نشریه پژوهش‌های زبان‌شناسی
سال هفتم، شماره دوم، شماره ترتیبی ۱۳، پاییز و زمستان ۱۳۹۴
تاریخ وصول: ۱۳۹۳/۱۱/۱
تاریخ اصلاحات: ۱۳۹۴/۶/۱۶
تاریخ پذیرش: ۱۳۹۵/۱/۳۰
صص ۱۴-۱

استخراج خودکار معادل‌های واژگانی از پیکره‌های دو زبانه موازی

طیبه موسوی میانگاه*
مهشید شکیبا**

چکیده

امروزه به مدد ظهور انواع فناوری‌های رایانه‌ای، پیکره‌های زبانی نقش بسیار مهمی در حل انواع مختلف مسائل زبانشناختی ایفا می‌کنند. پیکره‌های دو زبانه موازی در سطح جمله و در سطح واژه می‌توانند برای بازیابی واحدهای تک‌واژه‌ای و یا حتی چند واژه‌ای براحتی مورد استفاده قرارگیرند که این امر کاربردهای مفیدی در حوزه‌های مختلف رایانه و زبان خواهد داشت. هدف این مقاله به‌کارگیری یک پیکره موازی انگلیسی ° فارسی ازقبل طراحی‌شده درجهت ساخت یک مطابقه[1] (کشف اللغات) دو زبانهٔ کارآمد با استفاده از آمارهٔ اطلاعات متقابل است. در اینجا از آماره اطلاعات متقابل استفاده می‌شود تا همترازی در سطح واژه بین جملات انگلیسی و فارسی پیکرهٔ مورد نظر صورت‌گیرد. یک پیکرهٔ زبانی همترازشده در سطح واژه مسلماً کاربردهای زیادی از جمله در تهیه نرم‌افزار حافظهٔ ترجمه، مدیریت مجموعه اصطلاحات، بازیابی اطلاعات دوزبانه، سیستم ترجمه ماشینی مبتنی برآمار و مانند آن دارد. با استفاده از یک الگوریتم ابتکاری آزمایشی ترتیب داده شده و مقایسه‌ای بین برونداد همترازسازی خودکار با جملات همترازشده توسط مترجم انسانی صورت‌گرفت. نتایج این آزمایش نشان داد که برنامه مطابقه گزارش‌شده در این تحقیق می‌تواند صحتی معادل ۷۵ درصد را به دست‌آورد.

کلیدواژه‌ها: مطابقه دو زبانه، ترجمه خودکار، بازیافت اطلاعات، استخراج معادل‌های واژگانی، پیکره موازی انگلیسی-فارسی

[1] concordance

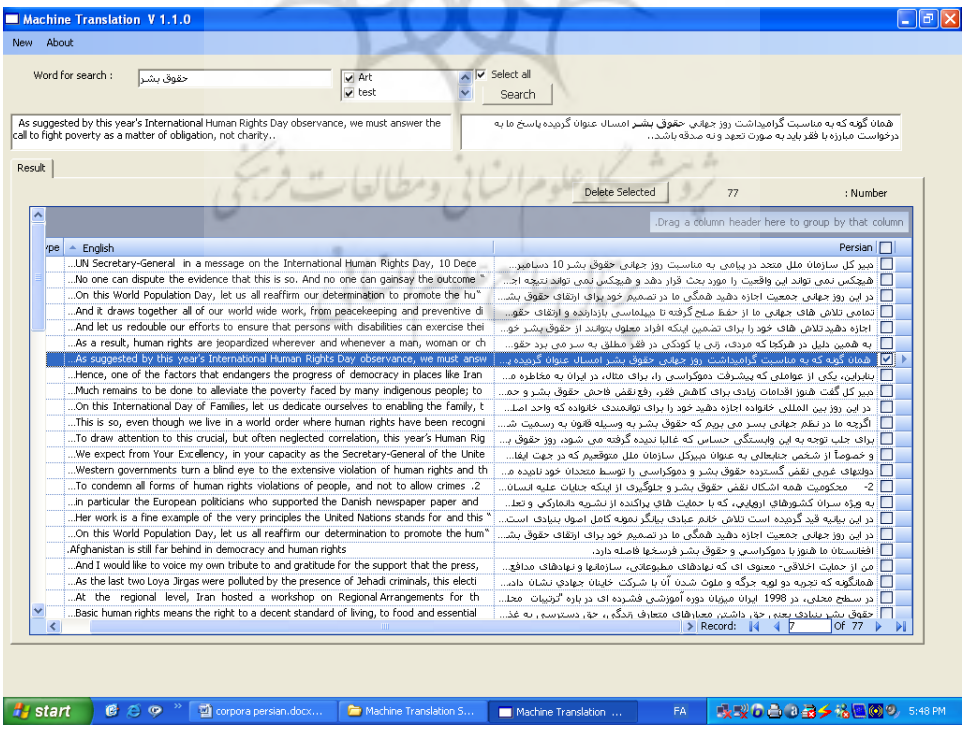
* دانشیار گروه زبان شناسی دانشگاه پیام نور Mosavit@pnu.ac.ir
** کارشناس ارشد گروه کامپیوتر دانشگاه پیام نور Mahshid.Shakiba@gmail.com

۱. مقدمه

در سالهای اخیر ساخت و استفاده از انواع مختلف پیکره‌های زبانی، از جمله کاربردهای رایانه‌ای هستند که در اختیار پژوهشگران حوزه ترجمه ماشینی، آموزش و یادگیری زبان، مطالعات ترجمه، بازیافت اطلاعات دو زبانه، واژه‌نگاری و غیره قرارگرفته‌اند. تا قبل‌از توسعه فناوری اطلاعات و ارتباطات، پیکره‌ها و نرم‌افزارهای کشف‌اللغات بندرت در دسترس پژوهشگران زبان قرارداشتند تا بتوانند در مورد زبان، محتوا، و یا امور مربوط به ترجمه اطلاعاتی را از درون آن‌ها استخراج نمایند.

