یک روش توزیع‌شده برای استخراج چندتایی‌های فارسی- انگلیسی

سیده سارا میرمبین، محمد قاسم‌زاده و امین نظارات

چکیده: این پژوهش در حوزه ترجمه ماشینی و در رابطه با استخراج چندتایی‌ها از پیکره‌های دوزبانه به وسیله اسپارک است. در این رابطه، مهم‌ترین چالش این است که عملیات بایستی بر روی پیکره‌های متنی بزرگ انجام شود لذا بایستی به صورت توزیع‌شده و با بهره‌گیری از راهکارها و ابزارهای تحلیل داده‌های حجیم، طراحی و پیاده‌سازی شود. در واقع هنگام ترجمه متون، به وفور با چندتایی‌هایی مواجه می‌شویم که بایستی چندتایی‌های متناظر با هر کدام را بیابیم و در ترجمه‌مان درج کنیم، این کار می‌تواند از طریق جستجو در پیکره‌هایی که شامل چندتایی‌ها و ترجمه متناظر با آنها است انجام شود. روش‌های موجود، این کار را به صورت غیر توزیع‌شده انجام می‌دهند، لذا ضمن این که نیاز به زمان زیادی دارند، نمی‌توانند از پیکره‌های خیلی بزرگ بهره ببرند. برای رفع این نارسایی، در این پژوهش یک روش توزیع‌شده ارائه گردیده که فاصله بین بخش‌های چندتایی‌ها را نیز لحاظ می‌کند. راه‌حل پیشنهادی به صورت توزیع‌شده، تمام چندتایی‌های ممکن را از جملات پیکره تک‌زبانه استخراج نموده و با استفاده از ضریب همبستگی، چندتایی‌های معتبر جداشده را با استفاده از پیکره دوزبانه ترجمه می‌کند. روش پیشنهادی روی یک کلاستر محاسباتی با ٦٤ گیگابایت حافظه اصلی و پردازنده ٢٤هسته‌ای، در محیط اسپارک پیاده‌سازی گردید. داده‌های آزمایش شامل پیکره‌های فارسی و انگلیسی تک‌زبانه و نیز پیکره دوزبانه، حاوی به طور متوسط ۱۰۰ هزار جمله بودند. نتایج آزمایشی نشان می‌دهند که بدین طریق، زمان اجرا به شدت کاهش و کیفیت ترجمه نیز به طور قابل ملاحظه‌ای بهبود می‌یابد.

کلیدواژه: **الگوریتم توزیع‌شده، پیکره‌های متنی، ترجمه ماشینی، چندتایی‌ها.**

۱- مقدمه

طی چند دهه اخیر و هم‌زمان با گسترش و پیشرفت زبان‌شناسی رایانه‌ای، در بسیاری از کشورهای توسعه‌یافته، تلاش‌های همه‌جانبه در جهت ترجمه متون از طریق کامپیوتر انجام گرفت که حاصل کار با توجه به تنگناها و مسایل خاص ترجمه، ارزشمند بوده است. البته بعضی از محققین به منظور تشخیص و کشف دانش مورد نظر از شیوه‌های استخراج دانش مانند خوشه‌بندی بهره گرفته‌اند [۱]. به هر حال ترجمه ماشینی در بعضی از زمینه‌ها، عملکردی کاملاً رضایت‌بخش دارد و در بعضی زمینه‌ها نتایج به‌دست‌آمده، علی‌رغم قابل فهم بودن، بایستی ویراستاری شوند.

با توسعه اینترنت و ورود افراد زیادی جهت استفاده از آن، نیاز به ترجمه خودکار بیش از پیش مشهود گردیده است. در حال حاضر، حدود هشتاد درصد از مطالب منتشرشده در اینترنت به زبان انگلیسی هستند. این در حالی است که در بسیاری از کشورهای جهان مانند ایران، افراد زیادی هستند که به زبان انگلیسی مسلط نیستند. برای این دسته از کاربران ترجمه ماشینی می‌تواند ابزار مناسبی برای بهره‌بردن از مطالب ارائه‌شده به زبان انگلیسی باشد.

در ترجمه ماشینی اولین عامل تأثیرگذار میزان شباهت دو زبان مورد نظر است. هرچه دو زبان به هم نزدیک‌تر باشند ترجمه ماشینی بین آن دو زبان سریع‌تر رشد کرده و به نتایج خوبی رسیده است اما هرچه دو زبان از لحاظ نحوی از هم دورتر بوده‌اند ترجمه ماشینی رشد کندتری داشته است. در این رابطه، بدیهی است که زبان فارسی و زبان انگلیسی از نظر ساختار نحوی به یکدیگر نزدیک نیستند، لذا برای بهبود ترجمه ماشینی این دو زبان به یکدیگر، نیازمند پژوهش بیشتر و عمیق‌تری هستیم.

یکی از مواردی که می‌تواند به طور مؤثر در ترجمه متون مفید واقع شود، ترجمه چندتایی‌های رایج در زبان‌هاست. استفاده از فرهنگ لغت نیز غالباً به دلیل عدم ارائه لیست کاملی از ترکیبات چندتایی رایج در زبان‌ها کمتر می‌تواند راهگشا باشد. با بررسی کلیدواژه‌های جستجوشده به زبان فارسی در سایت گوگل ملاحظه می‌شود که به طور معمول، کاربران نتوانسته‌اند معادل‌های مناسبی برای یافتن مطالب مورد نظر خود انتخاب و ترجمه نمایند. در واقع مشکل اصلی، عدم آگاهی کاربران از معادل مناسب این ترکیبات در زبان فارسی است. به عنوان مثال، با جستجوی عبارت انگلیسی shoot the breeze در سایت مترجم گوگل، عبارت "شلیک‌کردن نسیم" به عنوان معادل ارائه می‌شود، در حالی که معادل صحیح عبارت یادشده، "گپ‌زدن و اختلاط‌کردن" است. ترجمه ارائه‌شده توسط گوگل از روی فرهنگ لغت و بدون در نظر گرفتن این که این یک ترکیب چندتایی است، صورت گرفته است. بنابراین ساخت بانک اطلاعاتی و پیکره‌ای (Corpus) که حاوی ترجمه دقیق‌تری از ترکیبات چندتایی فارسی- انگلیسی باشد ضروری است.

با ساخت پیکره‌ای از چندتایی‌ها و ترجمه متناظر با آنها و به کارگیری آن در سیستم‌های ترجمه ماشینی، می‌توان دقت ترجمه را افزایش داد. همچنین در بازیابی اطلاعات نیز از این پیکره استفاده می‌شود. این پیکره به صورت جداگانه نیز می‌تواند برای افرادی که فقط به دنبال ترجمه چندتایی مورد نظرشان هستند مفید واقع شود.

هدف از انجام این پژوهش ساخت پیکره‌های تک‌زبانه و دوزبانه‌ای است که بتوان در پژوهش‌های ترجمه ماشینی زبان فارسی که مبتنی بر پیکره هستند، از آنها استفاده نمود. پس از آن به دنبال استخراج چندتایی‌ها و الحاق آنها به نزدیک‌ترین ترجمه هستیم. در این پژوهش قرارگرفتن فاصله بین بخش‌های یک کلمه یا عبارت نیز لحاظ شده است.

