۷.۱۲
	قیمت سهام	۱۶.۴۶
	رشد اقتصادی	۱۵.۹۲
south korea	کره جنوبی	۵۵.۷۷
	کره شمالی	۴۶.۴۳
	پیونگ یانگ	۴۶.۲۸
	جزیره کره	۳۵.۲۸
	شبه جزیره	۳۱.۷۴
plane crash	سقوط هواپیمای	۲۶.۶۴
	سانحه هوایی	۲۱.۲۶
	هواپیمای مسافربری	۱۵.۵۱
	برج مراقبت	۱۲.۷۵
	خبر سقوط	۱۱.۵۳

(کلمات و یا عبارات) می‌توانند به صورت کلمه یا عبارت ظاهر شوند. لازم به ذکر است که کلمات انگلیسی در آزمایش‌هایی که انجام داده‌ایم ریشه‌یابی شده‌اند در نتیجه ممکن است پسوندهای آنها تغییر کرده باشد.

از آنجایی که ترجمه اصطلاحات بدست آمده توسط نگرش دوم، حالات مختلفی می‌توانند داشته باشند، از آنها می‌توان به عنوان منبعی متفاوت برای ترجمه کلمات پرس‌وجو استفاده کرد. نتایج کارآیی آزمایش‌های بازیابی اطلاعات بین زبانی توسط این ترجمه‌های استخراج شده در جدول ۱۱.۵ آورده شده است. همان‌طور که جدول نشان می‌دهد میانگین متوسط دقت ۰.۱۶۲ شده است که نسبت به نگرش اول کاهش یافته است، این در حالی است که دقت در ۵ سند و دقت در ۱۰ سند افزایش داشته‌اند. نتایج بدست آمده ارزشمند هستند ولی انتظار می‌رفت که بتوان نتایج بهتری را با استفاده از ترجمه‌های استخراج شده توسط نگرش دوم بدست آورد. از آنجایی که در این نگرش حدودِ آستانه‌ای برای تشخیص عبارات در نظر گرفته‌ایم، شاید اگر بتوان این حدودِ آستانه را تنظیم کرد و عبارات بهتری را تشخیص داد، به عبارتی بتوان عبارات و کلمات را بهتر ترکیب کرد، کارآیی بازیابی اطلاعات بین زبانی نیز بهبود یابد. در آینده در نظر

جدول ۱۰.۵: نمونه‌هایی از ترجمه عبارات استخراج شده از پیکره تطبیقی توسط نگرش دوم

اصطلاح انگلیسی	اصطلاح فارسی	امتیاز
mobil	تلفن همراه	۱۶.۷۶
	تلفن	۱۵.۷۷
earthquak	زلزله	۲۹.۷۵
	زمین لرزه	۲۸.۰۵
	ریشتر	۲۶.۰۱
	لرزه	۲۴.۷۹
	زلزله زدگان	۲۲
cancer drug	شیمی درمانی	۲۴.۷۶
	سرطانی	۲۳.۳۲
presidenti elect	انتخابات ریاست	۲۵.۱۹
	ریاست جمهوری	۲۱.۷۸

جدول ۱۱.۵: کارآیی ترجمه عبارات

روش	MAP	%Mono	Impr.	Prec@۵	%Mono	Impr.	Prec@۱۰	%Mono	Impr.
تک‌زبانی	۰.۴۰۹			۰.۶۱۶			۰.۵۹		
COCOT	۰.۱۳	۳۱.۷۸		۰.۲۲۳	۳۶.۲۰		۰.۲۰۵	۳۴.۷۴	
نگرش اول	۰.۱۷۳	۴۲.۲۹	۲۱.۴%	۰.۲۵۸	۴۱.۸۸	۱۵.۴۳%	۰.۲۳۶	۴۰	۱۳.۳%
نگرش دوم	۰.۱۶۲	۳۹.۶	۱۳.۶%	۰.۲۸۴	۴۶.۱	۲۷%	۰.۲۴۷	۴۱.۸۶	۱۸.۶%

داریم بر روی بهبود و تنظیم پارامترهای این روش تمرکز کنیم.

۷.۳.۵ ترکیب با واژه‌نامه

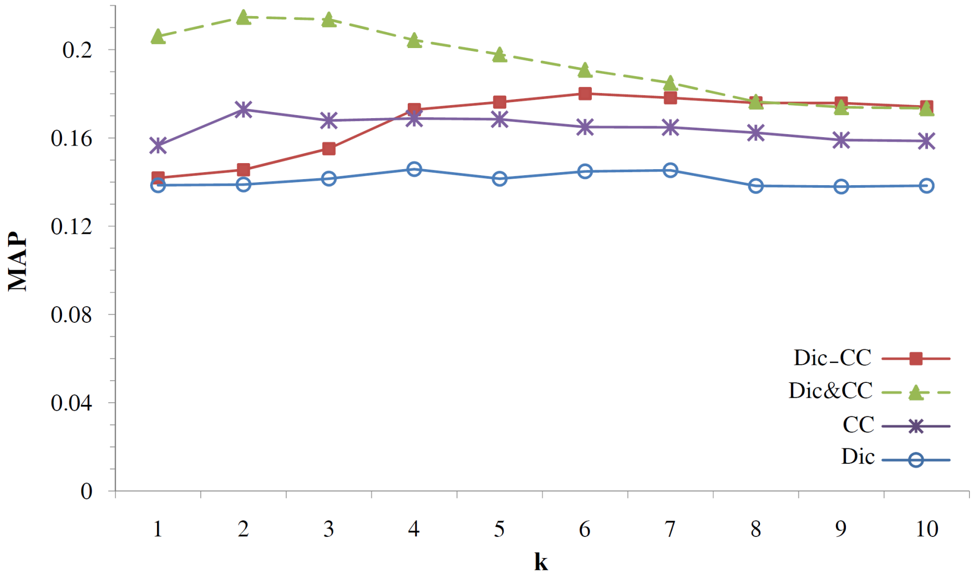
در ادامه پژوهش، از ترجمه‌های استخراج شده توسط پیکره تطبیقی برای بهبود بازیابی اطلاعات بین زبانی مبنی بر واژه‌نامه استفاده می‌کنیم. به دو روش کلی پیکره تطبیقی را با واژه‌نامه ترکیب می‌کنیم. در اولین روش تمام کلمات پرس‌وجوها ابتدا توسط واژه‌نامه ترجمه می‌شوند. سپس کلماتی که در واژه‌نامه یافت نشده‌اند (عموماً اسامی خاص) با استفاده از پیکره تطبیقی ترجمه می‌شوند. در آزمایش‌های انجام شده این روش با نماد Dic-CC نشان داده شده‌اند. در روش دوم، تمام کلمات پرس‌وجو هم توسط واژه‌نامه و هم توسط پیکره تطبیقی ترجمه می‌شود. این روش با نماد Dic&CC نشان

داده شده است. شکل ۳.۵ میانگین متوسط دقت بازیابی اطلاعات بین زبانی برای روش‌های مبنی بر واژه‌نامه بر اساس تعداد متفاوت ترجمه‌های کلمات (۱ تا ۱۰) برای هر کلمه پرس‌وجو را نشان می‌دهد. در این شکل محور افقی بر تعداد ترجمه‌های استفاده شده برای هر کلمه در واژه‌نامه دلالت دارد. لازم به ذکر است که تعداد ترجمه‌های استخراج شده از پیکره تطبیقی توسط روش انتخاب تعداد همسایه پویا محاسبه می‌شوند.

در این شکل نتایج استفاده از تنها واژه‌نامه و تنها پیکره تطبیقی به ترتیب به صورت Dic و CC نشان داده شده‌اند. نکته جالب توجه اینجاست که هنگام استفاده از واژه‌نامه برای بازیابی اطلاعات بین زبانی، عموماً با افزایش تعداد ترجمه‌های استفاده شده یک کلمه، کارآیی کاهش می‌یابد، ولی در آزمایش‌های انجام شده این پژوهش بر روی پیکره همشهری این اتفاق نیافتاده است. بررسی کارآیی تک تک پرس‌وجوها به طور مجزا نشان می‌دهد که با افزایش تعداد ترجمه‌های واژه‌نامه، دقت بعضی از پرس‌وجوها کاهش می‌یابد و به طور عکس در بعضی دیگر از پرس‌وجوها دقت افزایش می‌یابد به همین دلیل در مجموعه میانگین متوسط دقت آنها تقریباً عددی ثابت است.

به طور کلی، کارآیی ضعیف واژه‌نامه تا حدودی قابل پیش‌بینی بود، زیرا واژه‌نامه‌ها فاقد اسامی خاص هستند، در حالی که در بیشتر پرس‌وجوها اسامی خاص کلمات اصلی می‌باشند و عدم امکان ترجمه‌ی آنها آسیب جدی به کارآیی پرس‌وجو وارد خواهد کرد. بنابراین یکی از دلایل موفقیت پیکره تطبیقی قابلیت ترجمه این کلمات خارج از واژه‌نامه می‌باشد. از جمله مزایای دیگر پیکره تطبیقی پیدا کردن کلمات مرتبط مناسب برای گسترش پرس‌وجو است. قسمت‌هایی از نمودار که مقدار میانگین متوسط دقت با افزایش مقدار k، افزایش یافته است، به خوبی این فرضیه را نشان می‌دهند. علاوه بر این، همان گونه که در شکل مشاهده می‌شود، ترکیب واژه‌نامه با پیکره تطبیقی مفید بوده است و روش Dic&CC بهترین نتایج را نشان می‌دهد.