ولی امروزه با ظهور فناوری‌های رایانه‌ای و تولید حجم وسیع متون ترجمه، منابع پیکره‌ای قابل‌دسترس برای محققان اشتیاق زیادی را در آنان برای ساخت و بهره‌برداری از پیکره‌ها به وجود آورده است. پیکره‌های ترجمه یا پیکره‌های موازی در رشته زبانشناسی رایانه‌ای بویژه در سیستم‌های ترجمه ماشینی کاربردهای وسیعی دارند. در حقیقت، پیکره‌های موازی ابزار اصلی برای انتخاب ترجمه‌های ممکن واحدهای واژگانی و متعاقب آن یافتن محتمل‌ترین ترجمه‌ها در یک سیستم ترجمه ماشینی مبتنی بر آمار هستند.

مطابقه نرم‌افزاری است برای مشاهده پیکره که فهرستی از رخدادهای یک واژهٔ خاص، قسمتی از یک واژه یا ترکیبی از چند واژه را در بافت نشان‌می‌دهد. به واحد جستار کلیدواژه گفته می‌شود. رایج‌ترین شکل نمایش در یک مطابقه نشان‌دادن تعدادی سطر است با ساختار کلیدواژه در بافت که در آن هر کلیدواژه با طول بافت معینی از هر دو طرف در مرکز سطر نشان داده می‌شود. نمونه‌ای از یک مطابقه دو زبانه انگلیسی ° فارسی (موسوی میانگاه، ۲۰۰۹) در شکل (۱) نشان داده شده‌است.



شکل ۱) رکوردهای تولیدشده توسط مطابقه موازی درسطح جمله برای جستار "حقوق بشر"

[1] Keyword in Context

تمام پیکره‌های ماشین‌خوان[1] به محققان در محاسبه فراوانی رخداد واژه یا عبارت مورد جستجو (جستار) کمک می‌کنند تا با انتخاب واژه یا عبارت مورد نظر تمام رکوردها (سطور) مطابقه که در آن جستار مورد نظر وجود دارد در صفحه نمایش ظاهر شود. بعلاوه، اکثر بسته‌های تحلیل پیکره شامل مطابقه نیز می‌باشند که محققان را در یافتن تمام رخدادهای جستارهای موردنظر با امکان مرتب‌کردن داده‌های نمایش داده شده در صفحه و نیز همراه با محدوده بافت متعلقه از سمت چپ و راست یاری می‌نمایند (هیرویوکی و توشیکو، ۱۹۹۶). واژه‌های خاص، قسمتی از یک واژه، یا گروهی از واژه‌ها که توسط برنامه مطابقه نمایش داده می‌شوند همه از یک پیکرهٔ متنی استخراج می‌شوند. بدیهی است که هرچه حجم پیکره بیشتر باشد، برونداد برنامه دقیق‌تر خواهد بود. درواقع، مطابقه‌های دوزبانه از جمله ابزاری هستند که کار سخت ترجمه را آسان‌تر می‌نمایند.

یک مطابقه موازی از این جهت با یک مطابقه یک‌زبانه متفاوت است که درمورد اول برنامه مطابقه براساس مجموعه‌ای از متون یک‌زبانه ساخته می‌شود مثل برنامه مایکروکانکورد (جونز، ۱۹۸۶) و ورد اسمیت (اسکات، ۲۰۰۰)، درحالی‌که درمورد دوم برنامه مطابقه براساس مجموعه‌ای از متون موازی دو یا چندزبانه ساخته می‌شود. این بدان معناست که در مطابقه‌های موازی کاربر می‌تواند جستار خود را به یک زبان نوشته و جملات متناظر برای آن جستار را نه تنها به همان زبان بلکه به زبان یا زبان‌های دیگر دریافت نماید. دو نمونه از پرکاربردترین مطابقه‌های موازی پاراکانکورد (بارلو، ۲۰۰۲) و مولتی‌کانکورد (جونز، ۱۹۹۸) هستند.

از طرف دیگر، مطابقه موازی در سطح واژه نوع خاصی از مطابقه دوزبانه است که در آن جستجوی کاربر برای یک زنجیره در زبان مبدأ منتهی به یافتن جملات متناظر با آن جستار در هر دو زبان مبدأ و مقصد شده، درحالی‌که زنجیره جستار در جمله زبان مبدأ و ترجمه آن در جمله مقصد پررنگ و کاملاً مشخص‌شده است. داشتن این توانایی برای مطابقه‌ها مستلزم دستیابی به فنون پیشرفته همترازسازی در سطح واژه است. البته در بسیاری از موارد هم شناسایی ترجمهٔ دقیق جستار مورد نظر امکان پذیر نمی باشد.

در سال‌های اخیر بسیاری از سیستم‌های ترجمهٔ ماشینی با استفاده از همترازسازی در سطح واژه، پیکره‌های موازی که توسط مدل‌های آی‌بی‌ام (براون و همکاران، ۱۹۹۳) و بسته جیزاپلاس‌پلاس[3] (اوچ، ۲۰۰۰ و اوچ و نی، ۲۰۰۳) تولید شده طراحی شده‌اند. اما این پژوهش روشی نسبتاً ابتکاری و نوین برای استخراج تناظرها در سطح واژه از پیکره موازی همترازشده در سطح جمله و با استفاده از آماره اطلاعات متقابل ارائه می‌نماید. در بخش‌های زیر پس از مرور تحقیقات انجام‌گرفته در این حوزه، پیکره دوزبانه‌ای که برای این پژوهش مورداستفاده قرارگرفته معرفی‌می‌شود و متعاقب آن روش تحقیق بکاررفته برای انجام آزمایشی در رابطه با ارزیابی کیفی این برنامه مطابقه همراه با ارائه الگوریتم آن مورد بحث قرارخواهدگرفت. آزمایش و نتایج حاصل از آن در بخش ۴ و ۵ گزارش‌خواهندشد. این مقاله با ملاحظات جمع‌بندی و بحث در مورد کارهای آتی به پایان خواهد رسید.