پیکره خروجی که شامل چندتایی‌ها و معادل آنها می‌باشد می‌تواند در سیستم‌های ترجمه ماشینی مورد استفاده قرار گیرد و در ترجمه جملات هنگام برخورد با این چندتایی‌ها به جای ترجمه تک‌تک کلمات آن و ارائه ترجمه‌ای نامأنوس و گاهی غلط، ترجمه صحیح آنها ارائه شود. همچنین این پیکره می‌تواند به منظور آموزش چندتایی‌ها در تعلیم زبان انگلیسی مورد استفاده قرار گیرد.

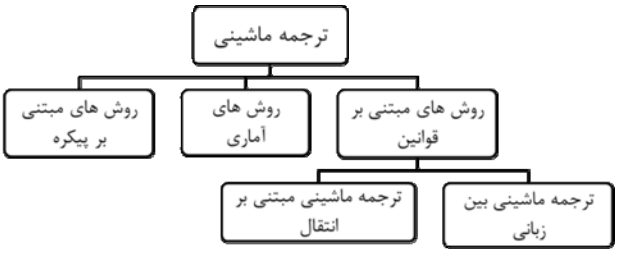
در ادامه، ساختار مقاله بدین شرح است: در بخش دوم مفاهیم اصلی

این مقاله در تاریخ ۵ خرداد ماه ۱۳۹۸ دریافت و در تاریخ ۲۸ آبان ماه ۱۳۹۸ بازنگری شد.

سیده سارا میرمبین، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: sara.mirmobin@stu.yazd.ac.ir).

محمد قاسم‌زاده (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: m.ghasemzadeh@yazd.ac.ir).

امین نظارات، گروه مهندسی کامپیوتر، دانشگاه پیام نور یزد، یزد، ایران، (email: aminnezarat@pnu.ac.ir).



شکل ۱: انواع روش‌های ترجمه ماشینی.

مربوط به ترجمه ماشینی، الگوریتم‌های توزیع‌شده و به کارگیری الگوریتم‌های توزیع‌شده در ترجمه ماشینی آمده است. در بخش بعدی به تعدادی از پژوهش‌های مرتبط که ذهنیت ارزشمندی در رابطه با پیاده‌سازی روش‌های معرفی‌شده ارائه می‌دهند می‌پردازیم. جزئیات ساخت پیکره‌ها، روش پیشنهادی، نحوه استخراج چندتایی‌ها از پیکره‌های تک‌زبانه و الحاق چندتایی‌ها به نزدیک‌ترین ترجمه‌شان در بخش چهارم شرح داده شده‌اند. در ادامه نحوه پیاده‌سازی روش پیشنهادی تشریح می‌گردد و نتایج حاصل از پیاده‌سازی مورد بحث و بررسی قرار می‌گیرند.

۲- دانش پس‌زمینه

در این بخش، معرفی دانش پس‌زمینه مرتبط با پژوهش انجام‌شده ارائه می‌گردد.

۲-۱ ترجمه ماشینی

ترجمه ماشینی یکی از جذاب‌ترین شاخه‌ها در زمینه پردازش زبان طبیعی است. با ماشینی‌شدن کارها و کاهش نقش مستقیم انسان در به انجام رساندن پروژه‌های مختلف، لزوم وجود نرم‌افزاری هوشمند برای ترجمه روان متون از یک زبان مبدأ به زبان مقصد شدت گرفت. در این رابطه، روش‌های موجود برای ترجمه ماشینی را می‌توان به روش‌های مبتنی بر قوانین، روش‌های آماری و روش‌های مبتنی بر پیکره تقسیم کرد [۲]. شکل ۱ روش‌های ترجمه ماشینی را به نمایش می‌گذارد.

۲-۱-۱ روش‌های مبتنی بر قوانین

در روش‌های مبتنی بر قوانین، قواعد زبان‌های مبدأ و مقصد به طور ثابت لحاظ می‌شوند، از این رو وابسته به زبان می‌باشند. به طور کلی روش‌های مبتنی بر قوانین، ابتدا متن را تجزیه می‌کنند و سپس یک نمایش نمادین واسطه‌ای (زبان میانی) برای آن برمی‌گزینند. متن زبان مقصد از روی این نمایش نمادین تولید می‌شود.

با توجه به ماهیت این نمایش میانی، عناوین متفاوتی برای روش‌های متناظر برگزیده شده که عبارتند از:

- ترجمه ماشینی بین زبانی
- ترجمه ماشینی مبتنی بر انتقال

در روش ترجمه بین زبانی، زبان مبدأ به یک زبان مصنوعی عمومی تبدیل می‌شود. در واقع این یک نمایش مجرد سطح بالا است که از روی آن می‌توان متن را به هر زبان دیگر ترجمه کرد.

در یک سیستم ترجمه مبتنی بر انتقال، ابتدا متن ورودی از نظر ریخت‌شناسی و نحوی تجزیه و تحلیل می‌شود تا یک نمایش معنایی برای آن به دست آید. این نمایش معنایی، سپس می‌تواند به یک سطح بالاتر تجرید، پالایش شود. در این پالایش بر بخش‌های مربوط به ترجمه تأکید می‌شود و سایر اطلاعات زاید نادیده گرفته می‌شود. در فرایند انتقال، این نمایش سطح بالا به نمایشی در همان سطح اما در زبان مبدأ تبدیل می‌شود. این دو نمایش با عنوان نمایش‌های میانی شناخته می‌شوند. برای تولید زبان خروجی فرایند نمایش به صورت معکوس انجام می‌شود [۲].

۲-۱-۲ روش‌های آماری

در روش‌های آماری سعی می‌شود تا ترجمه مورد نظر با استفاده از روش‌های آماری بر اساس متون دوزبانه موجود به دست آید. وقتی مقدار زیادی متون دوزبانه در دسترس باشند، به نتایج بسیار شگفت‌انگیزی در ترجمه می‌توان دست یافت. در این روش‌ها با توجه به هم‌نشینی کلمات در متون مبدأ و مقصد، کلمات مناسب انتخاب می‌شوند.

در سال‌های گذشته در توصیف روش‌های آماری گفته می‌شد زمانی که مقدار زیادی متون در دسترس باشند، می‌توان به نتایج بسیار شگفت‌انگیزی در ترجمه دست یافت، ولی متأسفانه حجم این متون هنوز بسیار اندک است یا در مقایسه روش‌های آماری در مقابل روش‌های مبتنی بر قوانین گفته می‌شد اشکال عمده روش‌های ترجمه آماری در این است که فراهم‌کردن متون دوزبانه برای آموزش آن بسیار دور از دسترس است و فقط شرکت‌ها و مؤسسات خاصی به متونی از این دست دسترسی دارند اما دقت این روش‌ها نسبتاً خوب است و ترجمه قابل قبولی ارائه می‌دهند. اما با گذشت زمان و پیشرفت چشم‌گیر تکنولوژی، آن قدر پیکره‌های مختلفی جمع‌آوری و ارائه شد و مورد استفاده قرار گرفت که در روش‌های ترجمه دسته جدیدی با نام روش‌های مبتنی بر پیکره [۲] برای آن لحاظ شد.