جدول ۱۲.۵ نتایج کارآیی کلی آزمایش‌هایمان را نشان می‌دهد. سه سطر اول جدول که شامل روش‌های پایه بازیابی تک‌زبانی، COCOT و واژه‌نامه تنها است به عنوان خط بنیادی برای مقایسه دیگر روش‌ها آورده شده‌اند. همان طور که جدول نشان می‌دهد، می‌توان تنها با استفاده از پیکره تطبیقی (CC) به نتایج قابل قبولی دست یافت و در مقایسه با بازیابی تک‌زبانی توانسته‌ایم به ۴۳.۲% میانگین متوسط دقت دست یابیم. علاوه بر این، شاید دور از پیش‌بینی نبود که روش‌های Dic-CC و Dic&CC که ترکیبی از دو منبع ترجمه می‌باشند، بتوانند بهترین کارآیی را کسب کنند. به طور کلی این دو روش مجموعه واژگان واژه‌نامه را با اضافه کردن ترجمه‌های استخراج شده از پیکره تطبیقی گسترش می‌دهند. در روش Dic-CC، آزمایش‌هایی برای تعیین تعداد ترجمه‌های استفاده شده توسط واژه‌نامه انجام داده‌ایم. در جدول ۵.۱۲ بهترین نتایج را گزارش داده‌ایم که مبنی بر انتخاب اولین ترجمه از واژه‌نامه می‌باشد. در آزمایش‌های انجام شده در بهترین حالت توانسته‌ایم به ۵۲.۳% میانگین متوسط دقت روش پایه تک‌زبانی برسیم.



شکل ۳.۵: میانگین متوسط دقت برای روش‌های مختلف مبنی بر واژه‌نامه

روش CC-Dic نشان دهنده روش بررسی صحّت ترجمه با استفاده از کشف برون‌هشته است. در این جدول تنها بهترین نتایج ترجمه کلمات برون‌هشته هنگامی که از واژه‌نامه استفاده می‌شود، آورده شده است. همان طور که در شکل ۵.۲ دیده شد، بهترین حالت هنگامی است که تنها پنج کلمه از برون‌هشته‌ها توسط واژه‌نامه ترجمه شوند. این نتایج جالب توجه است زیرا میانگین متوسط دقت ترجمه تنها این پنج کلمه توسط واژه‌نامه و ترجمه بقیه کلمات از پیکره تطبیقی توانسته به ۴۵.۲% کارآیی بازیابی اطلاعات تک‌زبانی برسد که بهبودی قابل توجه نسبت به واژه‌نامه است. به طور کلی می‌توان نتیجه گرفت که استفاده از ترجمه‌های استخراج شده از پیکره تطبیقی، کارآیی واژه‌نامه را برای بازیابی اطلاعات بین زبانی بهبود می‌دهد. در صد بهبودها نسبت به واژه‌نامه در ستون‌های Impr._{Dic} جدول ۱۲.۵ برای مقایسه آسان‌تر کارآیی این روش‌ها قید شده‌اند.

۸.۳.۵ گسترش پرس‌وجو توسط شبه بازخورد ارتباطی

همان طور که در جدول ۱۲.۵ مشاهده می‌شود، در تمام روش‌ها دقت در ۵ و ۱۰ سند بالا از میانگین متوسط دقت بهتر می‌باشند. همچنین هنگامی که دقت در اسناد بالای بازیابی شده نسبتاً زیاد باشد، روش شبه بازخورد ارتباطی[۱] به بهبود میانگین متوسط دقت کمک خواهد کرد. بنابراین در مجموعه‌ای دیگر از آزمایش‌ها، گزارشی از نتایج استفاده از گسترش پرس‌وجو توسط شبه بازخورد ارتباطی ارائه می‌کنیم. جدول ۱۳.۵ نتایج سه روش گسترش پرس‌وجو را نشان می‌دهد که

[۱] pseudo relevance feedback

جدول ۱۲.۵: کارآیی روش‌های مختلف ترجمه پرس‌وجو

روش	MAP	%Mono	Impr.$_{Dic}$	Prec@۵	%Mono	Impr.$_{Dic}$	Prec@۱۰	%Mono	Impr.$_{Dic}$
تک‌زبانی	۰.۴۰۹			۰.۶۱۶			۰.۵۹		
COCOT	۰.۱۳	۳۱.۷		۰.۲۲۳	۳۶.۲		۰.۲۰۵	۳۴.۷	
Dic	۰.۱۴۶	۳۵.۶		۰.۲۳۷	۳۸.۴		۰.۲۲۵	۳۸.۱	
CC	۰.۱۷۷	۴۳.۲	۲۱.۲%	۰.۲۸	۴۵.۴	۱۸.۱%	۰.۲۶۴	۴۴.۷	۱۷.۳%
Dic-CC	۰.۱۸۳	۴۴.۷	۲۵.۳%	۰.۳۱۷	۵۱.۴	۳۳.۷%	۰.۲۸۵	۴۸.۳	۲۶.۶%
Dic&CC	**۰.۲۱۴**	۵۲.۳	۴۶.۵%	**۰.۳۳۱**	۵۳.۷	۳۹.۶%	**۰.۳۱۸**	۵۳.۸	۴۱.۳%
CC-Dic	۰.۱۸۵	۴۵.۲	۲۶.۷%	۰.۳۰۳	۴۹.۲	۲۷.۸%	۰.۲۸۶	۴۸.۴	۲۷.۱%

جدول ۱۳.۵: کارآیی گسترش پرس‌وجو توسط شبه بازخورد ارتباطی

روش	MAP	%Mono	Prec@۵	%Mono	Prec@۱۰	%Mono
تک‌زبانی	۰.۴۲۴		۰.۶۱۸		۰.۶۰۷	
CC	۰.۱۹۴	۴۵.۷	۰.۲۹۸	۴۸.۲	۰.۲۸	۴۶.۱
Dic&CC	**۰.۲۶۲**	۶۱.۸	**۰.۳۶۲**	۵۸.۶	**۰.۳۷**	۶۱

شامل بازیابی تک‌زبانی، انحصاراً استفاده از پیکره تطبیقی CC و بهترین نتیجه بازیابی اطلاعات بین زبانی که تا کنون بدست آمده با استفاده از پیکره تطبیقی Dic&CC می‌باشد. همان طور که جدول نشان می‌دهد، با گسترش پرس‌وجو در مقایسه با روش پایه تک‌زبانی، می‌توانیم به ۴۵.۷% میانگین متوسط دقت، ۴۸.۲% دقت در ۵ سند و ۴۶.۱% دقت در ۱۰ سند با استفاده از پیکره تطبیقی به عنوان منبع ترجمه دست یابیم و با استفاده از ترکیب واژه‌نامه با پیکره تطبیقی به ۶۱.۸% میانگین متوسط دقت، ۵۸.۶% دقت در ۵ سند و ۶۱% دقت در ۱۰ سند نائل شویم. در این روش، با استفاده از شبه بازخورد ارتباطی توانسته‌ایم از گسترش پرس‌وجو در زبان مبدأ سود ببریم و همچنین هنگام استفاده از پیکره تطبیقی به منظور ترجمه پرس‌وجوها، از گسترش پرس‌وجو در زبان مقصد نیز استفاده کرده‌ایم.

۹.۳.۵ استفاده از پیکره INFILE

در آزمایش‌های انجام شده تا کنون از پیکره همشهری و پرس‌جوهای دو زبانه آن برای بازیابی اسناد فارسی توسط پرس‌وجوهای انگلیسی استفاده کرده‌ایم. در این بخش با استفاده از پیکره INFILE، بازیابی اسناد انگلیسی را بر اساس پرس‌وجوهای فارسی بررسی می‌کنیم.

جدول ۱۴.۵: کارآیی بازیابی تک زبانی انگلیسی-انگلیسی

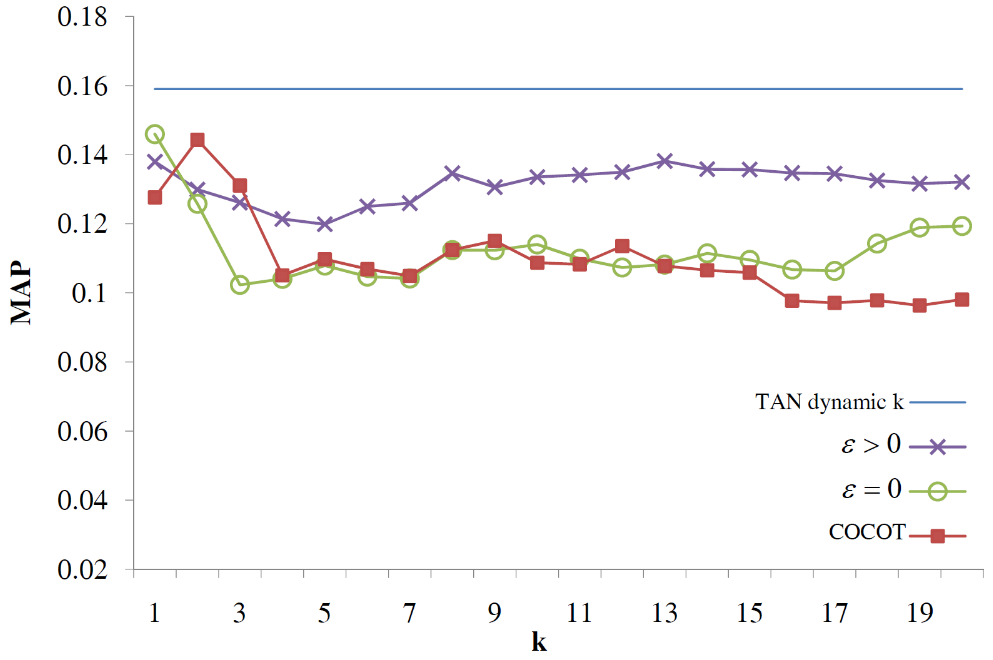
اجرا	MAP	Prec@۵	Prec@۱۰
بدون گسترش پرس‌وجو	۰.۳۶۶	۰.۴۸	۰.۴۴
گسترش پرس‌وجو	۰.۳۵۵	۰.۴۶۲	۰.۴۵۱