[1] Machine readable corpora
[2] aligning
[3] GIZ/A++ Toolkit
[4] Mutual information

۲. مروری بر پژوهش‌های پیشین

تاکنون تحقیقات زیادی در زمینه پیکره‌های موازی صورت گرفته‌است. تلاش تمام این تحقیقات به سمت استفاده از کاربردهای چنین پیکره‌هایی به منظور حل مسائل مختلف زبانشناسی رایانشی مانند ترجمه ماشینی مبتنی بر آمار (براون و همکاران، ۱۹۹۳)، بازیابی اطلاعات دوزبانه (نظارات و موسوی میانگاه، ۱۳۹۰)، یادگیری زبان (وو و همکاران، ۲۰۰۳)، ترجمهٔ انسانی (موسوی میانگاه، ۲۰۰۶) و نظیر آن بوده است. برخی از محققان سعی کرده‌اند تا روش‌های جدیدی برای همترازسازی پیکره‌های دوزبانه در سطح جمله بیابند (سیمارد و پلامندن، ۱۹۹۸ و اوچ و نی، ۲۰۰۱). برخی دیگر از آنها به سمت استفاده از روش‌های آماری (گیل و چرچ، ۱۹۹۱؛ کاپیک، ۱۹۹۳؛ داگان و همکاران، ۱۹۹۳؛ اینو و ناگایتو، ۱۹۹۳ و فانگ، ۱۹۹۵) یا زبانی (یاماماتو و ساکاماتو، ۱۹۹۳؛ کومانو و هیراکاوا، ۱۹۹۴ و ایشیماتو و ناگااو، ۱۹۹۴) برای همترازسازی پیکره‌های دوزبانه در سطح واژه روی‌آورده‌اند. روش‌های آماری از فراوانی واژه‌ها استفاده می‌کنند و میزان ارتباط بین واژه‌های متناظر در دوزبان موجود در پیکره موازی را محاسبه می‌نمایند. این در حالیست که روش‌های زبانی تناظرهای واژگانی در دو زبان را با استفاده از یک واژه‌نامه دوزبانه پیدا می‌کنند.

ونگ لکسم یک برنامه مطابقه موازی انگلیسی _ چینی طراحی نموده است و کاربردهای آموزشی آن در زبان‌های انگلیسی و چینی را از طریق مثال‌هایی از آزمایش‌های آموزش و یادگیری نشان داده است، و بدین ترتیب روش یادگیری داده‌محور را پیاده‌سازی نموده است. درواقع، در پژوهش او، این ارزش آموزشی مطابقه موازی است که مورد توجه قرارگرفته و نه روش یافتن تناظرهای واژگانی در پیکره موازی (ونگ، ۲۰۰۱).

همینطور یک سیستم مطابقه انگلیسی–چینی مبتنی بر شبکه بنام توتال‌ریکال[1] به منظور افزایش استفاده مجدد از ترجمه‌ها و نیز تشویق به استفاده از منابع موثق و طبیعی در نگارش زبان دوم توسط وو و همکارانش طراحی‌شده‌است. توتال‌ریکال پیشرفته‌تر از مطابقه‌های قبلی است چرا که در این سیستم کاربر نه تنها می‌تواند تمام سوابق مربوط به واحد واژگانی مورد جستجو را ببیند بلکه می‌تواند معادل‌های ترجمانی آن را به صورت برجسته مشاهده نماید. این سیستم در حقیقت نوعی مطابقه موازی همترازشده در سطح واژه است که نسبت به مطابقه‌های قبلی کاربردهای بسیار بیشتری دارد (وو و همکاران، ۲۰۰۳).

پاراکانکورد مطابقهٔ موازی و چندزبانه‌ای است که چهار متن موازی از چهار زبان متفاوت، یا یک متن اصلی و سه ترجمهٔ متفاوت از سه زبان دیگر را می‌پذیرد. این برنامه دربرگیرنده نوار ابزاری برای برجسته‌سازی ترجمه‌های کاندید شده و نیز یک بخش خودکار بنام هات وردز است که از اطلاعات بسامدی استفاده می‌کند تا اطلاعاتی در مورد ترجمه‌های ممکن جستار موردنظر فراهم نماید (بارلو، ۲۰۰۲).

موزر و همکارانش یک سیستم همترازسازی در سطح واژه برای ترجمهٔ ماشینی آماری ارائه‌دادند که بطور همزمان جملهٔ متن مقصد را طوری مجدداً مرتب‌سازی می‌کرد که با ترتیب واژه‌ها در جمله متن مقصد متناظر خود منطبق باشد. ایده اصلی کار آن‌ها تولید یک الگوریتم یکنواخت بین جملهٔ مقصد و یک شبکه‌بندی تبدیلی بود که نشان‌دهندهٔ مرتب‌سازی‌های مجدد متفاوت واژه‌ها در جملهٔ مبدأ باشد. آن‌ها نشان دادند که سیستم‌های ترجمه‌ای که با روش

[1] TotalRecall
[2] hot words
[3] monotonic

پیشنهادی آن‌ها کار می‌کنند بهتر و یا حداقل در همان حد سیستم‌هایی که با بستهٔ نرم افزاری جیزاپلاس‌پلاس کارمی‌کنند عمل می‌کنند (مازر و همکاران، ۲۰۰۶).

۳. پیکرهٔ موازی انگلیسی-فارسی

پیکرهٔ موازی انگلیسی-فارسی در ابتدا به صورت بانک داده‌ای متنی متشکل از متون اصلی به زبان انگلیسی و ترجمه‌های آن‌ها به زبان فارسی و نیز متون اصلی به زبان فارسی و ترجمه‌های آن‌ها به زبان انگلیسی گردآوری شد. اگرچه میزان دسترسی به متون دوزبانه که شامل زبان فارسی باشند به دلیل تراکم پایین متون فارسی در سراسر دنیا و وجودنداشتن چنین متونی در برخی گونه‌ها و زمینه‌های خاص بسیار پایین می‌باشد، ما موفق به تهیهٔ بانک داده‌ای دوزبانهٔ نسبتاً حجیمی متشکل از صدهزار جمله انگلیسی و فارسی شدیم. متون در این پیکره به انواع زیر طبقه‌بندی شده‌اند: مذهبی، ادبی، سیاسی، اقتصادی، علمی، شعر، اصطلاحات و ضرب‌المثلها، ورزشی، متفرقه، پزشکی، و فرهنگی. نوع متن با جستجوی هر جمله درکنار آن ظاهر می‌شود. پایگاه داده‌ای این پیکره به‌صورت اکسس[1] و اس کیو ال[2] موجود می‌باشد و مطابقه مبتنی بر آن نیز به‌صورت تحت ویندوز و تحت وب تهیه‌شده است. پیکرهٔ تولیدشده پیکره‌ای درحال پیشرفت است، بدین معنا که پیکره‌ایست باز که برحسب نیاز و با گذشت زمان مطالبی به آن اضافه خواهد شد (موسوی میانگاه، ۲۰۰۹).

