استخراج جملات موازی از دادگان وب

نسرین براتعلی‌پور[۱]
هشام فیلی[۲]
آزاده شاکری[۳]

چکیده

پیکره‌های موازی یکی از منابع با ارزش در بسیاری از کاربردهای پردازش زبان طبیعی و همچنین بازیابی هوشمند اطلاعات بین‌زبانی است. لازمه استفاده از این پیکره‌ها هم‌ترازی آنها در سطح جمله است، اما جمع‌آوری و یا تولید این پیکره‌ها و همچنین هم‌ترازی آنها بسیار پرهزینه است. با توجه به گستردگی و قابلیت دسترسی رایگان صفحات وب دوزبانه، جمع‌آوری پیکره‌های موازی از وب و هم‌ترازی آنها به صورت خودکار بسیار مطلوب است. در این مقاله برای تولید جملات موازی، ابتدا صفحات وب حاوی جملات موازی انتخاب، سپس ویژگی‌های هر زوج جمله فارسی–انگلیسی در این صفحات محاسبه و در نهایت به کمک طبقه‌بند بیشترین پراکندگی[۴] جملات موازی استخراج می‌شود. یکی از ویژگی‌های جملات استخراج شده، وابسته نبودن به دامنه و امکان پوشش حوزه‌های متفاوت معنایی است.

واژه‌های کلیدی: پیکره موازی، هم‌ترازی متون، داده‌کاوی وب

۱- مقدمه

امروزه پیکره‌ها یکی از منابع ضروری در پردازش زبان طبیعی و بازیابی اطلاعات بین زبانی به شمار می‌آیند. از جمله این پیکره‌ها می‌توان به پیکره‌های موازی و تطبیقی اشاره کرد. پیکره موازی مجموعه بزرگی از اسنادی است که ترجمه یکدیگراند. پیکره تطبیقی مجموعه اسنادی است که ترجمه یکدیگر نیستند اما شامل متونی با مفاهیم مشترکی‌اند.

با توجه به دانش غنی زبانی در پیکره‌های موازی، امروزه از این منابع در بسیاری از کاربردهای پردازش زبان طبیعی و بازیابی اطلاعات بین زبانی استفاده می‌شود. از جمله مهم‌ترین آنها، استفاده در ترجمه ماشینی و به طور خاص می‌توان به ماشین ترجمه آماری اشاره کرد. همچنین واژه‌نگاری

---

[۱]- دانشجوی کارشناسی ارشد، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران
Nasrinbaratali@ut.ac.ir
[۲]- استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران hfaili@ut.ac.ir
[۳]- استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران shakery@ut.ac.ir

[4]- Maximum Entropy Classifier

چندزبانه[۱]، ابهام‌زدایی معانی کلمات[۲]، استخراج اصطلاحات[۳] و یادگیری زبان به کمک کامپیوتر[۴] از کاربردهای دیگر این منابع است(تایدمن[۵]،۲۰۱۱: ۵).

جمله کوچک‌ترین واحد معنادار متن است و لذا برای استفاده از پیکره‌های موازی در بسیاری از کاربردها، به هم‌ترازی در سطح جمله پرداخته می‌شود. جملات موازی دربرگیرنده اطلاعات مفید نحوی و معنایی است اما به این دلیل که جملات توسط مترجم در طول فرایند ترجمه شکسته، ترکیب، حذف، اضافه یا جابه‌جا می‌شوند، هم‌ترازی آنها دشوار می‌شود.

در این پژوهش ابتدا صفحات حاوی جملات موازی انتخاب و سپس از میان دادگان استخراج شده از وب، جملات موازی استخراج می‌شود. برای این منظور، پژوهش به دو بخش کلی استخراج دادگان تطبیقی از وب(صفحاتی حاوی جملات موازی) و استخراج جملات موازی تقسیم می‌شود.

برای جمع‌آوری اسناد حاوی جملات موازی، از وب استفاده شد، زیرا وب به عنوان منبع وسیعی از اسناد دوزبانه محسوب می‌شود، علاوه بر آن اسنادی که در وب موجود است به دامنه خاصی[۶] محدود نمی‌شوند. از این رو می‌توان جملات موازی را در حوزه‌های متفاوت بررسی کرد. در این بخش برای هم‌ترازی اسناد، مستقل از محتوای اسناد، از اطلاعات افزوده[۷] صفحات وب استفاده شد. راهکار مورد نظر به تفصیل در بخش هم‌ترازی اسناد بیان می‌شود.

در هم‌ترازی جملات، یکی از روش‌های مطرح استفاده از طبقه‌بندها است، به طوری که با دریافت دو جمله مشخص می‌شود که آیا این دو جمله ترجمه یکدیگر است یا خیر. هر طبقه‌بند به ویژگی‌هایی نیاز دارد تا بر اساس آن طبقه‌بندی نماید. شناسایی این ویژگی‌ها و استفاده از آن در هم‌ترازی جملات بسیار مهم است. در این پژوهش علاوه بر پیاده‌سازی ویژگی‌های تعریف شده در زبان‌های دیگر، ویژگی‌های جدیدی تعریف می‌شود. از این رو دقت تشخیص جملات موازی افزایش می‌یابد. ابتدا از طبقه‌بند بیشترین پراکندگی به عنوان ابزاری در تعیین احتمال شباهت دو جمله بر اساس ویژگی‌های تعریف شده استفاده و پس از آن، با توجه به احتمالات به دست آمده از طبقه‌بند، موازی بودن دو جمله تعیین شد. طبقه‌بند و ویژگی‌های تعریف شده نیز به تفصیل در هم‌ترازی جملات بیان شده است.