نتایج بازیابی اطلاعات تک زبانی

در این بخش نیز همانند آزمایش‌های انجام شده بر روی پیکره همشهری، دو نوع اجرای تک زبانی انجام داده‌ایم، یکی با استفاده از مدل بازیابی KL-divergence بدون گسترش پرس‌وجو و دیگری استفاده از گسترش پرس‌وجو توسط شبه بازخورد ارتباطی. برای اجرای روش شبه بازخورد ارتباطی، از روش مدل ترکیبی پیاده سازی شده در ابزار Lemur توسط ۱۰ سند با بالاترین رتبه برای انجام بازخورد استفاده کرده‌ایم. به عنوان پارامترهای این مدل، از ۱۰۰ کلمه برای گسترش مدل پرس‌وجو و ۰.۸ به عنوان ضریب بازخورد استفاده کرده‌ایم. جدول ۱۴.۵ ، میانگین متوسط دقت، دقت در ۵ سند و دقت در ۱۰ سند را برای نتایج اجراهای تک زبانی نشان می‌دهد. همان طور که از جدول می‌توان دید، گسترش پرس‌وجو توسط شبه بازخورد ارتباطی بر روی پرس‌وجوهای این پیکره با معیارهای میانگین متوسط دقت و دقت در ۵ سند بالا، نتایج رضایت‌بخشی نداشته است، ولی با معیار دقت در ۱۰ سند بالا، این روش خوب عمل کرده است.

نتایج بازیابی اطلاعات بین زبانی توسط پیکره تطبیقی

شکل ۴.۵ کارآیی بازیابی اطلاعات بین زبانی را با استفاده از ترجمه‌های استخراج شده از پیکره تطبیقی بر اساس تعداد ترجمه‌های استفاده شده برای هر کلمه پرس‌وجو در ساخت پرس‌وجوی ترجمه شده (top-1 تا top-20) نشان می‌دهد. همانند نتایج بازیابی اطلاعات بین زبانی به دست آمده از پیکره همشهری، همان طور که انتظار می‌رفت روش مبنی بر شبکه با انتخاب تعداد پویای k (TAN dynamic k) بیشترین میانگین متوسط دقت را دارد. همچنین روش مبنی بر شبکه $\epsilon > 0$ که از اطلاعات همسایگی کلمات نیز استفاده می‌کند با افزایش تعداد k نسبتاً پایدار است و تا حدودی کارآیی را بهبود می‌دهد. تنها تفاوت این نتایج با پیکره همشهری در این است که روش شبکه $\epsilon = 0$ نتوانسته نسبت به روش پایه COCOT کارآیی را بهبود دهد. کارایی پایین $\epsilon = 0$ ممکن است به این دلیل باشد که کلمات فارسی ریشه‌یابی نشده‌اند و از آنجایی که پرس‌وجوهایی که باید ترجمه شوند به زبان فارسی هستند، دقت ترجمه‌ها برای بازیابی اطلاعات بین زبانی پایین است. به عبارت دیگر، کلمات هم‌ریشه‌ای که به شکل‌های مختلف نوشته شده‌اند، متمایز از هم در نظر گرفته شده‌اند و گروه‌بندی آنها با استفاده از روش‌های ریشه‌یابی می‌توانست ترجمه دقیق‌تری به دست دهد.

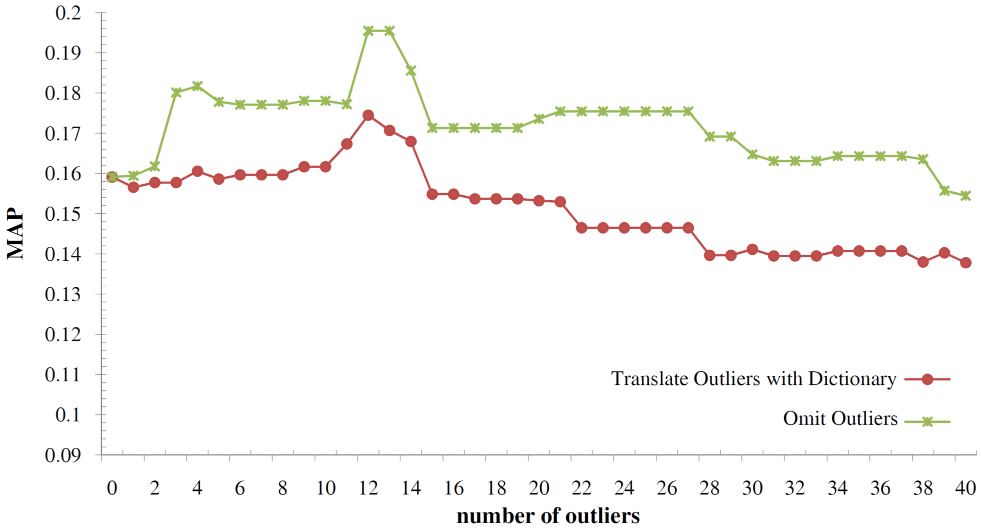


شکل ۴.۵: میانگین متوسط دقت برای روش‌های مختلف مبنی بر پیکره تطبیقی در پیکره INFILE

هنگامی که از اطلاعات همسایگی کلمات استفاده می‌شود این مشکل تا حدودی بر طرف شده و ترجمه‌های استخراج شده دقیق‌تر می‌شوند.

بررسی صحّت ترجمه

در این سری آزمایش‌ها، کلمات برون‌هشته در پیکره INFILE را بر اساس روش ارائه شده در ۳.۱.۴ مشخص کرده و همانند نتایج ارائه شده برای پیکره همشهری، آزمایش‌هایی بر روی آنها را با هدف بهبود کارآیی بازیابی اطلاعات بین زبانی انجام می‌دهیم. شکل ۵.۵، نتایج حذف و جایگزینی کلمات برون‌هشته را بر روی کارآیی بازیابی اطلاعات بین زبانی پیکره INFILE نشان می‌دهد. همان طور که در شکل دیده می‌شود، تنها با حذف یازده کلمه از پرس‌وجوها توانسته‌ایم نتایج بهتری را کسب کنیم. همچنین اگر ترجمه این کلمات را از منبع غنی‌تری همانند واژه‌نامه استخراج کنیم، آنگاه دقت بازیابی به میزان قابل توجهی افزایش یافته و به میانگین متوسط دقت ۰.۱۹۵ دست می‌یابیم که برابر ۴۴٪ بازیابی تک زبانی است. نتایج این آزمایش‌ها بر روی پیکره INFILE تا حدودی رفتاری مشابه پیکره همشهری دارند. در نتیجه می‌توان گفت روشی که در این پژوهش برای بررسی صحت ترجمه‌ها استفاده کرده‌ایم ارزشمند است و نتایج بدست آمده از اعمال آن بر روی دو مجموعه اسناد متفاوت نشان دهنده این حقیقت است که توانسته‌ایم با دقت نسبتاً خوبی کلماتی که ترجمه‌شان از پیکره تطبیقی استخراج نشده را مشخص کنیم.



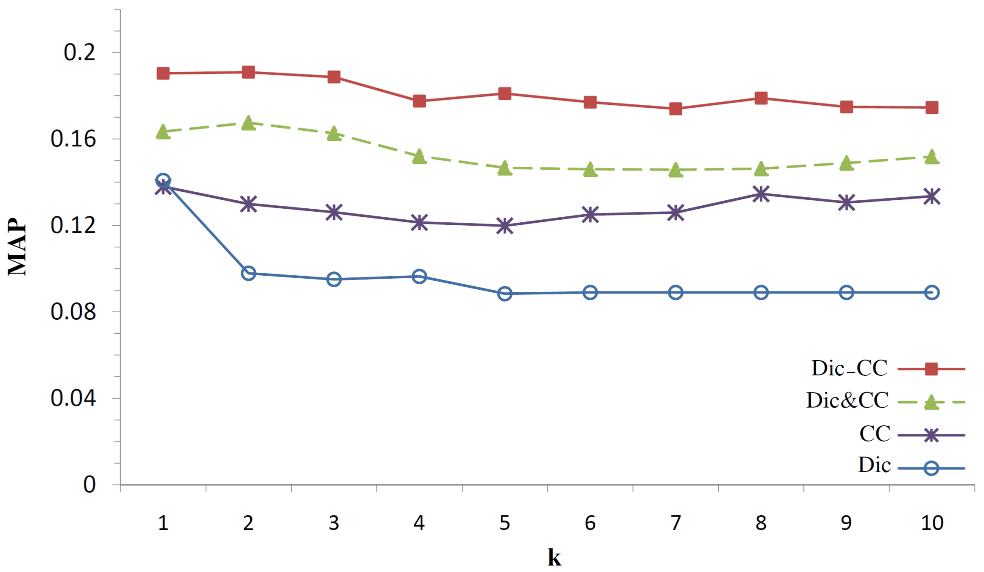
شکل ۵.۵: میانگین متوسط دقت با استفاده از کشف کلمات برون‌هشته در پیکره INFILE