آماده‌سازی و همترازسازی پیکره

متون خامی که از منابع گوناگون استخراج می‌شوند برای ورود به پیکره باید پیش‌پردازش شوند. دانلودکردن، تبدیل فرمت و هنجارسازی (عادی‌سازی) متون از جمله مراحل بسیار وقت‌گیر در آماده‌سازی پیکره به شمار می‌رود. برخی از صفحات که نامربوط هستند و نیز تمام شکل‌ها، جداول و عکس‌ها قبل از اینکه وارد پیکره شوند باید حذف‌گردند. در برخی موارد که یک جمله یا قسمتی از آن ترجمه نشده باشد، قسمت‌های ترجمه نشده باید حذف شوند. بعد از بازبینی و تأیید، تمام متون به طور هماهنگ به فرمت ایکس ام ال[3] رمزگذاری می‌شوند تا پیکره بتواند مستقل از کاربرد بوده و از طریق اینترنت به راحتی قابل مبادله باشد.

به منظور استخراج اطلاعات از پیکرهٔ موازی، لازم است متون دوزبانهٔ این پیکره در مرحلهٔ نخست در سطوح پاراگراف، جمله و واژه همتراز شوند. منظور از همترازسازی ایجاد وابستگی بین قطعاتی از متن در یک زبان و ترجمه‌ها یا متون معادل آن‌ها در زبان دیگر است. همترازسازی در سطح پاراگراف کار نسبتاً ساده‌ای است، چون مرزهای پاراگرافی معمولاً مشخص هستند، درضمن اینکه این نوع همترازسازی برای استفاده‌های بعدی از پیکره‌های موازی در امر پژوهش خیلی هم مفید به نظر نمی‌رسند. همترازسازی در سطح واژه به معنای تعیین‌کردن جفت‌واژه‌های متناظر در دو زبان می باشد. این نوع همترازسازی کاری بسیار دشوار بوده و نیاز به الگوریتم های پیچیده‌ای دارد. بنابراین اکثر پیکره‌های موازی در سطح جمله همتراز می‌شوند.

[1] Access
[2] SQL
[3] XML (Extendible Mark-up Language)

در پیکرهٔ موازی انگلیسی- فارسی طراحی‌شده توسط نویسنده، همترازسازی در سطح جملات گرچه می‌توانست با روش‌های کاملاً خودکار یا نیمه‌خودکار انجام شود (رنسیک، ۱۹۹۸، ۱۹۹۹)، لیکن تماماً به طور دستی انجام شده است تا صحت همترازسازی به ۱۰۰ درصد برسد. قصد ما این است که این پیکره در حل مسائلی مانند پردازش خودکار متن و ترجمهٔ ماشینی آماری و ساختن نرم افزار حافظهٔ ترجمه که در آن‌ها دقت بالا بسیار مهم می باشد مورداستفاده قرارگیرد.

پیکره‌ی موازی انگلیسی ° فارسی فوق در انجمن منابع زبانی اروپا به ثبت رسیده و دستیابی به آن از طریق مراجعه به کاتالوگ این انجمن که آدرس آن در زیر آورده می‌شود امکان‌پذیر می‌باشد:

http://catalog.elra.info/product_info.php?products_id=1111

ساخت یک پیکره به‌خودی‌خود هدف محسوب نمی‌شود بلکه معمولاً به عنوان قسمتی از یک پروژه تحقیقاتی درنظرگرفته می‌شود. درواقع هنگامی که ساخت پیکره به پایان می‌رسد کار اصلی شروع می‌شود.

٤. الگوریتم همترازسازی در سطح واژه

پس از این‌که همترازسازی پیکره مورد نظر در سطح جمله صورت‌گرفت، مرحلهٔ بعدی تشخیص تناظرهای واژگانی در جفت جملات متناظر است. به‌عبارت دیگر، هدف اصلی در اینجا این است که مشخص شود کدام واژه در جمله انگلیسی با کدام واژه در جمله فارسی متناظر آن مطابقت دارد. این دو واژه معادل در جفت جملات متناظر برجسته یا پررنگ می‌شوند.

از آنجاکه میان طول یک جستار و سرعت پاسخ‌دهی رابطهٔ عکس برقرار است، یعنی هرچقدر یک جستار تعداد واژه‌های بیشتری را دربرگیرد احتمال تهی بودن نتیجه بیشتر است، در این روش عمدتاً یک واژه به عنوان جستار درنظر گرفته شده است. همچنین، از آنجایی که واحد اصلی جستار در امور مربوط به نمایه از قبیل بازیابی اطلاعات در بیشتر موارد اسم‌های مجزا هستند، تمام واژه‌های قاموسی زبان فارسی باید با تمام واژه‌های محتوایی زبان انگلیسی مقایسه شوند تا از محاسبات اضافی و اختلال ممانعت به‌عمل آید. تشخیص واژه‌های محتوایی از واژه‌های دستوری ازطریق برچسب‌زن اجزای کلام که با استفاده از سیستم برچسب‌زن کلاز[3] درمورد واژه‌های انگلیسی اعمال می‌شود براحتی امکان‌پذیر است. درمورد زبان فارسی نیز از سیستم برچسب زن اجزای کلام فارسی‌تگ[4] استفاده می‌شود (موسوی میانگاه، ۲۰۱۵). از آنجا که واژه‌های دستوری در هر زبانی مقولات بسته‌ای را تشکیل می‌دهند با خارج‌کردن این مقولات از واژگان، بقیه واژه‌ها محتوایی درنظرگرفته شده‌اند. درمواردی که بین این دو گروه واژه‌های محتوایی و دستوری هم‌پوشانی وجود داشته نیز سیستم‌های برچسب زن مذکور که عمدتاً سیستم‌های مبتنی بر قاعده هستند تاحد زیادی قادر به تشخیص آن‌ها بوده‌اند.