۲-۱-۳ روش‌های مبتنی بر پیکره

در روش‌های مبتنی بر پیکره برای ترجمه متون جدید از مجموعه متون ترجمه‌شده قبلی استفاده می‌شود. در واقع پیکره زبانی عبارت است از مجموعه‌ای از متن‌های نوشتاری یا گفتاری که می‌توان در توصیف و تحلیل زبان از آن بهره گرفت. اصطلاح پیکره را به ویژه زبان‌شناسان ساختگرا به کار می‌بردند و همواره تأکید می‌کردند که توصیف یک زبان یا گویش باید مبتنی بر داده‌های گردآوری‌شده و تحلیل این داده‌ها باشد و فراگیری و بزرگی پیکره عامل تعیین‌کننده‌ای در افزایش دقت و اعتبار نتایج به شمار می‌آید [۳]. در روش مبتنی بر پیکره از الگوریتم‌های مختلفی برای رسیدن به هدف مورد نظر استفاده می‌شود. یکی از روش‌های متداول استفاده از همان روش‌های آماری است.

۲-۲ الگوریتم‌های توزیع‌شده

عصر حاضر با توسعه شتابان فناوری مواجه بوده است و با توجه به فراگیرشدن استفاده از خدمات الکترونیکی و همچنین استفاده از شبکه‌های اجتماعی، حجم زیادی از اطلاعات تولید می‌شود. به دلیل حجم بالا و عدم ساخت‌یافتگی این اطلاعات، پوشش آنها از طریق پایگاه داده‌های سنتی و رابطه‌ای، امکان‌پذیر نبوده و باید از راهکارهای نوین برای پردازش آنها استفاده می‌شد تا با افزایش سرعت پردازش، دستیابی به نتایج مورد نظر در زمان قابل قبولی میسر گردد.

از سوی دیگر با ارزان‌شدن ادوات ذخیره‌سازی و توسعه راهکارهای استخراج دانش، یافتن الگوهای پنهان از میان حجم عظیمی از داده‌های تولیدشده، تبدیل به روندی بسیار کاربردی و جذاب تحت عنوان کلی "داده‌های حجیم" شده است. بالابودن حجم داده‌ها، مدیریت و تجزیه و تحلیل این داده‌ها را متفاوت می‌کند و اگر از تعداد کمی گره‌های پردازشی استفاده شود، با توجه به این حجم بالا، پردازش با سرعت کمتری صورت می‌پذیرد. برای افزایش سرعت پردازش، گره‌های بیشتر و همچنین قدرت پردازش بیشتری مورد نیاز است که این نیز به نوبه خود، هزینه بالاتری را

به دنبال دارد [۴].

نگاشت- کاهش (Map-Reduce) روشی است که به صورت گسترده در حل مسایل داده‌های حجیم مورد استفاده قرار می‌گیرد. این روش مسایل داده‌های حجیم را بر روی ساختارهای توزیع‌شده سخت‌افزاری حل می‌نماید. در این روش اطلاعات بر روی نگاشتگرها و محاسبات بر روی هر نگاشتگر توزیع می‌گردد و محاسبات بر روی هر نگاشتگر به صورت جداگانه انجام می‌پذیرد و نتایج محاسبه به کاهنده ارسال می‌شود. در این رابطه ابزارهای متفاوتی ارائه شده‌اند که هدوپ یکی از موفق‌ترین و معروف‌ترین آنها است. هدوپ یک ابزار متن‌باز است که برای پردازش و ذخیره‌سازی داده‌های حجیم به کار می‌رود [۵].

مدل پردازشی نگاشت- کاهش با وجود مزایای فراوانی که دارد از جمله مقیاس‌پذیری، تحمل خطا و مدل ساده پردازشی برای تولید برنامه و پردازش داده، دارای دو عیب عمده است: اول این که این سیستم برای کار با دیسک طراحی شده و تمام نتایج در مراحل مختلف باید در دیسک ذخیره شوند که این خود سرعت پردازش را بسیار پایین می‌آورد و دوم این که توابع آماده آن بسیار محدود هستند و بار اصلی تولید برنامه و پردازش داده بر روی برنامه‌نویسان است.

در سال ۲۰۰۹ دانشگاه برکلی مدلی جدید برای پردازش داده‌های حجیم با نام آپاچی اسپارک ارائه داد که تمرکز آن بر روی انجام محاسبات درون حافظه بود یعنی تا حد امکان و با وجود ظرفیت محدود حافظه اصلی، محاسبات درون حافظه انجام می‌شود. این امر باعث می‌شود سرعت پردازش داده‌ها نسبت به هدوپ معمولی در پردازش‌های دیسک‌محور تا ده برابر و در پردازش‌های درون‌حافظه‌ای تا صد برابر افزایش پیدا کند که خود بهبود بسیار زیادی را نشان می‌دهد و برای الگوریتم‌های تکرارشونده بسیار عالی عمل می‌کند [۴].

۲-۳ الگوریتم‌های توزیع‌شده در ترجمه ماشینی

در رابطه با الگوریتم‌های یادگیری، به طور معمول هرچه حجم داده‌های به کار رفته بیشتر باشد، دقت و صحت خروجی حاصل بیشتر خواهد بود. ترجمه ماشینی نیز از این قاعده مستثنا نیست. با گسترش فناوری و تولید داده‌های بیشتر، پژوهشگران این حوزه نیز اقدام به استفاده از این داده‌ها نمودند. حتی نشان داده شده که در سال‌های اخیر، رشد تولید داده‌های آموزش برای ترجمه ماشینی بسیار سریع‌تر از عملکرد کامپیوترهای شخصی بوده است [۶].

تقسیم الگوریتم‌های مدل‌سازی که این داده‌ها را بر روی کلاسترهای محاسباتی، پردازش می‌کند با چالش‌هایی مانند هماهنگ‌سازی و مبادله داده‌ها روبه‌رو است. با این وجود، مدل پردازشی نگاشت- کاهش به عنوان یک راه حل برای این مسایل مطرح شده که دارای عملکردی قدرتمند است، اطلاعات مربوط به سطح سیستم را از محقق پنهان می‌کند و برنامه‌ها را در میان کلاسترهای محاسباتی توزیع می‌کند. در ادامه این پژوهش شرح به کارگیری مدل پردازشی نگاشت- کاهش، برای تخمین پارامتر مدل الحاق دو کلمه به هم و یک مدل ترجمه بر اساس اصطلاح، آمده است. به کارگیری این روش پردازشی برای ترجمه ماشینی در زبان فارسی بسیار کم بوده است هرچند به کارگیری این روش در سایر زبان‌ها بسیار موفق بوده است.