ترکیب با واژه‌نامه

همانند پیکره همشهری، روش‌های مختلف ترکیب واژه‌نامه با پیکره تطبیقی را بر روی پیکره INFILE نیز بررسی کرده‌ایم. شکل ۶.۵ نتایج بدست آمده از این آزمایش‌ها را نشان می‌دهد. در این شکل همان طور که انتظار می‌رفت ترکیب ترجمه‌های استخراج شده از پیکره تطبیقی با واژه‌نامه توانسته تا حد مطلوبی کارآیی را افزایش دهد. همان گونه که در بخش‌های قبلی نیز ذکر شد، استفاده از پیکره‌های تطبیقی برای بازیابی اطلاعات بین زبانی دو مزیت عمده دارد. اول اینکه می‌توان ترجمه کلمات که در واژه‌نامه نیستند همانند اسامی خاص و کلمات جدیدی که به زبان اضافه می‌شوند را توسط آنها بدست آورد. همچنین با استخراج ترجمه از پیکره‌های تطبیقی یکسری کلمات مرتبط نیز بدست می‌آیند که می‌توان از آنها به عنوان گسترش پرس‌وجو استفاده کرد.

در نتایج بدست آمده از پیکره INFILE، روش Dic-CC، که برای ترجمه کلماتی که در واژه‌نامه یافت نشده‌اند از پیکره تطبیعی استفاده می‌کند، بهترین کارآیی را دارد. طبق آمار آزمایش‌های انجام شده، از ۱۷۵ کلمه پرس‌وجو، ۱۰۵ کلمه ترجمه‌شان در واژه‌نامه بوده و تنها ۱۶ کلمه توسط پیکره تطبیقی ترجمه شده‌اند. همچنین لازم به ذکر است برای اینکه بتوانیم کلمات پرس‌وجو را توسط واژه‌نامه ترجمه کنیم، پرس‌وجوهای فارسی و کلمات فارسی واژه‌نامه را ریشه‌یابی کرده‌ایم تا بتوان ترجمه کلمات هم‌ریشه را از واژه‌نامه استخراج کرد.

به طور کلی، نتایج اعمال روش‌های ترجمه متفاوت بر روی پیکره INFILE در جدول ۱۵.۵ آمده است. همان گونه که در جدول مشاهده می‌شود، همانند نتایج کسب شده بر روی پیکره همشهری (جدول ۱۲.۵)، ترکیب پیکره تطبیقی با



شکل ۶.۵: میانگین متوسط دقت برای روش‌های مختلف مبنی بر واژه‌نامه در پیکره INFILE

جدول ۱۵.۵: ارزیابی روش‌های مختلف ترجمه پرس‌وجو در پیکره INFILE

روش	MAP	%Mono	$Impr._{Dic}$	Prec@۵	%Mono	$Impr._{Dic}$	Prec@۱۰	%Mono	$Impr._{Dic}$
تک‌زبانی	۰.۳۶۶			۰.۴۸			۰.۴۴		
COCOT	۰.۱۴۴	۳۹.۳		۰.۲۰۴	۴۲.۵		۰.۱۹۵	۴۴.۳	
Dic	۰.۱۴۴	۳۹.۳		۰.۱۸۶	۳۸.۷		۰.۱۸۴	۴۱.۸	
CC	۰.۱۵۹	۴۳.۴	۱۰.۴۱%	۰.۱۹۵	۴۰.۶	۴.۸۳%	۰.۱۹۳	۴۳.۸	۴.۸۹%
Dic-CC	**۰.۲**	**۵۴.۶**	**۳۸.۸۸%**	۰.۲۴۴	۵۰.۸	۳۱.۱%	**۰.۲۴۸**	**۵۶.۳**	**۳۴.۷%**
Dic&CC	۰.۱۷۲	۴۶.۹	۱۹.۴۴%	۰.۲۱۳	۴۴.۳	۱۴.۵%	۰.۱۹۵	۴۴.۳	۵.۹۷%
CC-Dic	۰.۱۹۵	۵۳.۲	۳۵.۴%	**۰.۲۴۹**	**۵۱.۸**	**۳۳.۸%**	۰.۲۳۸	۵۴	۲۹.۳%

واژه‌نامه، بهترین نتایج را به دست داده است که ۵۴.۶% میانگین متوسط دقت را نسبت به خط بنیادی بازیابی تک‌زبانی دارد و نسبت به روشی که تنها از اطلاعات واژه‌نامه استفاده می‌کند، ۳۸.۸۸% بهبود داشته است. همچنین در روش کشف برون‌هشته‌ها (CC-Dic) تنها استفاده از ترجمه ۱۱ کلمه در واژه‌نامه در مقایسه با خط بنیادی بازیابی تک‌زبانی، ۵۳.۳% از میانگین متوسط دقت را کسب کرده و نسبت به واژه‌نامه ۳۵.۴% بهبود داشته است.

فصل ۶

نتیجه‌گیری

۱.۶ جمع‌بندی

در سیستم‌های بازیابی اطلاعات بین زبانی، زبان پرس‌وجو و اسناد بازیابی شده متفاوت می‌باشند. مسأله‌ای که در این گونه سیستم‌ها مطرح می‌باشد گذر از مرز بین زبان‌ها است. توسط ترجمه پرس‌وجوها و یا اسناد می‌توان بازیابی اطلاعات بین زبانی را تبدیل به بازیابی اطلاعات معمولی کرد. منابع مطرح در این زمینه واژه‌نامه‌ها، ماشین‌های ترجمه، پیکره‌های موازی و تطبیقی می‌باشند. تهیه و ایجاد سه دسته از این منابع یعنی واژه‌نامه‌ها، ماشین‌های ترجمه و پیکره‌های موازی، نسبتاً هزینه‌بر است و ممکن است برای تعداد زیادی از جفت زبان‌های اقلیت در دسترس نباشند. ولی پیکره‌های تطبیقی عموماً بین زبان‌ها وجود دارند و راحت‌تر می‌توان آنها را جمع‌آوری کرد. پیکره‌های تطبیقی اسنادی در دو زبان را شامل می‌شوند که لزوماً ترجمه یکدیگر نیستند و ممکن است تنها تا حدودی با یکدیگر مربوط باشند. بنابراین مزیت اصلی این گونه از پیکره‌ها در دسترس بودن آنها است. در این پایان‌نامه، بازیابی اطلاعات بین زبانی با استفاده از پیکره تطبیقی فارسی-انگلیسی بررسی شده است.

در فصل ۲ ، ویژگی‌ها و کاستی‌های روش‌های مختلف بازیابی اطلاعات بین زبانی بررسی شدند. از این میان پیکره‌های تطبیقی به عنوان منبعی ارزشمند در این زمینه به طور کامل‌تر مورد بررسی قرار گرفتند. دو مسأله اصلی مطرح در بازیابی اطلاعات بین زبانی، چگونگی استخراج دانش ترجمه و استفاده از آن می‌باشد که در ادامه فصل دوم کارهای انجام شده در این زمینه بر روی پیکره‌های تطبیقی بررسی شدند.

در این پژوهش بر روی بازیابی اطلاعات بین زبانی دو زبان فارسی و انگلیسی تمرکز شده است. از آنجایی که تا کنون مطالعاتی بر روی پیکره‌های تطبیقی بین این دو زبان بدین منظور صورت نگرفته است، در گام نخست، پیکره تطبیقی نسبتاً بزرگِ فارسی-انگلیسی ایجاد شده است. این پیکره توسط دو مجموعه اخبار مستقل از هم ساخته و اسناد آنها را بر اساس شباهت موضوعی و تاریخ انتشار هم‌تراز شده‌اند. لازم به ذکر است که مراحل ساخت این پیکره در فصل ۳ آورده شده است. همچنین آزمایش‌های مختلفی برای ساخت پیکره تطبیقی با روش‌های متفاوتی از جمله روش‌های گوناگون ترجمه و مدل‌های بازیابی مختلف، انجام شده است. به منظور ارزیابی پیکره تطبیقی ساخته شده (UTPECC) چندین معیار در نظر گرفته شده‌اند. به عنوان اولین و مهمترین معیار، از یک مقیاس مرتبط بودن پنج سطحی برای ارزیابی کیفیت هم‌ترازی‌ها در یک ماه استفاده شده است. نتایج ارزیابی نشان می‌دهد که با استفاده از ترجمه‌ی کامل کلمات نمایش اسناد و مدل بازیابی Okapi با شبه بازخورد ارتباطی توانسته‌ایم به یک پیکره تطبیقی با کیفیت بالا دست پیدا کنیم که ۹۵% اسناد هم‌ترازی ارزیابی شده کاملاً و یا نسبتاً درباره رویداد مرتبطی هستند.