اکنون زمان آن رسیده تا مطابقه موازی در سطح جمله به شکل یک بسته نرم‌افزاری ساخته شود تا کاربر قادر باشد به وسیله آن واژهٔ خاصی را در زبان انگلیسی یا فارسی جستجوکرده و فهرستی از تمام جملات به زبان موردجستجو شامل

[1] content words
[2] function words
[3] CLAWS POS tagger
[4] FarsiTag

آن واژه خاص همراه با جملات متناظر به زبان دیگر را دریافت نماید و این در حالی است که جستار موردنظر و معادل آن به زبان دیگر به طور برجسته نشان داده‌شده است. به عبارت دیگر، جستجو می‌تواند براساس هرکدام از زبان‌های موجود در پیکره انجام شود.

اگر دو جملهٔ انگلیسی و فارسی ترجمهٔ یکدیگر باشند، انتظار می‌رود که یک واژهٔ فرضی در جملهٔ فارسی دارای یک ترجمهٔ انگلیسی در جملهٔ انگلیسی متناظر خود داشته باشد. جدول (۱) پیوندهای مناسب در دو جملهٔ متناظر انگلیسی و فارسی را که توسط مترجم انسانی همتراز شده نشان می‌دهد.

For true colonialism to exist two conditions are necessary.
۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸

برای وجود استعمار حقیقی دو شرط لازم است.
۸ ۷ ۶ ۵ ۴ ۳ ۲ ۱

جدول ۱) همترازسازی در سطح واژه در یک نمونهٔ ساده از جفت جملهٔ انگلیسی و فارسی

English→ Persian
1→1
2→۴
3→3
4→۲
5→۵
6→۶
7→۸
8→۷

لازم به ذکر است که زبان فارسی زبانی هسته آغازین است که از راست به چپ نوشته می‌شود. در دو جملهٔ همتراز انگلیسی °فارسی، یافتن ترجمه مناسب یک واژه فرضی فارسی در میان واژه‌های انگلیسی جمله معادل آن برای یک جستجوگر غیرانسانی تقریباً غیرممکن است، چرا که ترتیب عناصر در جملات انگلیسی و فارسی با هم تطابق ندارد. از این‌رو، ما روشی نسبتاً بدیع برای یافتن مناسب‌ترین و محتمل‌ترین معادل هر واژه با استفاده از آماره اطلاعات متقابل ارائه داده‌ایم. اطلاعات متقابل اساساً برای محاسبهٔ میزان همبستگی میان واژه‌ها با استفاده از آمار هم‌رخدادی واژه‌ها بکار برده می‌شود و به‌صورت فرمول شماره (۱) قابل تعریف است (چرچ و هنکس، ۱۹۹۰):

(۱)

$$MI(x,y) = Log_2 \frac{P(x,y)}{P(x)P(y)} = Log_2 \frac{Nf_w(x,y)}{f(x)f(y)}$$

در اینجا x و y واژه‌های فرضی در بافت هستند. احتمال‌های $p(x)$ و $p(y)$ با استفاده از شمارش تعداد رخدادهای x و y در پیکره یعنی $f(x)$ و $f(y)$ محاسبه می‌شوند. N حجم پیکره را نشان می‌دهد. $P(x, y)$ با شمارش تعداد دفعاتی که x و y در بافت یکسانی ظاهر می‌شوند (در اینجا منظور از بافت همان رکورد است) محاسبه می‌شود. از آنجا که الگوریتم پیاده‌سازی شده برای این برنامه جمله‌ها را شناسایی نموده و واژه‌های موجود در یک جمله که جستار به آن تعلق دارد را به عنوان بافت زبانی آن جستار در نظر می‌گیرد، در این تحقیق هر رکورد شامل یک جملهٔ انگلیسی و معادل فارسی آن است.

استفاده از مقادیر اطلاعات متقابل براساس این فرض است که هنگامی که دو واژهٔ انگلیسی و فارسی در محدودهٔ بافت معینی (که در اینجا یک رکورد است) با فراوانی بالایی هم‌رخداد می‌شوند، احتمال این که آن‌ها ترجمه یکدیگر باشند بیشتر می‌شود. درحقیقت، برای یافتن معادل فارسی هر واژهٔ محتوایی در زبان انگلیسی این برنامه جملاتی را در پیکره جستجو می‌کند که در آن‌ها واژهٔ انگلیسی مورد نظر وجود داشته باشد. سپس برنامه مقدار اطلاعات متقابل آن واژهٔ انگلیسی با تمام واژه‌های محتوایی زبان فارسی که در جملات فارسی متناظر وجود دارند را محاسبه می‌نماید. به‌عنوان مثال، جملهٔ همتراز شده زیر را که در آن ترجمه واژه انگلیسی که زیر آن خط کشیده‌شده باید از جمله فارسی متناظر آن استخراج شود را درنظر بگیرید:

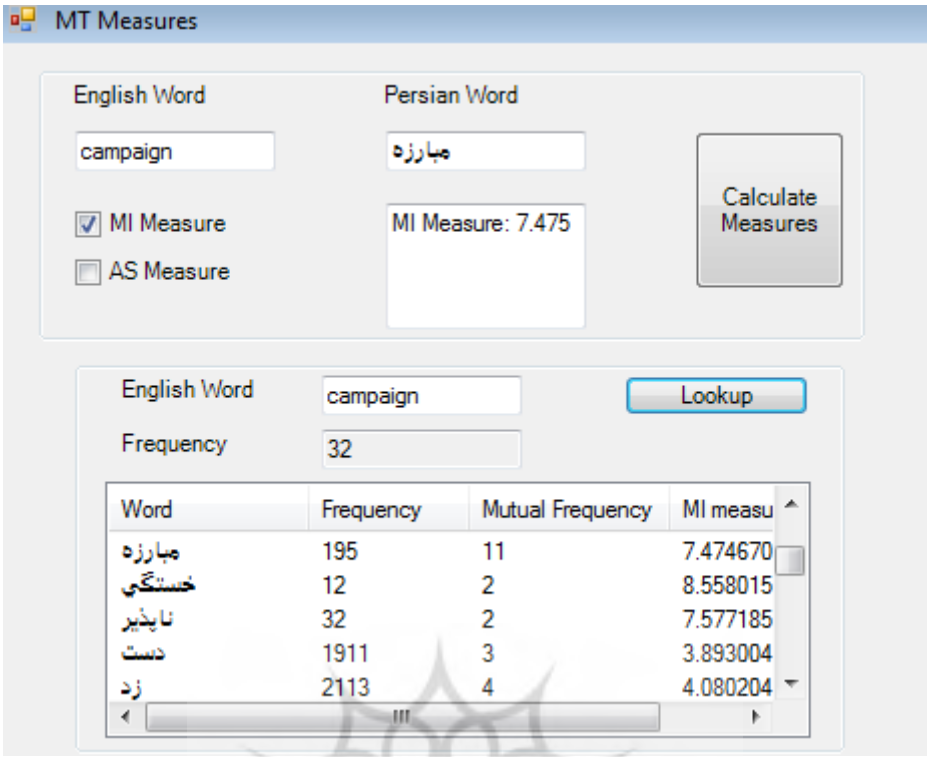
جدول ۲) نمایش یک رکورد تولیدشده توسط پیکره برای جستار "campaign"

English sentence	Persian sentence
UN has launched a global campaign against the poverty.	سازمان ملل مبارزه جهانی علیه فقر را آغاز کرده است.

در جدول (۲) مقدار اطلاعات متقابل واژه *campaign* و تمام واژه‌های محتوایی فارسی در طرف راست محاسبه شده‌اند و جفت واژه‌هایی که همبستگی زیادتری با یکدیگر دارند بعنوان ترجمه‌های یکدیگر انتخاب می‌شوند و بدین ترتیب ترجمه‌هایی که احتمال درست بودنشان کمتر است حذف می‌شوند. نتایج این محاسبات را می‌توان در جدول شمارهٔ (۳) مشاهده نمود. روش اجرای این برنامه برای واژه *campaign* در محیط نرم‌افزاری در شکل ۲ نمایش داده شده است.

جدول ۳) مقدار اطلاعات متقابل واژه "*campaign*" و ۸ واژه در جمله فارسی متناظر

Word x	Word y	f(x)	f(y)	f(x, y)	MI
campaign	سازمان	۳۲	۱۱۳۰	۴	۴/۷۰۶
campaign	ملل	۳۲	۱۶۰۴	۷	۴/۹۱۵
campaign	**مبارزه**	۳۲	**۱۹۵**	**۱۱**	**۷/٤٧٤٦**
campaign	جهانی	۳۲	۳۱۰	2	۴/۴۲۸
campaign	علیه	۳۲	۱۵۹	۴	۵/۴۹
campaign	فقر	۳۲	۱۷۳	۲	۵/۸۹
campaign	آغاز	۳۲	۱۷۷	۲	۵/۸۶۶
campaign	کرده است	۳۲	۴۶۳۵	۵	۵/۹۴۳



شکل ۲) روش اجرای برنامه برای واژه campaign در محیط نرم افزاری

از جدول شمارهٔ (۳) می‌توان دریافت که برای این جمله بیشترین مقدار اطلاعات متقابل به جفت "campaign" و "مبارزه" با مقدار ۷/۴۷۴۶ متعلق می‌باشد و مقدار اطلاعات متقابل دیگر جفت واژه‌ها بسیار پایین‌تر از آن است. بدین ترتیب، این الگوریتم قادر خواهد بود مناسب‌ترین ترجمه یک واژه انگلیسی را در جمله فارسی متناظر آن برجسته نماید و با این کار یک مطابقه مستقل از زبان کاملاً مبتنی بر آمار تولید شده است. البته جستار می‌تواند هم به زبان انگلیسی و هم به زبان فارسی باشد. تعداد محاسبات موردنیاز برای یافتن معادل یک واژهٔ خاص در یک جمله به یک زبان به تعداد واژه‌های محتوایی در جمله متناظر آن به زبان دیگر بستگی دارد.

۵. نتایج آزمایش

به منظور ارزیابی میزان تأثیر این الگوریتم، آزمایشی بر روی پیکره موازی موجود ترتیب داده شد. در این آزمایش تنها واژه‌های محتوایی درنظرگرفته شدند چراکه واژه‌های اصلی که توسط کاربران در تقریباً تمام مطابقه‌ها موردجستجو قرار می‌گیرند همین نوع واژه‌ها هستند. هرچند واژه‌های دستوری نیز به انسجام متن کمک می‌کنند، اما همانطور که می‌دانیم در موتورهای جستجو و به‌طور کلی در بازیابی اطلاعات چنین واژه‌هایی قابل ملاحظه نیستند و اکثر قریب به اتفاق جستارها در زبان‌های مختلف عمدتاً عباراتی متشکل از واژه‌های محتوایی مانند اسم و صفت هستند.

پیکره آزمون استفاده‌شده برای ارزیابی عملکرد این آزمایش مبتنی بر الگوی پیشنهادی شامل مجموعه‌ای از ۱۰۰ واژه انگلیسی در نقش جستار است که به‌عنوان درونداد به الگوریتم داده می‌شود و برونداد سیستم مطابقه در سطح جمله است که در آن کاربر قادر است مجموعه‌ای از جفت جملات به انگلیسی و فارسی را همراه با جستارهای مورد نظر و

ترجمه‌هایشان که برجسته نمایش داده می‌شوند را مشاهده نماید. با هر واژهٔ انگلیسی در پیکرهٔ آزمون فهرستی از ترجمه‌های ممکن ارائه می‌شود و انتخاب محتمل‌ترین آن‌ها هدف نهایی سیستم است.