۳- پیشینه تحقیق

در این بخش به تشریح پژوهش‌های مرتبط انجام‌شده در این حوزه می‌پردازیم. مطالعه این پژوهش‌ها، دانش مورد نیاز برای انجام پژوهش جاری را فراهم نموده است.

پژوهش [۷] با استفاده از روش آماری و مبتنی بر پیکره به استخراج چندتایی‌های فارسی- انگلیسی پرداخته است. در این پژوهش که با هدف دستیابی به یک پیکره دوزبانه فارسی به انگلیسی صورت گرفته است، توسط یک ربات نرم‌افزاری اقدام به جمع‌آوری متون و واژه‌های مختلف از وب‌سایت‌های متعدد شده است. با توجه به این که در یک پژوهش دیگر توسط موسوی میانگاه [۸]، یک پیکره متنی دوزبانه با بیش از سه میلیون و پانصد هزار واژه تهیه شده بود، در این پژوهش اقدام به گسترش همان پیکره نموده و تعداد واژه‌ها را به بیش از چهار میلیون و پانصد هزار کلمه رسانده و تنوع انواع جملات را مطابق پیکره تک‌زبانه گسترش داده‌اند. برای تشخیص عبارات چندتایی در زبان انگلیسی از پیکره Word Net که یکی از بزرگ‌ترین و بهترین پیکره‌هاست بهره برده‌اند.

به منظور استخراج همه چندتایی‌ها در زبان‌های مبدأ و مقصد بدین صورت عمل کرده‌اند: ابتدا کل ساختار یک جمله را استخراج و سپس، مقدار حرکت به جلو (g) را مشخص کرده‌اند. مقدار حرکت به جلو عددی است که بیشترین تعداد کلماتی را که می‌تواند در یک چندتایی معتبر وجود داشته باشد، مشخص می‌کند. در این پژوهش g برابر ۴ لحاظ شده که انتخاب این مقدار بر مبنای تجربه زبان‌شناسی صورت گرفته است. آن گاه با استفاده از یک نرم‌افزار کامپیوتری که با زبان برنامه‌نویسی C# نوشته شده است، اقدام به شناسایی تمامی ترکیبات ممکن از چندتایی‌های مستتر در جمله شده است.

برای تعیین چندتایی‌های معتبر و حذف داده‌های اضافه، روش x^2 به کار گرفته شده است. انتخاب این روش به این دلیل بوده که با استفاده از فرمول $x^2(d,c)$ می‌توان میزان وابستگی دو عبارت d,c در تمامی جملات یک پیکره را بررسی و میزان فراوانی ترکیبات مختلف وقوع یا عدم وقوع هر یک یا ترکیبی از هر کدام را محاسبه نمود. در ادامه، میزان وابستگی هر کدام به هم به دست می‌آید.

سپس به منظور انتخاب ترکیبات مناسب، نیاز به یک مقدار حد آستانه وجود داشته که مقدار آستانه ۶٫۶۳ پس از محاسبه مقادیر مختلف توسط پژوهشگر انتخاب شده است.

اگر مقدار x^2 محاسبه‌شده برای هر یک از ترکیب‌ها کوچک‌تر از مقدار حد آستانه بود، بدین معنی تفسیر شده که می‌توان وابستگی بین کلمات آن ترکیب را قبول کرد و مقادیر بالاتر از حد آستانه را رد نمود.

برای هریک از رکوردهای به‌دست‌آمده، الگوریتم معادل‌سازی زیر اعمال و نتیجه محاسبه شده است:

۱) تمامی ترکیب‌های از دو تا چهارتایی جمله متناظر استخراج می‌شود.
۲) مقدار فرمول x^2 برای هر یک از این ترکیبات با استفاده از پیکره تک‌زبانه محاسبه می‌شود.
۳) ترکیب‌های با مقدار x^2 کمتر از ۶٫۶۳ نگه داشته و مابقی حذف می‌شوند.
۴) سپس مقدار x^2 برای چندتایی معتبر و ترکیب معادلش محاسبه می‌گردد و ترکیبی که بیشترین همبستگی را دارد به عنوان معادل برگزیده می‌شود.

کاستی پژوهش فوق این است که زمان اجرا بالا بوده و به همین دلیل امکان استفاده از پیکره‌های با اندازه بزرگ برای آن وجود ندارد.

در پژوهش دیگری، با اشاره به قدرتمندبودن روش‌های آماری در ترجمه ماشینی این نکته را متذکر می‌شوند که روش‌های آماری به دلیل استفاده از ترجمه لغوی در دو زبان با ساختار متفاوت، از نقص‌های زیادی رنج می‌برند [۹].

جدول ۱: متغیرهای فرمول ضریب همبستگی x^2.

متغیر	توضیحات
n_{11}	تعداد جملاتی که هر دو بخش چندتایی در آنها به ترتیب پشت سر هم قرار دارند.
$n_{1.}$	تعداد جملاتی که بخش اول چندتایی فقط در آن وجود دارند.
$n_{.1}$	تعداد جملاتی که بخش دوم چندتایی فقط در آن وجود دارند.
$n_{..}$	تعداد جملاتی که هیچ یک از بخش‌های چندتایی در آنها وجود ندارند.

روش پیشنهادی این پژوهش برای تشخیص و ترجمه چندتایی‌ها به‌کارگیری الگوریتم فیلترینگ چندلایه است. هدف این الگوریتم الحاق بخش‌بندی‌های جمله‌های دو زبان متفاوت است. این الحاق به صورت یک به یک (one-to-one, one-to-null and null-to-one) است. در این روش ابتدا بخش‌هایی که پرتکرارتر هستند فیلتر شده و به عنوان بخش‌های کاندید انتخاب می‌شوند و به همین دلیل الگوریتم خوشه‌بندی که در مرحله بعد برای الحاق بخش‌های مختلف دو زبان به کار می‌رود، عملکرد مناسب‌تری خواهد داشت. برای تشخیص بخش‌های کاندید، به هر بخش عددی به نام $d(k)$ نسبت می‌دهیم که در آن، k طول بخش مورد نظر است (تعداد کلمات بخش مورد نظر) که طبق (۱) به دست می‌آید

$$\begin{aligned} d_k &= d(\omega_1, \omega_2, \ldots, \omega_k) = \\ &1 - \beta \times MI(\omega_1, \omega_2, \ldots, \omega_k) + \beta \times p(\omega_1, \omega_2, \ldots, \omega_k) \\ MI(\omega_1, \omega_2, \ldots, \omega_k) &= \\ &p(\omega_1, \omega_2, \ldots, \omega_k) \times \log^{\frac{p(\omega_1, \omega_2, \ldots, \omega_k)}{p(\omega_1) \times \ldots \times p(\omega_k)}} \end{aligned} \quad (1)$$

که $p(\square)$ احتمال وقوع این بخش از کلمات در کل متن و b ضریبی بین ۰ و ۱ است. حداکثر طول یک بخش، ۴ کلمه انتخاب شده است، چون بخش‌های با طول ۵ کم‌تکرار بوده‌اند. هر بخش که $d(k)$ بیشتری داشته باشد، کاندید مورد نظر است. برای مقایسه منصفانه بخش‌های با طول متفاوت، فرمول مورد نظر نرمال‌سازی شده است.