به منظور انجام بازیابی اطلاعات بین زبانی با استفاده از پیکره‌های تطبیقی نیاز است که دانش ترجمه را از این گونه پیکره‌ها استخراج کرده و پرس‌وجوهای ترجمه شده را ساخت. بدین منظور در فصل ۴، به شرح و چگونگی روش‌های ارائه شده برای استخراج ترجمه کلمات پرداخته‌ایم. به عنوان اصلی‌ترین روش، روشی را مبنی بر شبکه وابستگی اصطلاحات[1] معرفی کردیم که از ارتباط اصطلاحات در داده‌های تک زبانی علاوه بر ارتباط اصطلاحات بین دو زبان استفاده می‌کند. در این شبکه از میزان شباهت همسایگی اصطلاحات برای بهبود ترجمه‌های استخراج شده استفاده می‌شود. همچنین به منظور بررسی صحّت ترجمه‌های استخراج شده روشی را بر مبنای کشف برون‌هشته‌ها استفاده کرده‌ایم. در ادامه، در نگرشی دیگر، کلمات مجاور هم را در اسناد یک زبان به عنوان عبارت در نظر گرفته و روش ارائه شده برای استخراج دانش ترجمه را بر روی آنها اعمال کرده‌ایم. مسأله دیگری که در بازیابی اطلاعات بین زبانی مطرح است، ساخت پرس‌وجوهای ترجمه شده می‌باشد. در این پژوهش از روش مدل زبانی پرس‌وجو استفاده شده است که در فصل ۴، روشی برای چگونگی ساخت مدل زبانی بر اساس امتیازهای استخراج شده ارتباط اصطلاحات در دو زبان ارائه کرده‌ایم.

در فصل ۵، به ارائه نتایج آزمایش‌ها برای بازیابی اطلاعات بین زبانی پرداخته شده است. طبیعی است که هر چه کیفیت ترجمه‌های استخراج شده بهتر باشد، کارآیی بازیابی اطلاعات بین زبانی نیز بهتر خواهد بود. نتایج آزمایش‌ها بر روی مجموعه همشهری با پرس‌وجوهای CLEF-2008, 2009 برای بازیابی اطلاعات بین زبانی انگلیسی-فارسی، نشان می‌دهند که با روش پیشنهادی مبنی بر شبکه اصطلاحات می‌توانیم به ۴۳/۳% میانگین متوسط دقت، ۴۵/۴% دقت در ۵ سند و ۴۴/۷% دقت در ۱۰ سند در مقایسه با کارآیی بازیابی تک زبانی دست یابیم. در نگرشی دیگر کلمات برون‌هشته

[1] TAN: Term Association Network

را تشخیص داده و آنها را از پرس‌وجوها حذف می‌کنیم، در این حالت نتایج آزمایش‌ها نشان می‌دهد که کارآیی سیستم بازیابی اطلاعات بین زبانی افزایش یافته و میانگین متوسط دقت به ۴۵.۸% بازیابی تک زبانی رسیده است. همچنین نتایج قابل توجه دیگری از ترکیب واژه‌نامه با ترجمه‌های استخراج شده از پیکره تطبیقی بدست آمده است. با استفاده از گسترش پرس‌وجو با شبه بازخورد ارتباطی توانسته‌ایم به ۶۱.۸% میانگین متوسط دقت بازیابی تک زبانی دست یافته و نسبت به استفاده از تنها واژه‌نامه حدود ۴۶.۵% بهبود کسب شده است. این نتایج نشان می‌دهد که ترجمه‌های استخراج شده از پیکره تطبیقی می‌تواند دامنه واژگان واژه‌نامه را گسترش داده و علاوه بر استخراج ترجمه مستقیم کلمه، کلمات مرتبط با آن را هم کشف کند. آزمایش‌های انجام شده را بر روی پیکره INFILE به صورت بازیابی اسناد انگلیسی توسط پرس‌وجوهای فارسی نیز انجام شده‌اند. نتایج کسب شده در فصل ۵ بررسی شده و در کل روندی مشابه نتایج پیکره همشهری دارند.

۲.۶ کارهای آتی

کارهای آتی این پژوهش را می‌توان به سه دسته کلی تقسیم کرد. (۱) بهبود کیفیت و گسترش پیکره تطبیقی ساخته شده. (۲) ارائه روش‌هایی برای افزایش کیفیت ترجمه‌های استخراج شده. (۳) استفاده از ترجمه‌های استخراج شده در زمینه‌های دیگر از جمله ماشین‌های ترجمه. در ادامه شرح مختصری بر روی کارهایی را که می‌توان در هر یک از این سه دسته انجام داد، ارائه می‌شود.

- به منظور افزایش کیفیت پیکره تطبیقی ساخته شده می‌توان از ترجمه‌های استخراج شده به عنوان منبعی برای ترجمه کلید واژه‌های زبان مبدأ در هنگام ساخت پیکره استفاده کرد. در حقیقت ساخت پیکره تطبیقی به صورت فرآیندی تکراری انجام می‌شود و پس از هر تکرار کیفیت پیکره توسط ترجمه‌های استخراج شده از خودش بهبود می‌یابد.

- به منظور افزایش دامنه پیکره تطبیقی ساخته شده می‌توان از منابع دیگری به غیر از اخبار، همانند وب استفاده کرد. از جمله منابع دو زبانه موجود در وب ویکی‌پدیا است. حجم اسناد فارسی موجود در ویکی‌پدیا در حال افزایش است و از آنجایی که هر سند به زبان‌های دیگرِ آن موضوع پیوند دارد، می‌توان به راحتی اسناد مرتبط به هم را از این پایگاه اطلاعات استخراج کرد. از دیگر مزایای ویکی‌پدیا وجود عنوان[۱] هر سند است که می‌توان از آنها به صورت منبع غنی ترجمه استفاده کرد.

[۱] title

- ما در این پایان‌نامه پس از ارائه مدل کلی شبکه‌های ارتباطی کلمات، برای اندازه‌گیری شباهت همسایگی‌ها مسیرهای به طول یک، دو و سه را در نظر گرفته‌ایم. در آزمایش‌های انجام شده در بخش ۴.۳.۵ با همین حالات خاص نشان داده‌ایم که استفاده از اطلاعات همسایگی‌ها برای افزایش دقت ترجمه‌های استخراج شده، سودمند است. مطالعه ارزشمند دیگری که در ادامه این کار می‌توان انجام داد، استفاده از روش‌های پیچیده‌تر مقایسه همسایگی‌ها و تکرار آزمایش‌ها بر روی نتایج بدست آمده از آن می‌باشد.

- از جمله نوآوری‌های این پژوهش استخراج ترجمه عبارات از پیکره‌های تطبیقی می‌باشد. نتایج آزمایش‌های ما در بخش ۶.۳.۵ نشان داده‌اند که ترجمه‌های بدست آمده برای عبارات، دقت نسبتاً خوبی دارند. هدف ما از این پژوهش، ارائه یک روش استخراج ترجمه عبارات بود ولی در ادامه کار می‌توان با تنظیم پارامترهای استفاده شده کیفیت ترجمه‌ها را بهبود داد.

- در بیشتر آزمایش‌های انجام شده در زمینه بازیابی اطلاعات بین زبانی، پرس‌وجوها و پیکره‌های تطبیقی معمولاً متعلق به یک حوزه می‌باشند. در این پایان‌نامه نیز، پیکره‌تطبیقی، پرس‌وجوها و اسناد بازیابی شده، هر سه در حوزه اخبار می‌باشند. اگر که پرس‌وجوها و پیکره تطبیقی از یک حوزه نباشند، بازیابی اطلاعات بین زبانی به مراتب مشکل‌تر می‌شود. در نتیجه به عنوان یک ایده جدید می‌توان چگونگی تأثیر ترجمه‌های استخراج شده از پیکره‌ی تطبیقی را بر روی پرس‌وجوهایی در حوزه‌ی دیگر بررسی کرد. به عنوان مثال، در ادامه‌ی این پژوهش می‌توان از پیکره تطبیقی ایجاد شده توسط متون اخبار برای ترجمه پرس‌وجوهایی بر روی صفحات وب استفاده کرد. حتی شاید بتوان در این مورد خاص از اطلاعات پیوندهای بین صفحات نیز استفاده کرد. به این منظور در راستای این پژوهش، از پیکره‌ی محک وب dotIR[1] [۱۵] برای بازیابی اطلاعات بین زبانی استفاده کرده‌ایم. ولی به دلیل مناسب نبودن پرس‌وجوها برای بازیابی اطلاعات بین زبانی، نتایج خوبی کسب نشده است، در نتیجه نتایج آزمایش‌های انجام شده را در این پایان‌نامه گزارش نکرده‌ایم. همچنین به نظر می‌رسد که نیاز است مطالعات بیشتری بر روی بهینه‌سازی پیکره‌ی محک وب فارسی انجام شود [۳۰].

- ایده دیگری که در مسأله بازیابی اطلاعات بین زبانی به ذهن می‌رسد، استفاده از یک زبان میانی است. به عنوان مثال فرض کنید برای دو زبان فارسی و ژاپنی منابع زبان‌شناسی غنی وجود ندارد. در این حالت می‌توان از زبان میانی معروفی همانند انگلیسی استفاده کرد. مثلاً برای ترجمه پرس‌وجو از فارسی به ژاپنی می‌توان با فرض اینکه منابع زبان‌شناسی برای زوج زبان‌های فارسی-انگلیسی و انگلیسی-ژاپنی موجود هستند، ابتدا پرس‌وجو را به انگلیسی ترجمه کرد و سپس ترجمه انگلیسی را به ژاپنی ترجمه کرد. در اینجا ایده‌ای که به ذهن می‌رسد استفاده از

[1] http://ece.ut.ac.ir/DBRG/webir/

پیکره‌های تطبیقی بین این زبان‌هاست. یعنی اگر پیکره‌ای تطبیقی بین زبان‌های فارسی-انگلیسی و انگلیسی-ژاپنی وجود داشته باشد، آنگاه می‌توان ترجمه‌های استخراج شده از این پیکره‌ها را با یکدیگر ترکیب کرد و در نهایت ترجمه‌ی کلمات فارسی به ژاپنی را از این پیکره‌های تطبیقی استخراج کرد.