در این پژوهش ارزیابی به صورت دستی صورت گرفت. برای این کار به دو روش استفاده از داده آزمایشی و برچسب‌گذاری تصادفی جملات عمل شد.

---

[1]- Multilingual Lexicography
[2]- Word Sense Disambiguation
[3]- Terminology Extraction
[4]- Computer-Aided Language Learning
[5]- Tiedemann, J.
[6]- Specific-Domain
[7]- Meta-data

۲- پیشینه پژوهش

تولید پیکره‌های موازی به صورت دستی بسیار هزینه‌بر است لذا تحقیقات بسیاری در جهت تولید این منابع به صورت خودکار انجام شده است. به طور کلی روش‌های پیکره موازی به سه گروه تولید به کمک روش شباهت ساختاری، استفاده از الگوریتم‌های برگرفته از ماشین ترجمه خودکار و روش‌های مبتنی بر جستجوی هوشمند اطلاعات بین‌زبانی تقسیم می‌شوند. در ادامه چند مورد از روش‌های تولید خودکار پیکره موازی بررسی می‌شود.

رسنیک[۱] و اسمیت[۲](۲۰۰۳) روشی برای استخراج اسناد موازی با توجه به ساختار از وب ارائه کردند. در این راستا، پیکره‌ای به نام «استرند[۳]» ابداع شد، که در آن به کاوش خودکار متون موازی در وب از طریق شباهت‌های ساختار درونی صفحات وب پرداخته شده است. در این پیکره ساختاری همسان برای همه اسناد، در نظر گرفته می‌شود، سپس ساختارهای تولید شده با یکدیگر مقایسه می‌شوند. روش مطرح شده بسیار وابسته به ساختار است. این در حالی است که تمام مترجم‌ها سعی در حفظ قالب صفحه نمی‌کنند، بنابر این لازم است علاوه بر ساختار صفحات، به شباهت محتوای متون توجه شود. این شباهت مبتنی بر ترجمه کلمه به کلمه اسناد است و نتایج بازیابی را بهبود می-بخشد. با گسترش ماشین ترجمه، روش‌های مبتنی بر آن، برای بازیابی متون موازی مطرح شد. از این نمونه می‌توان به اوسکوریت[۴] و همکاران، ۲۰۱۰ اشاره کرد. در این کار با استفاده از ماشین ترجمه پایه، تمام اسناد جمع آوری شده، به یک زبان واحد تبدیل می‌شوند، سپس چندتایی‌های متداول و غیرمتداول از هر یک از متون استخراج و با شاخص‌گذاری متن بر اساس این چندتایی‌ها، بهترین متون موازی بر اساس معیار شباهت کسینوسی انتخاب می‌شود. در مقاله حاضر از روش هنزینگر[۵] (۲۰۰۶) استفاده و دو الگوریتم برای شناسایی متون کپی شده ارائه شد.

از نمونه کارهای انجام شده به کمک الگوریتم‌های جستجوی هوشمند اطلاعات می‌توان به پاتری[۶] و لانگلیز[۷](۲۰۱۱) اشاره کرد. در این مقاله پیکره‌ای به نام « پارادوکس[۸]» طراحی شده است. است. در این پیکره، ابتدا اسناد موازی از مجموعه اسناد ویکی‌پدیا استخراج و بر اساس سه معیار (کلمات با یک بار رخداد در متن، اعداد، علائم سجاوندی) شاخص‌گذاری می‌شوند، سپس با استفاده از این سه شاخص، سه ویژگی برای طبقه‌بند تعریف می‌گردد که بر اساس آن متون موازی تشخیص

[1]- Resnik, P.
[2]- Smith, N.A.
[3]- STRAND
[4]- Uszkoreit, J.
[5]- Henzinger, M.
[6]- Patry, A.
[7]- Langlais, P.
[8]- PARADOCS

داده می‌شوند. در مقاله اِنرایت[۱] و کوندارک[۲](۲۰۰۷)روش سریع برای تشخیص اسناد موازی ارائه شده شده است، نکته حائز اهمیت در اینجا، عدم نیاز به داده آموزش است. روش استفاده شده مبتنی بر جمع‌آوری تعداد کلمات با تعداد تکرار پایین از میان متون کاندید موازی است، به طوری که اسناد با داشتن کلمات با فرکانس پایین، کاندید بهتری برای اسناد موازی است. یانگ[۳] و وینگ‌لی[۴](۲۰۰۴) به به بررسی دو معیار مبتنی بر متن و اندازه آن برای هم‌ترازی اسناد پرداختند. مزیت استفاده از این دو معیار، پرهیز از دانش‌های قبلی مرتبط به زبان مثل واژه‌نامه دوزبانه است.