این پژوهش (به جهت این که ترجمه بخش‌های جمله مورد توجه بوده است) برای زبان‌های فارسی و انگلیسی که از لحاظ نحوی از هم دورند مناسب نمی‌باشد.

در پژوهش دیگری، محققان چینی برای سیستم ترجمه سخنرانی، ابتدا اقدام به ساخت پیکره مورد نظرشان کرده‌اند. این پیکره شامل متن سخنرانی‌ها و ترجمه آنهاست که توسط مترجمان حرفه‌ای ترجمه شده‌اند [۱۰]. البته چنین تفسیرهایی همیشه به عنوان داده برای این ماشین ترجمه مناسب نبوده است زیرا تفسیرهای هم‌زمان در محیط واقعی ممکن است شامل ترجمه‌های شفاهی باشد. بنابراین ترجمه‌های جدیدی را به این داده‌ها اختصاص داده‌اند. از آنجایی که اساساً یک جمله در یک سخنرانی طولانی است برای سیستم ترجمه ضروری است که واژگان کوتاه‌تر نسبت به جملات را به‌عنوان واحدهای ترجمه لحاظ کنند که این فرایند شامل ۳ مرحله است:

۱) تقسیم‌بندی: تقسیم یک جمله ورودی به واحدهای مناسب
۲) ترجمه: ترجمه واحدهای تولیدشده از مرحله قبل
۳) ترکیب: ترکیب واحدهای ترجمه به طوری که ترجمه نهایی یک جمله درست انگلیسی باشد.

این مراحل هم‌زمان با دریافت ورودی انجام می‌شوند. جمله به ۴ واحد تقسیم شده است. هم‌زمان با تقسیم‌بندی جمله، سیستم واحدهای ترجمه را به عبارات انگلیسی ترجمه می‌کند و سپس نتایج ترجمه‌ها را به هم متصل و جمله نهایی را تولید می‌کند.

برای تحقق این فرایند سیستم باید واحدهای کوچک‌تر از جمله را به عنوان واحد ترجمه در نظر بگیرد و این واحدها را به درستی شناسایی کند. یکی از شرط‌های واحد ترجمه این است که این چندتایی که به عنوان واحد انتخاب می‌شود، دارای معنی باشد. زیرا ترجمه برای هر چندتایی به صورت جداگانه انجام می‌شود. با این حال چندتایی‌هایی وجود دارد که قادر به ترجمه آنها به تنهایی نیستند لذا مجبور به در نظر گرفتن چندتایی‌های مجاور برای ترجمه شده‌اند. از شرط‌های دیگر، توجه به تأخیر در تفسیرهای واقعی توسط مترجمان حرفه‌ای بوده که یک واحد ترجمه را در طول ۴/۳ ثانیه تعریف کرده‌اند. پژوهش اخیر از لحاظ این که بخش مورد نظر دارای معنی باشد، ضعیف عمل می‌کند و به کارگیری یک پیکره از چندتایی‌ها می‌تواند برای رفع این کاستی به کار رود.

٤- روش پیشنهادی

در این پژوهش با استفاده از پیکره تک‌زبانه فارسی، اقدام به استخراج چندتایی‌ها در زبان فارسی و به کمک پیکره تک‌زبانه انگلیسی نیز اقدام به استخراج چندتایی‌ها در زبان انگلیسی می‌شود. سپس با استفاده از پیکره دوزبانه اقدام به معادل‌سازی هر یک از چندتایی‌های به‌دست‌آمده می‌کنیم. در این پژوهش به دنبال اجرای توزیع‌شده فرایند فوق، به روشی منطقی و مقرون به صرفه هستیم.

راه حل پیشنهادی این است که از پیکره‌هایی که هر خط آن یک جمله است، جمله به جمله تمام چندتایی‌های ممکن، استخراج و سپس اعتبارسنجی انجام شود و چندتایی‌هایی که معتبر هستند جدا گردند. پس از آن با استفاده از پیکره دوزبانه و طی یک روش آماری و مبتنی بر فرکانس تکرار نزدیک‌ترین ترجمه به چندتایی مورد نظر الحاق شود.

در مرحله استخراج چندتایی‌ها، فاصله‌افتادن بین بخش‌های چندتایی نیز در نظر گرفته می‌شود. زیرا به طور مثال خواب سبک یک چندتایی است که می‌تواند به شکل خواب خیلی سبک نیز در جمله به کار رود.

در واقع در روش پیشنهادی این پژوهش که به صورت توزیع‌شده عمل می‌کند و امکان به کارگیری پیکره‌های بزرگ زبانی را فراهم می‌کند، در استخراج چندتایی‌ها، امکان فاصله‌افتادن بین بخش‌های چندتایی‌ها را نیز در نظر گرفته است. برای استخراج تمام چندتایی‌های موجود در جمله، ابتدا باید کلمات جمله توسط جداکننده فاصله از هم جدا شوند و به ترتیب در لیستی قرار گیرند و سپس با پیمایش روی لیست تمام چندتایی‌های ممکن استخراج شوند. بر اساس تجربه زبان‌شناسی، تعداد چندتایی‌ها بهتر است تا چهار لحاظ شود [۷].

در این مرحله از پژوهش اقدام به ارائه راه حلی شده که فاصله نیز بین کلمات چندتایی‌ها در نظر گرفته شود تا امکان آمدن کلمات دیگر بین بخش‌های چندتایی‌ها بررسی شود و بدین نتیجه رسید که بخش‌های یک چندتایی می‌توانند گاهی بدون فاصله و پشت سر هم بیایند و گاهی با فاصله و این فاصله می‌تواند متغیر باشد.

پس از استخراج تمام چندتایی‌های ممکن از جملات باید اقدام به جداسازی آنهایی که معتبر هستند نمود. در این پژوهش برای انجام این مهم، از ملاک زیر استفاده شده است

$$x^2 = \frac{(n_{11} + n_{1.} + n_{.1} + n_{..}) \times ((n_{11} \times n_{..}) - (n_{1.} \times n_{.1}))^2}{(n_{11} + n_{1.}) \times (n_{.1} + n_{..}) \times (n_{11} + n_{.1}) \times (n_{1.} + n_{..})} \quad (2)$$

متغیرهای (۲) در جدول ۱ مشخص شده‌اند.

ضریب همبستگی x^2 نشان‌دهنده میزان همبستگی بخش‌های چندتایی

جدول ۲: اندازه پیکره‌های مورد استفاده پژوهش.