- به عنوان کارهای آینده‌ی این پایان‌نامه می‌توان از ترجمه‌های استخراج شده از پیکره تطبیقی در زمینه‌هایی به غیر از بازیابی اطلاعات بین زبانی استفاده کرد. به عنوان مثال توسعه ماشین‌های ترجمه، زمینه‌ای رو به رشد است و ترجمه‌های استخراج شده از پیکره‌های تطبیقی می‌توانند در بهبود کارآیی آنها نقش بسزایی ایفا کنند.

- به دلیل نبود ریشه‌یابِ استاندارد در زبان فارسی، در آزمایش‌هایی که انجام داده‌ایم کلمات فارسی را ریشه‌یابی نکرده‌ایم. در حالی که مطلوب این است که تمام کلمات ریشه‌یابی شوند. به عنوان مثال در آزمایش‌های انجام شده برای کلمه *economic* هم کلمه *اقتصاد* و هم کلمه *اقتصادی* استخراج شده است. اگر کلمات هم‌ریشه یکی شوند می‌توان با دقت بهتری ترجمه‌ها را استخراج کرد.

- یکی دیگر از تحقیقات جالبی که در زمینه بازیابی اطلاعات بین زبانی می‌توان انجام داد ترکیب ویژگی‌های اسناد به منظور رتبه‌بندی می‌باشد. در راستای تحقق این هدف، تحقیقاتی بر روی رتبه‌بندی در بازیابی اطلاعات توسط روش‌های مختلف ترکیب اطلاعات[۱] انجام داده‌ایم [۳۰] و در نظر داریم که در ادامه کارهای آینده این روش‌ها را به منظور بازیابی اطلاعات بین زبانی بکار ببریم.

[۱] information fusion

پیوست آ

پرس‌وجوهای پیکره INFILE

۴۵ پرس‌وجوی انگلیسی پیکره INFILE همراه با ترجمه‌های فارسی در جدول آ.۱ آمده است.

جدول آ.۱: پرس‌وجوهای پیکره INFILE

عنوان پرس‌وجوی ترجمه شده به فارسی	عنوان پرس‌وجوی انگلیسی	شماره پرس‌وجو
مبارزه با دوپینگ در ورزش	Fight against doping in sport	۱۰۱
اقتصاد ورزش	Sport economy	۱۰۲
سازمان مسابقات ورزشی بین المللی	Organization of international sports competitions	۱۰۳
ورزش ، خشونت و نژادپرستی	Sport, violence and racism	۱۰۵
ورزش معلولان	The disabled, adapted sports	۱۰۶
رای گیری الکترونیکی	Electronic voting	۱۰۷
سیاستمداران و رسانه	Politicians and media	۱۰۹
سیاست‌های نظر سنجی‌ها	Polls policies	۱۱۱
شکاف دیجیتال	Digital Divide	۱۱۳
سینمای هنر	Art-house cinema	۱۱۴
موزه‌های رایگان	The free museums	۱۱۵

روزنامه‌های رایگان در فرانسه	Free Newspapers in France	۱۱۶
کاریکاتور	The cartoon	۱۱۷
افزایش قیمت نفت	Rising oil prices	۱۱۸
فروش در فرانسه	The sales in France	۱۲۰
سوپرمارکت	The Supermarkets	۱۲۱
سرنوشت مغازه‌های محلی	The fate of local shops	۱۲۲
چالش‌های تجارت الکترونیک	The challenges of e-commerce	۱۲۳
مسئله تبت	The issue of Tibet	۱۲۵
سیاست مهاجرت اروپا	The European immigration policy	۱۲۶
بحران دارفور	The crisis in Darfur	۱۲۷
درگیری اسرائیل و فلسطین	The Israeli-Palestinian conflict	۱۲۸
شورش فارک	The FARC rebelion	۱۲۹
سازمان بازی‌های المپیک ۲۰۱۲	2012 Olympic Games organization	۱۳۰
مخاطرات دولت الکترونیک	E-government stakes	۱۳۱
شبکه بی‌سیم و سلامتی	Wireless network and health	۱۳۲
رقابت برای حق امتیاز پخش فوتبال	Competition over soccer broadcating rights	۱۳۳
پروتکل کیوتو	Issues of the Kyoto Protocol	۱۳۵
آلودگی هوا و کیفیت هوا	Air pollution and air quality	۱۳۶
مبارزه با تغییر آب و هوا	Fight against climate change	۱۳۷
داروها و بیوتکنولوژی	Drugs and biotechnology	۱۳۸
سوخت زیستی	Biofuels	۱۳۹
مصرف میوه‌ها و سبزیجات در پیشگیری از سرطان	Fruits and vegetables intakes and cancer prevention	۱۴۰
دوپینگ و سلامت ورزشکاران	Doping and athletes’s health	۱۴۱
حفظ تنوع زیستی	Biodiversity preservation	۱۴۲
آنفلونزای مرغی	Avian influenza	۱۴۳

فناوری نانو و علم نانو	Nanotechnologies and nanosciences	۱۴۴
اصلاح ژنتیکی گیاهان	Genetically modified plants	۱۴۵
سرطان پروستات	Prostate cancer	۱۴۶
انرژی تجدید شدنی	Renewable energy	۱۴۸
تحقیقات علمی در قطب شمال	Scientific research in Arctic	۱۴۹
زنان در جهان کار	Woman in the world of work	۱۵۰

مراجع

[1] S. Abdul-Rauf and H. Schwenk, “On the use of comparable corpora to improve SMT performance,” in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 16–23.

[2] S. Abdul-Rauf and H. Schwenk, “Exploiting comparable corpora with ter and terp,” in *BUCC ’09: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora*, 2009, pp. 46–54.

[3] S. Abedinzadeh, F. Taghiyareh, and F. Oroumchian, “A Collection of Persian-English Documents,” in *Iranian Conference on Electrical Engineering (ICEE 2007)*, 2007, pp. 119–132.

[4] E. Airio, “Who benefits from CLIR in web retrieval?” *Journal of Documentation*, vol. 64, no. 5, pp. 760–778, 2008.

[5] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian, “Hamshahri: A standard Persian text collection,” *Knowledge-Based Systems*, vol. 22, no. 5, pp. 382–387, 2009.

[6] A. AleAhmad, E. Kamalloo, A. Zareh, M. Rahgozar, and F. Oroumchian, “Cross language experiments at persian@ clef 2008,” *Evaluating Systems for Multilingual and Multimodal Information Access*, pp. 105–112, 2009.

[7] J. Allan, J. Callan, W. Croft, L. Ballesteros, J. Broglio, J. Xu, and H. Shu, “Inquery at TREC-5,” in *Proceedings of TREC*, vol. 5, 1997, pp. 119–132.

[8] L. Ballesteros and W. Croft, “Resolving ambiguity for cross-language retrieval,” in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 64–71.

[9] B. Bekavac, P. Osenova, K. Simov, and M. Tadić, “Making monolingual corpora comparable: a case study of Bulgarian and Croatian,” in *LREC*, 2004, pp. 1187–1190.

[10] R. Besançon, S. Chaudiron, D. Mostefa, O. Hamon, I. Timimi, and K. Choukri, “Overview of CLEF 2008 INFILE pilot track,” *Evaluating Systems for Multilingual and Multimodal Information Access*, pp. 939–946, 2009.

[11] M. Bijankhan, “Role of language corpora in writing grammar: introducing a computer software,” *Iranian Journal of Linguistics*, no. 38, pp. 38–67, 2004.

[12] M. Braschler and P. Schäuble, “Multilingual information retrieval based on document alignment techniques,” in *ECDL*, 1998, pp. 183–197.

[13] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[14] N. Collier, A. Kumano, and H. Hirakawa, "An application of local relevance feedback for building comparable corpora from news article matching," *NII J (Natl Inst Inform)*, vol. 5, pp. 9–23, 2003.

[15] E. Darrudi, H. Baradaran Hashemi, A. Aleahmad, A. Habibian, A. Zareh Bidoki, A. Shakery, and M. Rahgozar, "A Standard Web test collection for .ir domain," *ITRC technical report*, Fall 2009.

[16] M. W. Davis, "On the effective use of large parallel corpora in cross-language text retrieval," *Cross-language information retrieval*, pp. 11–22, 1998.

[17] L. Dimitrova, N. Ide, V. Petkevic, T. Erjavec, H. J. Kaalep, and D. Tufis, "Multext-east: parallel and comparable corpora and lexicons for six central and eastern european languages," in *ACL*, 1998, pp. 315–319.

[18] A. Eisele and J. Xu, "Improving machine translation performance using comparable corpora," in *Proceedings of 3rd Workshop on Building and Using Comparable Corpora , Malta*, pp. 35–41.

[19] R. Fattahi and M. Davarpanah, "Applying Natural Language Processing Techniques for Effective Persian-English Cross-Language Information Retrieval," 2010.

[20] N. Ferro and C. Peters, "CLEF 2009 Ad Hoc Track Overview: TEL and Persian Tasks," *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pp. 13–35.

[21] K. Fløttum, T. Gedde-Dahl, and T. Kinn, *Academic voices: across languages and disciplines*. John Benjamins Publishing Co, 2006.