آنتونووا[۵] و میسیوریو[۶](۲۰۱۱) به این مسئله پرداختند که مجموعه اسناد موازی دو زبانه در وب، لزوماً ترجمه دستی یکدیگر نیست، بلکه ترجمه ماشینی از هم است. ماشین ترجمه می‌تواند کلمات و عبارات بسیار متداول را به خوبی ترجمه نماید، اما زمانی که در اسناد کلمات، عبارات و جملات کم کاربرد استفاده شده باشد، ترجمه‌های اشتباه خواهیم داشت و به دلیل این که این داده‌ها به عنوان داده آموزش مورد استفاده قرار می‌گیرند باعث ترجمه‌های نادرست متون در آینده خواهند شد. بنابراین لازم است چنین ترجمه‌هایی شناسایی و از مجموعه داده‌ای حذف شود. خدیوی[۷] و نی[۸] (۲۰۰۵) نیز این مسئله را بیان نمودند و تأثیر داده خطادار در پیکره بر ترجمه ماشین آماری را مطالعه کردند. علاوه بر آن در این مقاله روش‌هایی برای فیلتر نمودن این خطاها ارائه شده است.

در بعضی از زبان‌ها حجم اسناد موازی قابل دسترس بسیار محدود است، بنابر این محققان برای تولید پیکره‌های موازی، اسناد تطبیقی را نیز بررسی کردند. در کار تالونساری[۹] و همکاران(۲۰۰۸) راهکارهای مربوط به خزش در وب به منظور جمع‌آوری اسناد قابل مقایسه استفاده شد. پس از کاوش متون موازی و تطبیقی از وب باید جملات موازی از آن استخراج شود. در مقاله مونتنو[۱۰] و مارکو[۱۱](۲۰۰۵) برای استخراج جملات موازی، طبقه‌بندکننده بیشینه پراکندگی، آموزش داده شده است به طوری که به ازای هر دو جمله تعیین می‌نمایند که آیا آن دو ترجمه یکدیگرند یا خیر. در این کار، ابتدا با مجموعه داده‌ای اولیه، واژه‌نامه دو زبانه‌ای استخراج و توسط آن جملات موازی کاندید انتخاب می‌شود. سپس به کمک واژه‌نامه و طبقه‌بندکننده آموزش داده شده، جملات موازی

[1]- Enright, J.
[2]- Kondrak, G.
[3]- Yang, C.
[4]- Wing Li, K.
[5]- Antonova, A.
[6]- Misyurev, A.
[7]- Khadivi,
[8]- Ney, H.
[9]- Talvensaari, T.
[10]- Munteanu,D.S.
[11]- Marcu, D.

تعیین می‌شود. در مقاله آبدوئی- راوف[1] و شونک[2](۲۰۰۹) جملات موازی از متون تطبیقی، با روش بازیابی اطلاعات بین زبانی به دست آمده است. در این کار، از ترجمه ماشینی برای ترجمه متون در زبان مبدا به زبان مقصد استفاده شده است. به طوری که ابتدا اسناد به زبان مقصد ترجمه و سپس جملات به عنوان پرس‌وجوی اسناد هم تراز در زبان مبدا، استفاده می‌شوند. از این طریق جملات کاندید انتخاب می‌شوند. حال با مقایسه طول جملات و حذف اعداد و جداول با دو معیار نرخ خطای کلمه و نرخ خطای جمله، جملات موازی انتخاب می‌شوند. اسمیت و سایرین(۲۰۱۰) معتقدند کارهایی که تاکنون جهت تولید پیکره موازی صورت گرفته است، بر روی پیکره‌هایی چون اخبار با دوره زمانی یکسان و یا بر روی صفحات وب با ساختار یکسان بوده است. منبعی که تاکنون در این زمینه کمتر مورد توجه قرار گرفته است ویکی‌پدیا است، مزیت ویکی‌پدیا این است که مقالات مرتبط، به یکدیگر پیوند داده شده‌اند. در این مقاله پس از اینکه اسناد هم تراز می‌شوند با توجه به اینکه جملات موازی در نزدیکی یکدیگر رخ می‌دهند، جملات با هم، تراز می‌شوند. در مقاله فونگ[3] و چوینگ[4](۲۰۰۴) استخراج جملات موازی از سطوح مختلف همترازی اسناد بررسی شد. اساس کار آنان بر روی متون با اندازه‌های متفاوت که ممکن است عناوین یکسانی نداشته باشند قرار دارد. در این مقاله ادعا می‌شود، همان‌طور که هرچه تطابق میان متون بیشتر باشد نتایج بهتری از استخراج جملات به دست می‌آید، تطابق بیشتر جملات نیز، ما را به سوی کشف اسناد همترازتر هدایت می‌کند.

در رابطه با این پژوهش کارهای متعددی در زبان‌های فارسی- انگلیسی صورت گرفته است. برادران هاشمی و همکاران(۲۰۱۰) در مقاله خود از زمان‌های انتشار و عناوین به عنوان معیار شباهت میان دو اخبار دوزبانه استفاده و پیکره تطبیقی فارسی- انگلیسی را منتشر کردند. تحقیقات متعددی جهت تولید پیکره موازی در متون دو زبانه صورت گرفت، این در حالی است که بسیاری از پیکره‌های موجود به صورت دستی تهیه گردیده است، علاوه بر آن این پیکره‌ها به حوزه معنایی خاصی همچون رمان، اخبار، زیرنویس فیلم اختصاص دارند. در مقاله موسوی(۲۰۰۹)روش ترجمه با بهره‌گیری از پیکره موازی پیشنهادی تولید شده است. در این روش از ماشین جستجو جهت جمع‌آوری صفحات وب موازی استفاده و همچنین برخی از مراحل جهت استخراج جملات موازی، به صورت دستی صورت گرفته است. منصوری و فیلی(۲۰۱۲) نیز روش ترجمه ماشینی را ایجاد کردند، این روش نیز مبتنی بر کتاب‌های رمان ترجمه شده است. در کار رسولی و همکاران(۲۰۱۱) تمرکز بر روی صفحات ترجمه شده است. جملات و پاراگراف‌های موازی با ارائه الگوی خطی مبتنی

[1]- AbduI-Rauf, S.
[2]- Schwenk, H.
[3]- Fung, P.
[4]- Cheung, P.