نوع پیکره	اندازه پیکره (تعداد جمله)
پیکره تک‌زبانه فارسی	۱۳۲۹۱۲۱
پیکره تک‌زبانه انگلیسی	۳۹۳۹۱۹۵
پیکره دوزبانه فارسی- انگلیسی	۶۶۳۰۰۹۶

مورد نظر است. هرچه بخش‌های یک چندتایی بیشتر در کنار هم و کمتر جدا از هم آمده باشند، میزان همبستگی آن‌ها بیشتر است. در پژوهش [۷] حد آستانه پس از محاسبه مقادیر مختلف توسط پژوهشگر، ۶٫۶۳ اعلام شده است بدین صورت که برای هر یک از چندتایی‌ها اگر مقدار x^2 کوچک‌تر از مقدار حد آستانه باشد، می‌توان وابستگی بین بخش‌های چندتایی را قبول کرد و مقادیر بالاتر از حد آستانه را رد نمود.

پس از آن که پیکره‌ای از چندتایی‌های معتبر حاصل شد، باید به دنبال نزدیک‌ترین ترجمه‌شان با استفاده از پیکره دوزبانه باشیم. برای این کار باید به ازای هر چندتایی، جملاتی از پیکره دوزبانه که در آن حضور دارد را جدا کرد، سپس باید تمام جمله‌های معادل در زبان مقابل را به چندتایی‌ها تبدیل کرد و برای چندتایی معتبر و هر یک از چندتایی‌های جمله زبان مقابل، ضریب همبستگی محاسبه شود و چندتایی‌ای که همبستگی بیشتری دارد به عنوان ترجمه بازگردانده شود.

۵- پیاده‌سازی و اجرا

روش پیشنهادی بر روی کلاستری با ۶۴ گیگابایت حافظه داخلی و ۲۴ گره در محیط اسپارک و به زبان اسکالا پیاده‌سازی گردید. در این رابطه سه پیکره زیر به کار گرفته شدند: ۱) پیکره تک‌زبانه فارسی، ۲) پیکره تک‌زبانه انگلیسی و ۳) پیکره دوزبانه فارسی- انگلیسی.

برای پیکره تک‌زبانه فارسی، از پیکره یک پژوهش مرتبط [۷] و نیز ترجمه قرآن کریم از سایت http://tanzil.net/trans بهره گرفتیم. برای پیکره تک‌زبانه انگلیسی، پیکره منتشرشده توسط سایت ویکی‌پدیا را لحاظ نمودیم. برای پیکره دوزبانه فارسی- انگلیسی علاوه بر پیکره مورد استفاده در پژوهش [۷]، پیکره دوزبانه ویکی‌پدیای منتشرشده توسط دانشگاه تحصیلات تکمیلی زنجان و پیکره دوزبانه منتشرشده توسط سایت http://www.manythings.org/anki را به کار بردیم.

از ترکیب پیکره‌های جمع‌آوری‌شده، پیکره‌های جدیدی حاصل شد که برای این پژوهش و پژوهش‌های آتی مفید خواهد بود. اندازه پیکره‌های مورد استفاده به طور خلاصه در جدول ۲ آمده است.

پیکره‌های جمع‌آوری‌شده در هر دسته جداگانه به یک شکل واحد درآمده و در کنار هم قرار گرفته‌اند. سپس چون واحد پردازش در این پژوهش جمله است پیکره‌ها به هر خط یک جمله تبدیل شده‌اند. هر سه پیکره مورد استفاده از فرمت‌های csv، accdb، bak و غیره به فرمت txt تبدیل شده‌اند. با انجام این تبدیل انجام هر گونه عملیات مانند تبدیل هر خط به یک جمله به راحتی امکان‌پذیر گردید. پیکره‌های تک‌زبانه هر خط فقط به یک جمله تبدیل شد که در انتهای آن هیچ علامتی وجود ندارد. در پیکره دو زبانه، هر خط آن شامل یک جمله انگلیسی و معادلش به زبان فارسی است که توسط علامت '|' از یکدیگر جدا می‌شود.

در مرحله اول پیاده‌سازی، چندتایی‌ها از جملات استخراج می‌شوند. برای استخراج چندتایی‌های جملات از تابع زیر استفاده می‌کنیم. در الگوریتم extractionChunkOfSentences پارامتر args جمله ورودی، range مشخص‌کننده تا چندتایی بودن و distance مشخص‌کننده فاصله بین بخش‌های چندتایی است.

الگوریتم extractionChunkOfSentences

```
Require: args: String, range: Int, distance: Int
01: for each arg ← args.split(" ") do
02:   if arg != "" then
03:     res.add(arg)
04: end for
05: for i ← 0 to res.size()-1 do
06:   for j ← 2 to range do
07:     sub ← res.subList(i, i + j + y)
08:       if j == 2 then
09:         word ← word + sub(0) + " "
10:         word ← word + sub(sub.size() – 1)
11:         type ← (j * 100) + ((sub.size() - j) * 10)
12:         word ← word + "@" + type
13:         res.add(word)
14:         word ← " "
15:       end if
16:       do for other types
17:     end for
58:   end for
59: return r
```

در انتهای چندتایی نوع آن بر حسب فاصله قرار داده شده است. جدول ۲ نوع چندتایی‌ها را به نمایش می‌گذارد.

پس از استخراج تمام چندتایی‌های جملات با استفاده از ضریب همبستگی x^2 که در قسمت قبل به آن اشاره شد، چندتایی‌های معتبر جدا و وارد مرحله ترجمه می‌شوند.

```
chunksWithCorrelation ← sentencesChunks.map(t
=>findCorrelation( t._1.split("@")(0),
t._1.split("@")(1).toInt, t._2 , broadcastedChunks.value,
monolingualCorpusSize ) )
```

پیاده‌سازی تابع findCorrelation به صورت زیر است:

الگوریتم findCorrelation

```
Require: aChunk: String, typee: Int, n11:Int,
fullMap:scala.collection.Map[String, Int], all: Long):
ArrayList[String]
01: if type == 200 | type == 210 | type == 220 | type == 230
02:   n10 ← fullMap.get(tuple2.split(" ")(0)).getOrElse(0) - n11
03:   n01 ← fullMap.get(tuple2.split(" ")(1)).getOrElse(0) - n11
04:   n00 ← all - n10 - n01 - n11
08:   correlationCoefficient ← correlation (n11, n10, n01, n00)
09:   word←chunk+"@"+type+"@"+ correlationCoefficient
10:   res.add(word)
11: end if
12: else if type == 300|type == 311|type == 321|type == 331
13:   s1 ← chunk.split(" ")(0)
14:   s2 ← chunk.split(" ")(1) + " "+chunk.split(" ")(2)+"@" + 200
15:   n10 ← fullMap.get(s1).getOrElse(0) - n11
16:   n01 ← fullMap.get(s2).getOrElse(0) - n11
17:   n00 ← all - n10 - n01 - n11
18:   correlationCoefficient ← correlation(n11, n10, n01, n00)
19:   word ← chunk+"@"+type+"@"+correlationCoefficient
20:   res.add(word)
21: end else
22: do for all types
23: return res
```

پس از آن که ضرایب همبستگی محاسبه گردید، جداسازی چندتایی‌های معتبر با توجه به آستانه انجام می‌شود. سپس چندتایی‌های معتبر به مرحله ترجمه می‌روند.