[22] M. Franz, J. McCarley, and S. Roukos, "Ad hoc and multilingual information retrieval at IBM," *NIST special publication SP*, pp. 157–168, 1999.

[23] P. Fung, "A pattern matching method for finding noun and proper noun translations from noisy parallel corpora," in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1995, pp. 236–243.

[24] P. Fung and P. Cheung, "Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em," in *Proceedings of EMNLP*, 2004, pp. 57–63.

[25] P. Fung and L. Yee, "An IR approach for translating new words from nonparallel, comparable texts," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 1998, pp. 414–420.

[26] F. Gey and A. Chen, "Phrase discovery for English and cross-language retrieval at TREC-6," in *In Proceedings of the sixth text retrieval conference (TREC-6*. Citeseer, 1998.

[27] M. Ghayoomi, S. Momtazi, and M. Bijankhan, "A study of corpus development for Persian," *International Journal of Asian Language Processing*, vol. 20, no. 1, pp. 17–33, 2010.

[28] C. Goutte, N. Cancedda, and M. Dymetman, *Learning machine translation*. The MIT Press, 2009.

[29] G. Grefenstette, *Cross-language information retrieval*. Kluwer Academic Publishers Norwell, MA, USA, 1998.

[30] H. Hashemi, N. Yazdani, A. Shakery, and M. Naeini, “Application of Ensemble Models in Web Ranking,” in *to be appear in proceedings of the 5th International Symposium on Telecommunications (IST2010)*, 2010.

[31] A. Hassan, H. Fahmy, and H. Hassan, “Improving named entity translation by exploiting comparable and parallel corpora,” in *Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP), AMML Workshop*, 2007.

[32] D. He, D. Oard, J. Wang, J. Luo, D. Demner-Fushman, K. Darwish, P. Resnik, S. Khudanpur, M. Nossal, M. Subotin *et al.*, “Making MIRACLEs: Interactive translingual search for Cebuano and Hindi,” *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 2, no. 3, pp. 219–244, 2003.

[33] D. He and J. Wang, “Cross-language information retrieval,” *Information Retrieval: Searching in the 21st Century*, pp. 233–254, 2009.

[34] T. Hedlund, E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola, and K. Järvelin, “Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000–2002,” *Information Retrieval*, vol. 7, no. 1, pp. 99–119, 2004.

[35] S. Hewavitharana and S. Vogel, “Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources,” in *Proceedings of the Workshop on Comparable Corpora, LREC08*, 2008, pp. 7–10.

[36] D. Huang, L. Zhao, L. Li, and H. Yu, “Mining large-scale comparable corpora from chinese-english news collections,” in *Coling 2010: Posters*. Beijing, China: Coling 2010 Organizing Committee, August 2010, pp. 472–480.

[37] S. Karimi, “Machine transliteration of proper names between english and persian,” Ph.D. dissertation, RMIT University, Melbourne, Victoria, Australia, 2008.

[38] K. Kishida, “Technical issues of cross-language information retrieval: a review,” *Information Processing & Management*, vol. 41, no. 3, pp. 433–455, 2005.

[39] K. Koskenniemi, “Two-level morphology: A general computational model for word-form recognition and production,” *Publications of the Department of General Linguistics, University of Helsinki*, vol. 11, 1983.

[40] J. Lafferty and C. Zhai, “Document language models, query models, and risk minimization for information retrieval,” in *SIGIR*, 2001, pp. 111–119.

[41] A. Laroche and P. Langlais, “Revisiting context-based projection methods for term-translation spotting in comparable corpora,” in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, August 2010, pp. 617–625.

[42] L. Lee, A. Aw, M. Zhang, and H. Li, "Em-based hybrid model for bilingual terminology extraction from comparable corpora," in *Coling 2010: Posters*. Beijing, China: Coling 2010 Organizing Committee, August 2010, pp. 639–646.

[43] B. Li and E. Gaussier, "Improving corpus comparability for bilingual lexicon extraction from comparable corpora," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, August 2010, pp. 644–652.

[44] B. Maia and S. Matos, "Corpógrafo v. 4–tools for researchers and teachers using comparable corpora," in *LREC 2008 Workshop on Comparable Corpora (LREC 2008)*, 2008, pp. 79–82.

[45] A. McEnery and R. Xiao, "Parallel and comparable corpora: What are they up to?" *In Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters, Clevedon, UK, 2007.*

[46] P. McNamee, "JHU ad hoc experiments at CLEF 2008," *Evaluating Systems for Multilingual and Multimodal Information Access*, pp. 170–177, 2009.

[47] P. McNamee and J. Mayfield, "Comparing cross-language query expansion techniques by degrading translation resources," in *SIGIR*, 2002, pp. 159–166.

[48] T. Miangah, "Solving the Polysemy Problem of Persian Words Using Mutual Information Statistics," in *Proceedings of the Corpus Linguistics Conference (CL2007)*, 2007, pp. 1–8.

[49] T. Miangah, "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)*, 2008, pp. 81–84.

[50] T. Miangah, "Constructing a Large-Scale English-Persian Parallel Corpus," *Meta: Translators' Journal*, vol. 54, no. 1, pp. 181–188, 2009.

[51] T. Miangah and A. Nezarat, "A Method for Cross-Language Retrieval of Chunks using Monolingual and Bilingual Corpora," *International Journal of Computer Applications*, vol. 10, no. 10, pp. 20–23, 2010.

[52] C. Monz and B. Dorr, "Iterative translation disambiguation for cross-language information retrieval," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005, pp. 520–527.

[53] E. Morin, B. Daille, K. Takeuchi, and K. Kageura, "Bilingual terminology mining-using brain, not brawn comparable corpora," in *45th Annual Meeting of the Association for Computational Linguistics,*, vol. 45, no. 1, 2007, pp. 664–671.

[54] E. Morin, B. Daille, K. Takeuchi, and K. Kageura, "Brains, not brawn: The use of 'smart' comparable corpora in bilingual terminology mining," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 1, pp. 1–23, 2010.

[55] D. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Computational Linguistics*, vol. 31, no. 4, pp. 477–504, 2005.

[56] D. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Comput. Linguist*, vol. 31, no. 4, pp. 477–504, 2005.

[57] D. Munteanu and D. Marcu, "Extracting parallel sub-sentential fragments from non-parallel corpora," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 81–88.

[58] D. Oard and A. Diekema, "Cross-language information retrieval," *Annual Review of Information Science and Technology*, vol. 33, pp. 223–256, 1998.

[59] P. Otero and J. Campos, "Automatic Generation of Bilingual Dictionaries Using Intermediary Languages and Comparable Corpora," in *Computational Linguistics and Intelligent Text Processing: 11th International Conference, CICLing 2010, Iasi, Romania, March 21-27, 2010, Proceedings*. Springer Verlag, 2010, p. 473.

[60] S. Papadimitriou and C. Faloutsos, "Cross-outlier detection," *Advances in Spatial and Temporal Databases*, pp. 199–213, 2003.

[61] E. Picchi and C. Peters, "Cross-language information retrieval: A system for comparable corpus querying," in *SIGIR'96, Workshop on Cross-Linguistic Information Retrieval*, 1996, pp. 24–33.

[62] M. T. Pilevar and H. Feili, "PersianSMT: A first attempt to english-persian statistical machine translation," in *JADT*, 2010.

[63] A. Pirkola, E. Leppanen, and K. Järvelin, "The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval," *Information Research*, vol. 7, no. 2, 2002.

[64] S. Pourmahmoud and M. Shamsfard, "Semantic Cross-lingual Information Retrieval," in *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on*. IEEE, 2008, pp. 1–4.

[65] C. Quirk, R. Udupa, and A. Menezes, "Generative models of noisy translations with applications to parallel fragment extraction," *Proc. of the MT Summit XI*, pp. 321–327, 2007.

[66] R. Rapp, "Identifying word translations in non-parallel texts," in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1995, pp. 320–322.

[67] R. Rapp, "Automatic identification of word translations from unrelated English and German corpora," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 1999, pp. 519–526.

[68] R. Rapp and M. Zock, "Automatic Dictionary Expansion Using Non-parallel Corpora," *Advances in Data Analysis, Data Handling and Business Intelligence*, pp. 317–325, 2010.

[69] P. Resnik, "Mining the web for bilingual text," in *ACL*, 1999, pp. 527–534.

[70] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval," in *SIGIR*, 1994, pp. 232–241.

[71] F. Sadat, "Exploiting Comparable Corpora for Cross-Language Information Retrieval," *PRICAI 2010: Trends in Artificial Intelligence*, pp. 662–667, 2010.

[72] F. Sadat, "Using Comparable Corpora to Improve the Effectiveness of Cross-Language Information Retrieval," in *Advances in Natural Language Processing: 7th International Conference on Nlp, Icetal 2010, Reykjavik, Iceland, August 16-18, 2010, Proceedings*. Springer, 2010, p. 320.

[73] F. Sadat, M. Yoshikawa, and S. Uemura, "Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*. Association for Computational Linguistics, 2003, pp. 141–144.

[74] F. Sadat, M. Yoshikawa, and S. Uemura, "Learning bilingual translations from comparable corpora to cross-language information retrieval: hybrid statistics-based and linguistics-based approach," in *Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11*. Association for Computational Linguistics, 2003, pp. 57–64.

[75] F. Sadat, M. Yoshikawa, and S. Uemura, "Enhancing cross-language information retrieval by an automatic acquisition of bilingual terminology from comparable corpora," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 397–398.