در سیستم‌های بازیابی اطلاعات، معیار دقت و بازخوانی و معیارهایی شبیه به آن‌ها به عنوان معیارهای اصلی ارزیابی به‌کار می‌روند: معیار دقت به حاصل تقسیم «تعداد مستندات بازیابی شده واقعاً با ربط» بر «تعداد کل مستندات بازیابی‌شده» گفته می‌شود. *معیار بازخوانی* به حاصل تقسیم «تعداد مستندات بازیابی شده با مرتبط بر تعداد تعداد مستندات مرتبط موجود در مجموعهٔ اطلاعاتی » گفته می‌شود. درواقع دقت درصد متون بازیابی‌شدهٔ مرتبط و فراخوانی درصد ارتباط متون بازیابی شده را نشان می‌دهند.

در این آزمایش عملکرد برنامه‌ای که سعی دارد واحدهای واژگانی زبان انگلیسی را با استفاده از پیکره ترجمه نماید براساس دونوع معیار ارزشیابی، یعنی دقت و فراخوانی، طبق فرمول‌های (۲) و (۳) موردارزیابی قرارگرفت.

(۲)

$$\text{Precision} = \frac{\text{Number of correctly translated strings}}{\text{Total number of strings translated by program}}$$

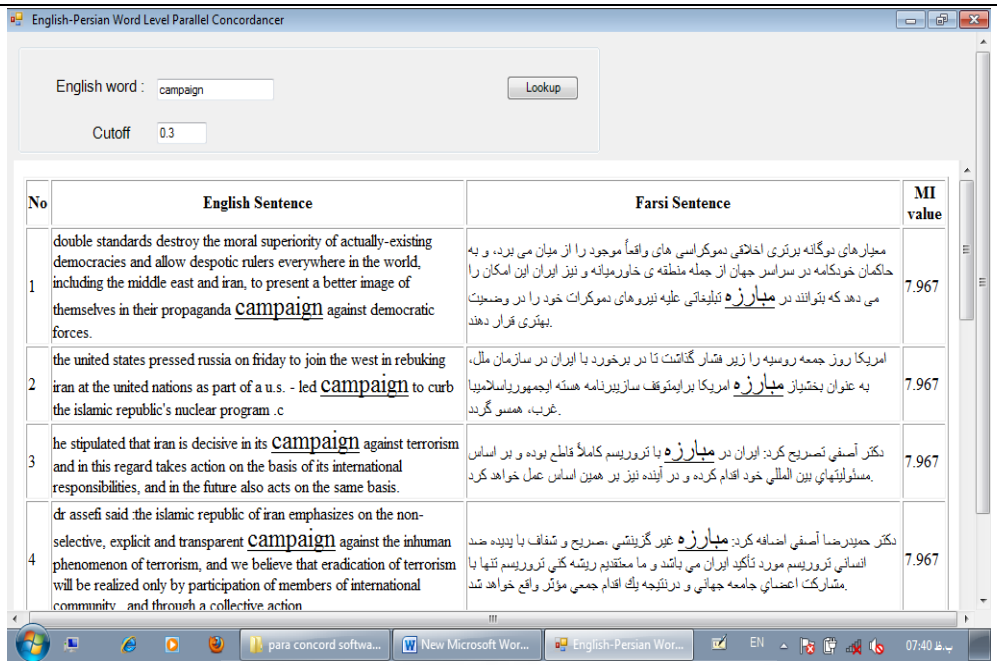
(۳)

$$\text{Recall} = \frac{\text{Number of correctly translated strings}}{\text{Total number of strings in the test set}}$$

دقت میزان دقیق بودن یا درستی را نشان می دهد، درحالی که فراخوانی میزان کامل بودن را نشان می‌دهد. از آنجایی که در این آزمایش پاسخ‌های تولیدشده (تعداد زنجیره‌های تولیدشده توسط برنامه) با کل پاسخ‌های موردانتظار (تعداد زنجیره‌ها در پیکرهٔ آزمون) یکسان هستند، دقت و فراخوانی یکی بوده و عملکرد برنامه با صحت اندازه‌گیری می‌شود:

$$\text{Accuracy} = \frac{\text{Number of correct outputs proposed by program}}{\text{Total number of English queries in the test set}} = \frac{75}{100} = 75\%$$

نتایج به دست آمده از معادل‌یابی خودکار با همان مجموعه از واژه‌ها که از جفت جملات انگلیسی و فارسی در پیکره به‌صورت دستی در سطح واژه همتراز شده بودند مقایسه شد. نتایج آزمایش نشان داد که برنامه مطابقه ما صحتی معادل ۷۵ درصد بدست آورده که بسیار امیدوار کننده است. طبیعی است در حالتی که این الگوریتم بر روی یک پیکره طبیعی از زبان انگلیسی که در آن انواع مختلف واژه‌ها (هم واژه‌های محتوایی و هم واژه‌های دستوری) یافت می‌شوند اعمال شود، صحت این روش مسلماً تا حد زیادی افزایش خواهدیافت. شکل (۳) نمونه‌ای از یک مطابقه موازی در سطح جمله که شامل ۵ جفت جمله با جستار انگلیسی *campaign*، معادل‌های فارسی آن و مقدار اطلاعات متقابل آن‌ها (در ستون چهارم) است را نشان می‌دهد



شکل ۳) نمونه‌ای از یک مطابقه موازی در سطح جمله

٦. نتیجه‌گیری و چشم‌انداز آینده

بی‌شک مطابقه‌های دوزبانه منابع ارزشمندی برای زبانشناسان، مترجمان و نیز کاربران رایانه در حوزهٔ پردازش متن و بویژه بازیابی اطلاعات هستند. این مطابقه‌ها از مجموعه حجیمی از جملات همترازشده در سطح جمله به‌عنوان پایگاه داده‌ای قابل‌جستجو استفاده می‌کنند. کاربر می‌تواند یک واحد زبانشناختی یا زنجیره‌ای از واحدها در یک زبان را جستجو کند و مطابقه می تواند تمام جملات به آن زبان که در آن‌ها واحد (های) مورد جستجو ظاهر شده‌اند همراه با جملات متناظر آن‌ها به زبان دیگر را نمایش می‌دهد.