در مرحله ترجمه، به ازای هر چندتایی معتبر جملاتی از پیکره دوزبانه که در آن حضور دارد را جدا کرده و جمله زبان مقابل را توسط الگوریتم extractionChunkOfSentences به چندتایی‌ها تقسیم می‌کنیم. این بار به ازای چندتایی معتبر زبان مبدأ و چندتایی‌های جمله ترجمه، ضریب همبستگی محاسبه می‌شود و آن چندتایی که بیشترین همبستگی را داشته باشد به عنوان ترجمه بازگردانده می‌شود. جدول ۳ نوع چندتایی‌ها را به‌نمایش می‌گذارد.

جدول ۳: نمایش نوع چندتایی‌ها.

نوع	چندتایی
۲۰۰	* *
۲۱۰	* - *
۲۲۰	* - - *
۲۳۰	* - - - *
۲۴۰	* - - - - *
۳۰۰	* * *
۳۱۰	* - * *
۳۱۱	* * - *
۳۲۰	* - - * *
۳۲۱	* * - - *
۳۳۰	* - - - * *
۳۳۱	* * - - - *
۳۴۰	* - - - - * *
۳۴۱	* * - - - - *
۴۰۰	* * * *
۴۱۰	* * * - *
۴۱۱	* * - * *
۴۱۲	* - * * *
۴۲۰	* * * - - *
۴۲۱	* * - - ❋ *
۴۲۲	* - - ❋ ❋ *
۴۳۰	* * * - - - *
۴۳۱	* * - - - * *
۴۳۲	* - - - * ▯ * *
۴۴۰	* * * - - - - *
۴۴۱	* * - - - - ❋ *
۴۴۲	* - - - - * * *

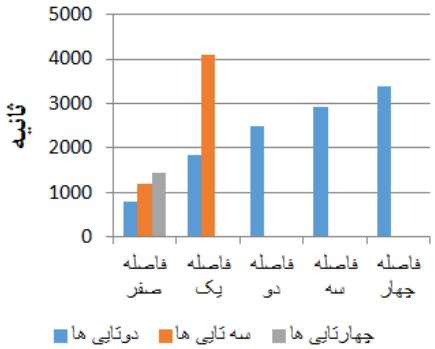
```
chunksWithPossibleTranslation←
validChunks.cartesian(bilingualCorpus) .filter(t =>
 isChunkThere(t._1, t._2.split('|')(0))) .map(t => (t._1,
extractionChunkOfSentences(t._2.split('|')(1), 4, 4)))
.flatMap { case (c, innerList) => innerList.map(c -> _)
}.flatMap(line =>
extractionChunkOfSentences(line.split('|')(0),4,4)).map(word
=> (word, 1)) .reduceByKey(_ + _) .map(t => ((t._1,
t._2), 1)).reduceByKey(_ + _) .map(t => (t._1._1, (t._1._2,
translationCorrelation(t._1._1, t._1._2, t._2,
bilingualCorpusSize,
broadcastedOriginSentencesChunks.value,
broadcastedDestinationSentencesChunks.value)))).groupBy
Key().map(t => (t._1(0), t._2.toList.sortBy(r => (r._2,
r._1)))).map(t => (t._1, t._2(0)))
```

٦- نتایج آزمایشی و تحلیل

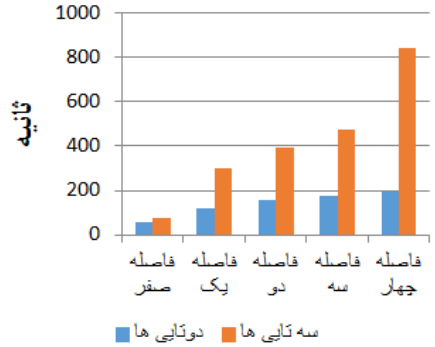
در این بخش، کارایی سیستم پیشنهادی، شامل زمان اجرا و نیز نتایج به دست آمده از پیش‌پردازش ریشه‌یابی تشریح می‌گردد.

در تمام اجراها مشخص شده که بیشتر زمان اجرا و حافظه مصرفی در بخش ترجمه چندتایی‌ها بوده است. لذا ابتدا فقط چندتایی‌ها استخراج شدند تا مشخص شود کدامین چندتایی‌ها رایج‌ترند.

در پیکره‌ای حاوی ده هزار جمله، چهارتایی‌ها تا فاصله چهار، در پیکره



شکل ۲: زمان اجرای استخراج و ترجمه دوتایی‌ها در پیکره ورودی ۵۰ هزار جمله‌ای.



شکل ۳: زمان اجرای استخراج و ترجمه دوتایی‌ها در پیکره ورودی ۱۰ هزار جمله‌ای.

حاوی پنجاه‌هزار جمله، چهارتایی تا فاصله سه، در پیکره حاوی صدهزار جمله و چهارتایی‌ها تا فاصله یک استخراج شدند. با تحلیل خروجی مشخص گردید که چهارتایی‌ها اساساً رایج نیستند اما در چهارتایی‌های بدون فاصله تعداد کمی چندتایی رایج به چشم می‌خورد.

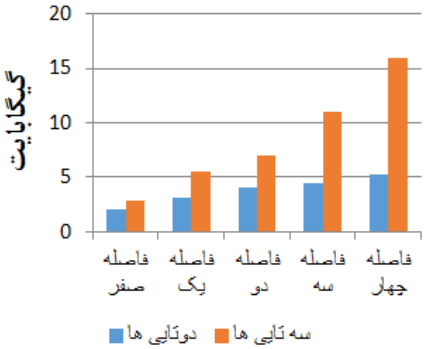
برای تحلیل سه‌تایی‌ها تا فاصله چهار، از پیکره حاوی صدهزار جمله استفاده نمودیم. نتایج خروجی نشان می‌دهند که سه‌تایی‌ها کمی بیشتر از چهارتایی‌ها رایج هستند و فاصله بیش از یک در آنها مناسب نیست و بهترین آنها همان نوع بدون فاصله است.

برای تحلیل دوتایی‌ها نیز مانند سه‌تایی‌ها از پیکره حاوی صدهزار جمله استفاده نمودیم. دوتایی‌ها بیشتر از همه رایج هستند و در بین آنها دوتایی‌های بدون فاصله کاربرد بیشتری دارند و فاصله چهار حتی در دوتایی‌ها نیز مناسب نیست و هرچه فاصله کمتر باشد چندتایی‌های رایج‌تری به چشم می‌خورند.

زمان اجرای پیاده‌سازی برای چندتایی‌های فارسی و انگلیسی با پیکره‌ی ورودی پنجاه‌هزار جمله‌ای در شکل ۲ آمده است.

در پیکره ورودی با ده هزار جمله امکان استخراج و ترجمه سه‌تایی‌ها تا فاصله چهار نیز بوده است. شکل ۳ زمان اجرا برای استخراج دوتایی‌ها و سه‌تایی‌ها با فاصله‌های مختلف را نشان می‌دهد. میانگین حافظه مصرفی در حین اجرا برای استخراج و ترجمه چندتایی‌ها در پیکره ورودی با ده‌هزار جمله در شکل ۴ قابل مشاهده است.