بر شباهت طول جملات و پاراگراف‌ها، علائم سجاوندی و واژه‌نامه دوزبانه استخراج می‌شود. پیکره موازی دیگر در کار پیله‌ور و فیلی(۲۰۱۱) تولید شده است که از تطابق زیرنویس فیلم‌ها به دست آمده است.

۳- استخراج جملات موازی

هدف این پژوهش ارائه روشی جهت استخراج جملات موازی از منابع موجود است. در این پژوهش وب به عنوان منبع غنی‌ای از اسناد موازی مورد توجه قرار گرفته است. همان‌طور که گفته شد، راهکار استخراج جملات موازی از وب، دو مرحله اصلی تولید دادگان تطبیقی از وب و همترازی جملات را شامل می‌شود.

۳-۱- تولید دادگان تطبیقی در وب

دادگان تطبیقی مجموعه‌ای از اسناد در دو زبان است که ترجمه یکدیگر نیستند اما عناوین یکسانی دارند. برخی از این دادگان منبع با ارزشی از جملات موازی‌اند.

روش‌های متفاوتی برای استخراج دادگان تطبیقی وجود دارد. یکی از روش‌های مطرح، خزش در وب است. به طوری که ابتدا توسط روش‌های خزش بخشی از صفحات موجود در وب استخراج می‌شود. سپس با توجه به شباهت‌های ساختاری و محتوایی، دادگان تطبیقی تشخیص داده می‌شوند.

در این پروژه تمرکز بر روی پیکره «دات آی آر»[1](درودی و دیگران، ۱۳۸۸)است و از خزش و جمع‌آوری صفحات در وب به طور مستقیم پرهیز شد. علت این امر موجود بودن پیکره حاوی صفحات وب در دامنه عمومی و هزینه زیاد در جمع آوری مجدد این دادگان است.

پیکره «دات آی آر» مجموعه صفحاتی است که با خزش بر وب در دامنه «.ir»به دست آمده است، این مجموعه شامل یک میلیون صفحه است، که برای هر صفحه آدرس[2]، عنوان و متن آن موجود است. مرحله اول در تولید جملات موازی همترازی اسناد است، این امر سبب کاهش فضای جستجو در جستجو و استخراج جملات موازی است. برای همترازی اسناد از اطلاعات افزوده صفحات وب استفاده و به همین جهت دو روش مبتنی بر شباهت آدرس[3] و لنگر[4] متن پیاده‌سازی شد. علت نامیدن دادگان جمع‌آوری شده به دادگان تطبیقی این است که برخی از پایگاه‌ها از صورت معیار آدرس‌دهی استفاده می‌کنند اما محتوای صفحات، ترجمه یکدیگر یا به عبارتی موازی نیستند.

۳-۱-۱- هم‌ترازی اسناد مبتنی بر شباهت آدرس

از مجموعه پایگاه‌های حاوی دادگان تطبیقی می‌توان به پایگاه‌های دو زبانه اشاره کرد. صفحات موجود در این پایگاه‌ها به دو زبان است، علاوه بر آن برخی از این صفحات، محتوای یکسانی دارند.

---

[1]- dotIR
[2]- URL
[3]- URL Similarity matching
[4]- Anchor

محتوای این صفحات می‌تواند منبع با ارزشی از جملات موازی باشند. شناسایی این صفحات و استخراج محتوای آنها، مجموعه‌ای از دادگان تطبیقی را فراهم می‌سازد.

در صفحات حاوی اطلاعات تطبیقی در یک پایگاه، غالباً از ساختارهای مشابه استفاده می‌شود. یکی از اطلاعات افزوده کارا در تشخیص این شباهت، آدرس صفحات است. در این پژوهش الگوهایی برای شباهت میان دو آدرس تعریف می‌شود. بر این اساس آدرس صفحات مربوط به هر پایگاه جمع‌آوری و بر اساس الگوهای تعریف شده، صفحات متناظر هر یک جستجو می‌شود. برای مثال دو صفحه http://ece.ut.ac.ir/en به عنوان متناظر صفحه http://ece.ut.ac.ir/fa شناسایی شد و در اصطلاح به یکدیگر تراز شدند.

۳-۱-۲- همترازی اسناد مبتنی بر لنگر متن

روش دیگری که برای استخراج صفحات تطبیقی بررسی شد، بهره‌گیری از لنگر صفحات جهت همترازی آنها به یکدیگر است.

لنگرهای متن قسمتی از متن‌اند که قابلیت کلیک دارند و به اصطلاح ابر پیوند[1]اند. برخی از صفحات موجود در وب برای ارجاع به صفحه متناظرشان در زبانی دیگر از لنگرهای متن استفاده می‌کنند. در این صفحات نام آن زبان را در قسمتی از متن به صورت پیوند قرار می‌دهند، به طوری که با کلیک بر روی آن به صفحه متناظر در زبان مورد نظر ارجاع داده می‌شود.