[76] G. Salton, "Automatic processing of foreign language documents," in *Proceedings of the 1969 conference on Computational linguistics*. Association for Computational Linguistics, 1969, p. 28.

[77] A. Shakery, "Probabilistic score propagation in information retrieval," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, Illinois, 2008.

[78] S. Sharoff, "Creating general-purpose corpora using automated search engine queries," in *WaCky! Working papers on the Web as Corpus*, 2006.

[79] P. Sheridan and J. Ballerini, "Experiments in multilingual information retrieval using the SPIDER system," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 58–65.

[80] P. Sheridan and J. P. Ballerini, "Experiments in multilingual information retrieval using the spider system," in *SIGIR*, 1996, pp. 58–65.

[81] J. Sinclair, "Preliminary recommendations on corpus typology," *EAGLES Document TCWG-CTYP/P (available from http://www. ilc. pi. cnr. it/EAGLES/corpustyp/corpustyp. html)*, 1996.

[82] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996, pp. 21–29.

[83] R. Steinberger, B. Pouliquen, and C. Ignat, "Navigating multilingual news collections using automatically extracted information," *CIT*, vol. 13, no. 4, pp. 257–264, 2005.

[84] T. Talvensaari, "Effects of aligned corpus quality and size in corpus-based CLIR," *Advances in Information Retrieval*, pp. 114–125, 2008.

[85] T. Talvensaari, M. Juhola, J. Laurikkala, and K. Järvelin, "Corpus-based cross-language information retrieval in retrieval of highly relevant documents," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 3, pp. 322–334, 2007.

[86] T. Talvensaari, J. Laurikkala, K. Järvelin, and M. Juhola, “Creating and exploiting a comparable corpus in cross-language information retrieval,” *TOIS*, vol. 25, no. 4, 2007.

[87] T. Talvensaari, A. Pirkola, K. Järvelin, M. Juhola, and J. Laurikkala, “Focused web crawling in the acquisition of comparable corpora,” *Information Retrieval*, vol. 11, pp. 427–445, 2008.

[88] K. Tanaka and H. Iwasaki, “Extraction of lexical translations from non-aligned corpora,” in *Proceedings of the 16th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1996, pp. 580–585.

[89] T. Tao and C. X. Zhai, “Mining comparable bilingual text corpora for cross-language information integration,” in *SIGKDD*, 2005, pp. 691–696.

[90] D. Trieschnigg, D. Hiemstra, F. deJong, and W. Kraaij, “A Cross-lingual Framework for Monolingual Biomedical Information Retrieval,” in *CIKM10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 2010.

[91] R. Udupa, K. Saravanan, A. Kumaran, and J. Jagarlamudi, “Mining named entity transliteration equivalents from comparable corpora,” in *CIKM ’08: Proceeding of the 17th ACM conference on Information and knowledge management*. New York, NY, USA: ACM, 2008, pp. 1423–1424.

[92] M. Utiyama and H. Isahara, “Reliable measures for aligning Japanese-English news articles and sentences,” in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 72–79.

[93] T. Utsuro, T. Horiuchi, Y. Chiba, and T. Hamamoto, “Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on WWW news sites,” in *AMTA*, 2002, pp. 165–176.

[94] K. Yu and J. Tsujii, “Extracting bilingual dictionary from comparable corpora with dependency heterogeneity,” in *NAACL ’09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 121–124.

[95] B. Zhao and S. Vogel, “Adaptive Parallel Sentences Mining from Web Bilingual News Collection,” in *Proceedings of the 2002 IEEE International Conference on Data Mining*. IEEE Computer Society, 2002, pp. 745–748.

واژه‌نامه فارسی به انگلیسی

ابر پیوند	hyper link
اصطلاح	term
اطلاعات متقابل	mutual information
ایست واژه	stop word
اینترنت	Internet
بازخورد	feedback
بازخورد ارتباطی	relevance feedback
بازیابی اطلاعات	Information Retrieval (IR)
بازیابی اطلاعات بین زبانی	Cross-Language Information Retrieval (CLIR)
بایگرام	bigram
برخط	online
بررسی صحّت ترجمه	Translation Validity Check
برون‌خط	offline
برون‌هشته	outlier
برون‌هشته متقاطع	cross-outlier
برچسب	tag
برچسب گذاری	tagging
بسامد	frequency
بسامد اصطلاح	Term Frequency (TF)
بسامد سند	document frequency
بسامد مجموعه	collection frequency
بن واژه	lemma
بن واژه سازی	lemmatization
بیشینه بی‌نظمی	maximum entropy
بیشینه‌ی درست‌نمایی	maximum likelihood
بی‌بی‌سی	BBC
بی‌نظمی	entropy
تخمین احتمال	probability estimation
ترکیب اطلاعات	information fusion
تصفیه	filter
تطابق رشته	string matching
تطابق واژه هم‌ریشه	cognate matching
توافقی	adaptive
تک مقداره	single-value
تک‌زبانی	monolingual
حد آستانه	threshold
خزش	crawl
خط بنیادی	baseline
خودکار	automatic
دانش ریخت‌شناسی	morphological knowledge
دستی	manual

دقت .... precision
دو زبانه .... bilingual
رتبه .... rank
رخداد همزمان .... co-occurrence
ریشه یابی .... stemming
زمینه .... context
زوج هم‌ترازی .... alignment pair
سربرگ .... header
سند .... document
شبه بازخورد ارتباطی .... pseudo relevance feedback
صریح .... explicit
ضمنی .... implicit
طرح .... scheme
علامت .... token
عکس بسامد سند .... Inverse Document Frequency (IDF)
غیر موازی .... non-parallel
فراخوانی .... recall
فرهنگ جامع .... thesaurus
فرهنگ نگار .... lexicographer
فضای بردار .... vector space
ماشین ترجمه .... Machine Translation (MT)
ماشین‌خوان .... machine readable
متن آزاد .... free text
متوسط دقت .... Average Precision (AP)
مجموعه دادگان .... dataset
مدل ترکیبی .... mixture model
مدل زبانی .... language model
مدل زبانی پرس‌وجو .... query language model
مقوله دستوری .... part of speech
مقیاس .... scale
موتور جستجو .... search engine
میانگین متوسط دقت .... Mean Average Precision (MAP)
نرمال‌سازی بردار محوری .... Pivoted vector normalization
نرمال‌سازی کسینوسی استاندارد .... standard cosine normalization
نمایه .... index
نمایه سازی .... indexing
نویسه‌گردانی .... transliteration
هرس .... pruning
همبستگی .... correlation
همبستگی پیرسون .... Pearson correlation
هم‌ترازی .... alignment
هم‌ریشه .... cognate
وابستگی .... association
واژگان .... lexical
واژگان نظارت شده .... controlled vocabulary
وب .... web
پردازش زبان طبیعی .... natural language processing
پرس‌وجو .... query
پیش زمینه .... background
پیکره تطبیقی .... comparable corpus
پیکره موازی .... parallel corpus
چارچوب .... framework
کاوش .... mining

کلمات خارج از واژگان .... Out Of Vocabulary (OOV)
کلمه .... word
کلید واژه .... keyword
گره .... node
گسترش پرس‌وجو .... query expansion
گوگل .... Google

Abstract

The fast growth of the World Wide Web and the availability of information in different languages have attracted much attention in research on cross-language information retrieval (CLIR). CLIR concerns the problem of expressing queries in one language and finding documents in different languages. Since queries and documents are in different languages, some kind of translation must take place before matching. Dictionaries, machine translators, parallel corpora and comparable corpora are common translation resources in CLIR. However, dictionaries, machine translators and parallel corpora are expensive resources and do not exist for all language pairs. While, comparable corpora are far easier to obtain.

In this thesis, cross-language information retrieval using comparable corpora is studied in Persian and English languages. Comparable corpora are collections of texts in different languages that are about similar topics. For example, news articles about the same events published in different languages create comparable corpora. The availability of such linguistic resources is one of its advantages that causes the construction and use of comparable corpora to be considered as a valid field in CLIR. So far, there has not been conducted any research on doing cross-language information retrieval in Persian language using comparable corpora.

In this research, the first big Persian-English comparable corpus (UTPECC) is built from two independent news collections: BBC News in English and Hamshahri news in Persian. The similarity of the document topics and their publication dates are used to align the documents in these sets and several alternatives are tried for constructing and assessing the comparable corpora. We mine the comparable corpus to extract high quality translation knowledge by presenting a novel way of extracting translations based on Terms Association Network (TAN) which exploits word co-occurrences in monolingual data as well as word associations in different languages. The extracted translation knowledge may be in forms of words and phrases, thus they can be effective for cross-language information retrieval. Further, we do translation validity check by detecting the mistranslated terms as outliers. Evaluation results on different data sets show that translating queries using UTPECC and constructing query language models significantly outperforms using simple dictionary-based method. Moreover, the presented methods are specially effective in translating Out Of Vocabulary words and also expanding query words based on their related words.

Keywords: *cross-language information retrieval, Persian-English comparable corpus, term association network.*

University of Tehran
School of Electrical and Computer Engineering

Using Comparable Corpora for English-Persian Cross Language Information Retrieval

By
Homa Baradaran Hashemi

Under Supervision of
Dr. Azadeh Shakery

A thesis submitted to the Graduate Studies Office
in partial fulfillment of the requirements
for the degree of M.Sc
in
Computer Engineering

February 2011