در این مقاله تلاش بر این بود تا روشی نسبتاً نوین برای همترازسازی خودکار مجموعه‌ای از جملات موازی با استفاده از رویکردی آماری بنام آمارهٔ اطلاعات متقابل ارائه شود. برای این منظور یک پیکرهٔ موازی انگلیسی ° فارسی که به‌طور دستی در سطح جمله همترازسازی شده است مورد استفاده قرارگرفت. مقدار اطلاعات متقابل نشان‌دهندهٔ درجه همبستگی معنایی بین واژه‌ها است.

همترازسازی در سطح واژه مطابقه‌های دوزبانه را غنی‌تر می‌سازد و کاربردهای گوناگونی از جمله ترجمهٔ ماشینی آماری، بازیابی اطلاعات دوزبانه، یادگیری زبان، واژه‌نگاری رایانه‌ای و مانند آن دارد. روش ارائه‌شده در این تحقیق روشی مستقل از زبان است که صرفاً تکیه بر مقیاس‌های آماری دارد. از این‌رو، الگوریتم این روش می‌تواند برای هر جفت زبان دیگری که پیکره موازی آن‌ها موجود باشد بکار رود.

هنوز هم اصلاحات زیاد دیگری به این سیستم اضافه شود تا به تولید بروندادهای بهتری منجر شود. یکی این‌که نویسندگان قصد دارند پایگاه داده‌ای پیکره موجود را با اطلاعات جدید بیشتری بطور مداوم بروز رسانی نمایند بطوری که مطابقه موازی موجود که در سطح جمله قادر به معادل‌یابی است به مطابقه قدرتمندتری تبدیل شود. به‌علاوه، تفکیک حوزه‌های موضوعی نیز می‌تواند به بازدهی بیشتر و مؤثرتری منجر شود.

قدردانی

این پژوهش بااستفاده از اعتبارات دانشگاه پیام نور در قالب طرح پژوهشی انجام شده است.

منابع

نظارات، امین و موسوی میانگاه، طیبه (۱۳۹۰). طراحی و پیاده‌سازی یک سامانه بازیابی اطلاعات دوزبانه با استفاده از پیکره‌های زبانی. *پردازش و مدیریت اطلاعات، ویژه نامه ذخیره، بازیابی و مدیریت اطلاعات*: ۲۱۲-۱۹۷.

Barlow, M. (2002). ParaConc: Concordance software for multilingual parallel corpora. In: *Proceedings of the Third International Conference on Language Resources and Evaluation. Workshop on Language Resources in Translation Work and Research.*, Las Palmas, Spain, pp. 20-24.

Brown, P. F., Della Pietra, V. S. A.; Della Pietra, V. J.; and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19 (2), 263˚312.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1) , 22-29.

Fung, P. (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *Proc. of the 33rd Annual Meeting of the ACL*, 236-243.

Gale, .. A. and K. .. Church, (9991). IIdentifying word correspondences in parallel texts *Proceedings of the 4th DARPA Speech and Natural Language Workshop*::152-157, Pacific Grove, CA.

Dagan, I. Kenneth W. Ch, and William A. G. (1993). Robust bilingual word alignment for machine-aided translation. In: *Proc. of the Workshop on Very Large Corpora*: Academic and Industrial Perspectives.

Inoue, N. and Nogaito, I. (1993). Automatic construction of the Japanese-English dictionary from bilingual text. *Technical Report of IEICE, NLC*: 39-93.

lshimoto, H and Nagao, M. (1994). Automatic construction of a bilingual dictionary of technical terms from parallel texts. *Technical Report of IPSJ, NL*: 102-11.

Johns, T. (1986). Microconcord: A language-learner's research tool. *System* 14(2), 151-162.

Johns, T. (1998). Multiconcord*:* the lingua multilingual parallel concordancer for windows. Available on: http://web.bham.ac.uk/johnstf/l_text.htm. Accessed Feb 03

Kaji, H and Aizoni, T. (1996). Extracting word correspondences from bilingual corpora based on word co-occurrences information. *The 16th International Conference on Computational Linguistics,* pp. 23 ˚ 28. Copenhagen, Denmark

Kumano, A. and Hirakawa, H. (1994). Building an MT dictionary from parallel texts based on linguistic and statistical information. *Proc. of COLING'94*: 76-81.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. *Proceedings of the 31st Annual Meeting of the ACL:* 17-22, Columbus, Ohio.

Mauser, A., Matusov, E and Ney, H. (2006). Training a Statistical Machine Translation System without GIZA++. *International Conference on Language Resources and Evaluation (LREC):* 715-720, Genoa, Italy.

Mosavi Miangah, T. (2006). Applications of corpora in translation. *Translation Studies* 12, 43-56.

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

Mosavi Miangah (in Press). FarsiTag: A part of speech tagging system for Persian. Journal of Quantitative Linguistics.

Och, Franz J. (2000). Giza++: Training of statistical translation models. Available at: http://www-i6.informatik.rwthaachen.de/~och/software /GIZA++.html.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1), 1-19.

Och, F. and Ney, H. (2001). Improved Statistical Alignment Models, *Proceedings of ACL* 2001.

Rensik, .. (9998). PParallel strands: A preliminary investigation into mining the web for bilingual text.. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529:* 28-31, Langhorne, PA, October,.

Rensik, .. (9999): Mining the web for bilingual text . *Proc. of 37th Meeting of the ACL.* Maryland: 527-534.

Scott, M. (2000). *WordSmith Tools Version 3.0* [Computer software]. Oxford: Oxford University Press.

Simard, M. and Plamondon, P. (1998). Bilingual sentence alignment: balancing robustness and accuracy. *Machine Translation* 13, 59˚80.

Wang L. (2001). Exploring parallel concordancing in English and Chinese. *Language Learning & Technology* 5(3), 174-184.

Wu, J. C.; Yeh, K.; Chuang, C.; Thomas C., Shei, W. C. and Chang, J. (2003). TotalRecall: A bilingual concordance for computer assisted translation and language learning. *Association for Computational Linguistics*: 201-204.

Yamamoto, Y. and Sakamoto, M. (1993). Extraction of technical term bilingual dictionary from bilingual corpus. *Technical Report of IPSJ, NL*: 12-94. (in Japanese).