در این پژوهش لنگرهای موجود برای هر صفحه بررسی و بدین طریق صفحات متناظر در دو زبان انگلیسی و فارسی استخراج شد.

۳-۲- همترازی جملات

در این پروژه برای تعیین این که دو جمله با چه احتمالی ترجمه یکدیگرند، روش‌های یادگیری ماشین در نظر گرفته شد. برای این منظور مسئله استخراج جملات موازی با یک طبقه‌بند باینری الگو می‌شود. این طبقه‌بند با در نظر گرفتن ویژگی‌های متفاوت تعریف شده میان دو جمله، احتمال موازی و غیر موازی بودن را محاسبه می‌نماید. در ابتدا برای هر جمله در سند مبدا، تمام جملات سند مقصد به عنوان کاندید آن در نظر گرفته می‌شود. برای کاهش فضای محاسباتی، جملاتی که معانی بسیار دوری از یکدیگر دارند حذف می‌شود. برای این منظور جملات کاندید از دو فیلتر هم‌پوشانی کلمات و فیلتر اندازه عبور داده می‌شود. طراحی این طبقه‌بند مبتنی بر کار مونتنو و مارکو(۲۰۰۵)است. در این مقاله طبقه‌بند بیشترین پراکندگی، به عنوان بهترین طبقه‌بند برای داده‌های متنی مطرح شده است. علاوه بر آن ویژگی‌هایی برای این طبقه‌بند تعریف شد که ضمن پیاده‌سازی آن، ویژگی‌های جدید نیز تعریف شد. ویژگی‌های بررسی شده در این پژوهش به چهار گروه دسته‌بندی شد. این ویژگی‌ها به شرح زیر است:

[1]- hyperlinke

۱- ویژگی‌های مربوط به طول جملات

این ویژگی‌ها از تعداد کلمات دو جمله کاندید انگلیسی- فارسی محاسبه شده است.

- طول جمله انگلیسی
- طول جمله فارسی
- نسبت طول جمله انگلیسی به طول جمله فارسی
- اختلاف طول جمله انگلیسی از جمله فارسی

۲- ویژگی‌های مربوط به هم‌پوشانی کلمات

در این دسته از ویژگی‌ها، از دو منبع واژگان، واژه‌نامه و شبکه واژه[1] استفاده شده است.

۲-۱- مبتنی بر واژه‌نامه

برای محاسبه این ویژگی‌ها از واژه‌نامه دو زبانه انگلیسی- فارسی استفاده شد.

- نسبت تعداد زوج کلمه پیدا شده در واژه‌نامه دوزبانه به تعداد کلمات جمله انگلیسی
- نسبت تعداد زوج کلمه پیدا شده در واژه‌نامه دوزبانه به تعداد کلمات جمله فارسی

۲-۲- مبتنی بر شبکه واژه

در این کار نسبت هم‌پوشانی مترادف[2]های هر کلمه در جمله انگلیسی با کلمه در جمله فارسی محاسبه می‌شود.

- نسبت تعداد کلمات تطبیقی به تعداد کلمات جمله انگلیسی
- نسبت تعداد کلمات تطبیقی به تعداد کلمات جمله فارسی

۳- ویژگی‌های حاصل از تطابق کلمات

در این بخش« IBM model-1 »میان دو جمله محاسبه و ویژگی‌های زیر استخراج شد:

«IBM model-1 » الگوریتمی است که بهترین تطابق میان دو جمله را از تطابق میان کلمات آن دو جمله به دست می‌آورد(براون و همکاران، ۱۹۹۳).

هر یک از این ویژگی‌ها در دو جهت تطابق فارسی-انگلیسی و انگلیسی- فارسی محاسبه شدند.

- نتیجه عدد« IBM model-1»: بالاترین امتیازی که به تطابق دو جمله می‌دهد. این نتیجه ضرب احتمال تطابق میان کلمات است.
- سه کلمه اول با حاصلخیزی[3] بالاتر: در بهترین تطابق میان دو جمله، کلمات بر اساس این که به چند کلمه در زبان دیگر تطبیق داده شده است مرتب می‌شود و سپس سه کلمه‌ای که بیشترین تعداد تطابق را دارد برگردانده می‌شود.
- تعداد کلمات در جمله انگلیسی که با هیچ کلمه در جمله فارسی کاندید تطابق ندارد.

---

[1]- Wordnet
[2]- Synset
[3]- Fertility

۴- ویژگی‌های متفرقه

در این دسته از ویژگی‌ها واحدهایی از جمله را در نظر گرفتیم که در تشخیص ترجمه بودن دو جمله بسیار موثر است. در این دسته از ویژگی‌ها، نسبت و اختلاف اشتراک تعداد این واحدها به کل این واحدها در دو جمله کاندید محاسبه شده است.

- اعداد مشترک
- کلمات خارجی مشترک
- علائم سجاوندی

پس از تعریف و پیاده‌سازی ویژگی‌ها، جملات از نظر موازی یا غیر موازی بودن طبقه‌بندی شد. برای هر جفت جمله فارسی و انگلیسی کاندید، طبقه‌بند بیشترین پراکندگی مطابق زیر احتمال موازی بودن دو جمله را مشخص می‌نماید:

فرمول Π (۱.$\frac{1}{Z(sp)}$) =$P(Class_i \mid sp)$

$$P(Class_i|sp) = \left(\frac{1}{Z(sp)}\right)\prod_{j=1}^{k}\lambda_j^{f_{i,j}(class,sp)}$$

در این فرمول، $Class_i$ دو طبقه موازی و غیر موازی را شامل می‌شود، SP بیانگر دو جمله فارسی و انگلیسی کاندید است.Z(SP)جهت عملیات نرمال‌سازی و f تابع ویژگی‌ها است و λ وزنی است که بر هر ویژگی اختصاص داده شده است.

یکی از دلایل انتخاب این طبقه‌بندکننده این است که در آن، فرض استقلال ویژگی‌ها از هم وجود ندارد.

۴- ارزیابی نتایج

پیش از به‌کارگیری روش ارائه شده برای استخراج جملات موازی باید ابتدا عملیات پیش‌پردازش بر روی صفحات جمع‌آوری شده صورت گیرد. در این بخش، پس از توضیح پیش‌پردازش نحوه تولید داده آموزش و داده آزمایشی بیان و در نهایت نتایج و ارزیابی آنها ارائه خواهد شد.

۴-۱- پیش‌پردازش

تولید هر پیکره به یک سری عملیات پیش‌پردازش بر روی اسناد جمع‌آوری نیاز دارد. از جمله کارهای اصلی در مرحله پیش‌پردازش، نرمال‌سازی متون، حذف محتوای غیر متنی، تشخیص زبان متن، یکسان سازی متون، تعیین محدوده جمله و غیره است.

در پردازش پیکره دات آی آر باید محتوای صفحات را از میان صفحاتHTML استخراج شود. پس از استخراج بدنه صفحات، لازم است محتوای صفحات نرمال‌سازی شود.

برای داده‌های متنی کدینگ‌های متفاوتی[1] وجود دارد، در نتیجه در ابتدا همه آنها به یک صورت تبدیل شد. در ادامه کاراکترهای غیر مجاز حذف و کاراکترهای عربی به فارسی تبدیل شد. به این گونه عملیات،«نرمال‌سازی متن» می‌گویند.

یکی از مشکلات عمده در پژوهش‌های وابسته به جملات فارسی، تشخیص محدوده جملات است. متأسفانه در این زمینه با توجه به تحقیقات انجام شده، کار قابل اطمینانی انجام نشده است. لذا در این کار بر اساس نقطه «.»، علامت سؤال«؟» و علامت تعجب «!» جمله‌بندی انجام شد.

۴-۲- تولید داده آموزش

برای تولید داده آموزش از پیکره اسناد موازی موجود رمان‌ها(منصوری و فیلی، ۲۰۱۲)استفاده شد. پیکره رمان‌ها به صورت دستی تهیه شد و شامل مجموعه جملات فارسی و انگلیسی است که ترجمه یکدیگرند. این پیکره ۴۰۰ هزار جمله دارد که در مجموع حاوی شش میلیون کلمه در سمت فارسی است.

برای بیان این که روش ارائه شده وابستگی زیادی به داده آموزش ندارد، تنها از ۶۰ هزار جمله اول آن استفاده شد. مشخصات آن در جدول ۱ آورده شده است:

فارسی	انگلیسی	
۶۰۰۰۰		تعداد جملات
۵۸۸۶۰	۵۹۲۹۲	تعداد جملات یکتا
۱۰۰۶۳۴۵	۸۶۸۴۸۱	تعداد کلمات
۹۵۸۸۷۶	۸۵۴۷۴۴	تعداد کلمات یکتا
۱۷	۱۵	میانگین طول جملات

جدول ۱- اطلاعات پیکره رمان

[1]-UTF-8 /Windows-1256

داده مورد نیاز برای آموزش طبقه‌بند از جملات موازی و غیر موازی‌ای تشکیل شده است که ویژگی‌های آنها محاسبه شده است. برای تولید این داده، زوج جمله را از دو فیلتر هم‌پوشانی کلمات و اندازه جملات عبور داده شد. سپس در صورتی که این دوجمله در داده رمان وجود داشتند برچسب جمله موازی و در غیر این صورت برجسب غیر موازی زده شد. مشخصات داده آموزش تولید شده در جدول ۲ است. همان‌طور که از جدول مشخص است، پنج هزار جمله از فیلترها عبور کرده است و نسبت داده منفی به مثبت در حد دو به یک نگه داشته شده است.

نسبت تعداد جملات غیر موازی به جملات موازی	تعداد جملات غیر موازی	تعداد جملات موازی
۲	۱۱۰۰۰۰	۵۵۰۰۰

جدول ۲- اطلاعات مرتبط به داده آموزش

۴-۳- تولید داده آزمایشی

مشکلی که در پیکره دات آی آر وجود دارد این است که جملات در این پیکره برچسب خورده نیستند، بنابر این انجام آزمایش از این طریق هزینه‌بر است. زیرا باید به صورت دستی تمام جملات کاندید از نظر موازی و غیر موازی بودن برچسب گذاری شود که با توجه به بالا بودن تعداد جملات کاندید غیرممکن است. همان طور که گفته شد یک راه حل آن این است که به صورت تصادفی تعدادی صفحه موازی انتخاب و جملات آن برچسب گذاری شد. سپس نتایج الگوریتم پیشنهادی بر روی این جملات بررسی و به کل پیکره تعمیم داده می‌شود.

راه حل دیگر تولید داده آزمایشی به صورت دستی است. برای این منظور تعدادی از صفحات موازی در وب انتخاب و جمله‌بندی این متون به صورت دستی انجام شد. سپس برای هر صفحه، صفحه هم ارز آن و جملات موازی میان هر یک مشخص شد. مشخصات داده آزمایشی تولید شده مطابق جدول ۳ است:

۲۶	میانگین جملات هر صفحه
۴۸	تعداد کل جملات موازی در ۱۰ صفحه
4	میانگین جملات موازی در هر صفحه
۱۰	تعداد جملات موازی نویزی
۱۰	تعداد صفحات بررسی شده

جدول ۳: مشخصات داده آزمایشی

۴-۴- نتایج حاصل از همترازی اسناد در پیکره دات آی آر

با روش‌های بیان شده، به ۲۲۰۰ زوج صفحه متناظر استخراج شد. مسئله مطرح در این روش این است که در بسیاری از موارد از صورت‌های معیار برای نمایش آدرس صفحات متناظر استفاده نشده

است. علاوه بر آن تمرکز پروژه پیکره دات آی آر، بر روی اسناد فارسی بود، لذا تعداد صفحات انگلیسی موجود در این پیکره محدود است.

۴-۵- نتایج حاصل از اجرای الگوریتم بر پیکره دات آی آر

در این مرحله ابتدا جملات کاندید استخراج شد. برای این منظور به ازای هر دو صفحه تراز شده، ضرب کارتزین جملات آنها در نظر گرفته شده است و سپس این جملات از دو فیلتر هم‌پوشانی کلمات و اندازه جملات عبور داده شد. برای مجموعه جملات باقی‌مانده ویژگی‌های هر جفت محاسبه و سپس جملات و ویژگی‌ها به طبقه‌بند داده شد.

از مجموعه جملات کاندید داده شده به طبقه‌بند، ۱۱۲۰ جمله موازی تشخیص داده شد. برای ارزیابی دقت و یادآوری این جملات، به دو طریق ارزیابی شد.

در روش اول، نتیجه اجرای الگوریتم بر روی ۱۰ صفحه برچسب زده شده که در بخش قبل توضیح داده شد، ارزیابی شد. جدول ۴حاصل این ارزیابی است.

صحت[1]	یادآوری[2]	دقت[3]
٪۹۳,۵	۰,۳	۰,۹۲

جدول ۴- نتایج ارزیابی روی داده آزمایشی

دقت معیاری است برای محاسبه میزان دقت الگوریتم و یادآوری میزان جملات صحیح بازگردانده شده از کل جملات موازی موجود را نشان می‌دهد. صحت نیز ترکیبی از دو معیار دقت و یادآوری است.

جملات موازی با دقت بالایی از صفحات وب استخراج شد. بنابر این در کاربردهایی چون ماشین ترجمه که به دقت بالا نیاز دارد این جملات بسیار سودمند خواهند بود. علت پایین بودن یادآوری به این خاطر است که واژه‌نامه دوزبانه از پیکره رمان تولید شده است، بنابر این بسیاری از کلمات در حوزه‌های مختلف را پوشش نداده است و در نتیجه بسیاری از جملات نتوانسته‌اند از فیلتر هم‌پوشانی کلمات عبور کنند. حجم بالای صفحات تطبیقی در وب سبب می‌شود که بتوان به پیکره قابل توجهی از جملات موازی رسید و در نتیجه پایین بودن یادآوری تاثیر بدی در کل نخواهد داشت.

در روش دوم، ۱۱۲ جمله(۱۰ درصد) از جملات استخراج شده به صورت دستی ارزیابی شد. برای جلوگیری از وارد شدن سلیقه‌های شخصی به این نتایج، از دو نفر برای برچسب‌گذاری استفاده شد. برای این کار جملات به پنج دسته تقسیم و مطابق با این تقسیم‌بندی برچسب‌گذاری انجام شد:

---

[1]- Accuracy
[2]- Recall
[3]- Precision

۱- (موازی) جملاتی که ترجمه دقیقی از یکدیگرند.
۲- (نیمه موازی) جملاتی که ترجمه یکدیگرند و تنها چند کلمه به نسبت تعداد کلمات کل در طول عملیات ترجمه حذف یا اضافه شده است.
۳- (تطبیقی) جملاتی که ترجمه یکدیگر نیستند اما شباهت معنایی بسیار زیادی دارند. برای مثال:
۱) خانه سه اتاق دارد.
۲) The house has two rooms
۴- (هم دامنه) جملاتی که ترجمه یکدیگر نیستند اما حوزه مربوطه به دو جمله یکسان است.
۵- (نامرتبط) جملاتی که هیچ گونه ارتباط معنایی با یکدیگر ندارند.
نتایج این برچسب‌گذاری در جدول ۵ نشان داده‌شده است:

برچسب‌زننده۲	برچسب‌زننده۱	
۶۸,۸	۶۷,۹	موازی
۸,۹	۹,۸	نیمه موازی
۱۳,۴	۱۳,۴	تطبیقی
۸,۹	۸,۹	هم دامنه
۰	۰	نامرتبط

جدول ۵- نتایج ارزیابی دستی خروجی

در آزمایشات صورت گرفته جملات طولانی با دقت بالایی استخراج شد و این به سبب تعداد زیاد ویژگی‌هایی است که تعریف شد. تعداد کلمات موجود در جملات موازی در این ۱۱۲ جمله بررسی شده برابر ۳۰۷۱ است. این بدین معناست که میانگین تعداد کلمات جملات استخراج شده تقریباً برابر ۳۰ است.
مشخصات پیکره تولید شده جملات موازی در جدول ۶ آورده شده است:

۱۱۴۲	تعداد جملات
۱۱۱۸	تعداد جملات یکتا
۲۹۲۵	تعداد کلمات یکتا فارسی
۳۰۵۵	تعداد کلمات یکتا انگلیسی
۱۷	میانگین طول جملات فارسی
۱۵	میانگین طول جملات انگلیسی

جدول ۶ مشخصات پیکره استخراج شده

پس از بررسی نتایج، دو مشکل اصلی به وجود آمد: اول این که میزان داده‌های نویز و تبلیغاتی در این پیکره زیاد است، بنابر این محتوای آنها سودمند نیست. علاوه بر آن، این پیکره بسیاری از وبگاه‌های دو زبانه فارسی- انگلیسی را پوشش نداده است و در نتیجه از آن نمی‌توان به عنوان مجموعه کامل صفحات دات آی آر استناد کرد.

۵- نتیجه‌گیری و کارهای آینده

هدف از این پژوهش ارائه روشی جهت استخراج جملات موازی در وب است. با پیاده‌سازی و تعریف ویژگی‌هایی، جملات موازی با دقت بالایی از وب استخراج شد. همچنین این جملات دامنه‌های متفاوت را پوشش می‌دهد. همان‌طور که گفته شد یکی از مشکلات محدود بودن تعداد صفحات انگلیسی موجود در پیکره دات آی آر است. بنابر این می‌توان با گسترش صفحات فارسی-انگلیسی مورد بررسی به حجم بیشتری از جملات موازی رسید. علاوه بر آن ویژگی‌های تعریف شده به صورت محلی است به طوری که به مکان جملات متن و جملات تراز شده توجهی نشده است. لذا می‌توان با تعریف ویژگی‌هایی مرتبط با مکان جملات دقت جملات موازی استخراج شده را افزایش داد.

منابع

- درودی, ا. و همکاران(۱۳۸۸) مجموعه محک استاندارد برای تحقیقات بازیابی اطلاعات وب مجله مهندسی برق و کامپیوتر ایران.
- AbduI-Rauf, S. and H. Schwenk(2009). "On the use of comparable corpora to improve SMT performance". Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 16–23).
- Antonova, A. and A. Misyurev(2011). "Building a Web-based parallel corpus and filtering out machine-translated text". Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (pp. 136–144).
- Baradaran Hashemi, H. , A. Shakery and H. Faili(2010)." Creating a Persian-English comparable corpus". Multilingual and Multimodal Information Access Evaluation, 27–39.
- Brown, P. F., V. J. D. Pietra, S. A. D. Pietra and R. L. Mercer(1993). "The mathematics of statistical machine

translation: parameter estimation". *Comput. Linguist*, 19(2), 263–311.
- Enright, J. and G. Kondrak(2007)." A fast method for parallel document identification". Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers (pp. 29–32).
- Fung, P. and P. Cheung(2004)." Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em". Proceedings of EMNLP (Vol. 4, pp. 25–26).
- Henzinger, M. (2006). "Finding near-duplicate web pages: a large-scale evaluation of algorithms". Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 284–291).
- Mansouri, A. and H. Faili(2012). "State-of-the-art English to Persian Statistical Machine Translation System". AISP 2012. Presented at the AISP 2012.
- Mosavi Miangah, T. (2009). "Constructing a large-scale english-persian parallel corpus". Meta: Journal des traducteurs, 54(1).
- Munteanu, D. S. and D. Marcu(2005). "Improving machine translation performance by exploiting non-parallel corpora". *Computational Linguistics*, 31(4), 477–504.
- Patry, A. and P. Langlais(2011)." Identifying parallel documents from a large bilingual collection of texts: application to parallel article extraction in Wikipedia". Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (pp. 87–95).
- Pilevar, M., H. Faili and A. Pilevar(2011). "TEP: Tehran English-Persian Parallel Corpus". *Computational Linguistics and Intelligent Text Processing*, 68–79.
- Rasooli, M., O. Kashefi and B. Minaei-Bidgoli(2011). "Extracting Parallel Paragraphs and Sentences from English-Persian Translated Documents". *Information Retrieval Technology*, 574–583.
- Resnik, P. and N. A. Smith(2003). "The web as a parallel corpus". *Computational Linguistics*, 29(3), 349–380.

- Smith, J. R., C. Quirk and K. Toutanova(2010). "Extracting parallel sentences from comparable corpora using document level alignment". Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 403–411).
- Talvensaari, T. and others (2008). "Focused web crawling in the acquisition of comparable corpora". *Information Retrieval*, 11(5), 427–445.
- Tiedemann, J. (2011). "Bitext Alignment". Morgan & Claypool Publishers.
- Uszkoreit, J., J.M. Ponte, A. C. Popat and M. Dubiner(2010). "Large scale parallel document mining for machine translation". Proceedings of the 23rd International Conference on Computational Linguistics (pp. 1101–1109).
- Yang, C. C. and Li, K. Wing (2004). "Building parallel corpora by automatic title alignment using length-based and text-based approaches". *Information processing & management*, 40(6), 939–955.
- Khadivi, S. and H. Ney.(2005) “Automatic Filtering of Bilingual Corpora for Statistical Machine Translation.” *Natural Language Processing and Information Systems*, 263–274.