برای ارزیابی دقت خروجی از دوتایی‌های بدون فاصله با پیکره ورودی صد و پنجاه‌هزار جمله‌ای، نمونه‌گیری تصادفی بدون جایگذاری انجام شد. بدین صورت که ۵۰ چندتایی همراه با ترجمه‌شان انتخاب و به صورت تجربی ارزیابی شدند. نتایج حکایت از آن داشت که به دلیل تفاوت دو زبان فارسی و انگلیسی اگر در چندتایی‌هایی که در زبان معادل به عنوان ترجمه استخراج می‌شود، فاصله لحاظ شود، دقت ترجمه به طور قابل ملاحظه‌ای بهبود می‌یابد. بخشی از پیکره‌های خروجی در جدول‌های ۴ و ۵ نشان داده شده است.



شکل ۴: میانگین حافظه مصرفی در استخراج و ترجمه چندتایی‌ها با پیکره ورودی ۱۰ هزار جمله‌ای.

جدول ۴: بخشی از پیکره خروجی در زبان فارسی.

چندتایی فارسی	نوع چندتایی	نزدیک‌ترین ترجمه
ای‌داد	۲۰۰	ah dear
کنار دیوار	۲۰۰	against the wall
غیر قابل تحمل	۳۰۰	an intolerable
و با هر کس	۴۰۰	with whoever
خوش‌لباس	۲۰۰	dressed well

جدول ۵: بخشی از پیکره خروجی در زبان انگلیسی.

چندتایی انگلیسی	نوع چندتایی	نزدیک‌ترین ترجمه
be cool	۲۰۰	راحت باش
heavy hours	۲۰۰	ساعت متوالی
after supper	۲۰۰	بعد شام
let it go	۳۱۰	ولش کن
Well now	۲۰۰	خب حالا

۷- نتیجه‌گیری و پیشنهادها

این پژوهش نشان می‌دهد که با استخراج چندتایی‌ها و الحاق آنها به نزدیک‌ترین ترجمه از پیکره متناظر می‌توان به نتایج ارزشمندی دست یافت. ضمناً با بهره‌گیری از پردازش توزیع‌شده و در دست داشتن کلاسترهای محاسباتی قوی و بزرگ‌ترشدن پیکره‌های ورودی، سرعت و دقت نتایج به طور مضاعف بهبود می‌یابد. لحاظنمودن فاصله بین بخش‌های چندتایی، عامل جدید دیگری بود که در انجام آزمایش‌ها لحاظ شد و نتایج ارزشمند آن مشهود گردید.

مراجع

[۱] ا. سادات علوی، ه. مشایخی، ح. حسن‌پور و ب. رحیم‌پور کامی، "استفاده از خوشه‌بندی تکاملی برای تشخیص موضوع در بلاگ‌نویسی کوچک با لحاظنمودن اطلاعات شبکه اجتماعی،" *نشریه مهندسی برق و مهندسی کامپیوتر ایران، ب- مهندسی کامپیوتر*، جلد ۱۷، شماره ۴، صص. ۲۸۶-۲۷۷، زمستان ۱۳۹۸.

[۲] دبیرخانه شورای عالی اطلاع‌رسانی، *جمع‌آوری اطلاعات چالش‌ها و روش‌های ترجمه ماشینی زبان انگلیسی به فارسی و بالعکس*، شورای عالی اطلاع‌رسانی، مستند شماره ۱۹۰/۲۵۳۷/۱/۱، دانشگاه علم و صنعت ایران، تهران، ۱۳۸۸.

[۳] م. عاصی، "پردازش دستوری زبان فارسی با رایانه،" *نامه فرهنگستان*، جلد ۱، شماره ۱، صص. ۵۱-۲۹، اسفند ۱۳۸۳.

[۴] ش. عباسی، "داده‌های عظیم تعاریف و چالش‌ها،" *مجموعه مقالات کنفرانس بین‌المللی سیستم‌های غیر خطی و بهینه‌سازی کامپیوتر*، ۱۳ صص.، شیراز، دبی، امارات متحده عربی، خرداد ۱۳۹۴.

[۵] م. جهانی، "نو پرداز"، شرکت نوپرداز، ۱۳۹۷/۰۳/۱۹. [درون‌خطی]. Available: https://nopardazco.com. [دستیابی در ۲۲/ ۰۵ /۱۳۹۸].

[6] C. Dyer, A. Cordova, A. Mont, and J. Lin, "Fast, easy, and cheap: construction of statistical machine translation models with mapreduce," in *Proc. of the 3rd Workshop on Statistical Machine Translation,* pp. 199-207, Columbus, OH, USA, Jun. 2008.

[۷] ا. نظارات و ط. موسوی میانگاه، "طراحی و پیاده‌سازی یک سامانه بازیابی اطلاعات دوزبانه با استفاده از پیکره‌های زبانی،" *پژوهش‌نامه پردازش و مدیریت اطلاعات (علوم و فناوری اطلاعات سابق)*، جلد ۲۷، شماره ۲، صص. ۲۱۱-۱۹۸، زمستان ۱۳۹۰.

[8] T. Mousavimiyangah, "Constructing a large-scale English-Persian parallel corpus," *Meta*, vol. 54, no. 1, pp. 181-188, Jan. 2009.

[9] Y. Zhou, C. Zong, and B. Xu, "Bilingual chunk alignment in statical machine translation," in *Proc. Int. Conf. on System Man and Cybernetics,* pp. 1401-1406, Hague, The Netherlands, 10-13 Oct. 2004.

[10] M. Murata, T. Ohno, S. Matsubara, and Y. Inagaki, "Construction of chunk-aligned bilingual lecture corpus for simultaneous machine translation," in *Proc. of the 7th Conf. on International Language Resources and Evaluation, LREC'10,* pp. 1765-1770, Valletta, Malta, 19-21 May 2010.

سیده سارا میرمبین در سال ۱۳۹۵ مدرک کارشناسی مهندسی کامپیوتر (نرم‌افزار) خود را از دانشگاه اصفهان دریافت نمود. از مهرماه ۱۳۹۵ الی بهمن‌ماه ۱۳۹۷ نامبرده جهت انجام دوره کارشناسی ارشد مهندسی کامپیوتر (هوش مصنوعی) در دانشگاه یزد مشغول به تحصیل و پژوهش بودند. زمینه‌های علمی مورد علاقه نامبرده عبارتند از: بازیابی هوشمند اطلاعات، داده‌های حجیم، داده‌کاوی و شبکه‌های عصبی مصنوعی.

محمد قاسم‌زاده در سال ۱۳۶۸ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه شیراز و در سال ۱۳۷۴ مدرک کارشناسی ارشد مهندسی کامپیوتر (هوش ماشین و رباتیک) خود را از دانشگاه صنعتی امیرکبیر دریافت نمود. از بهمن ماه ۱۳۸۰ الی بهمن ماه ۱۳۸۴ نامبرده جهت انجام دوره دکتری (علوم کامپیوتر) در دانشگاه تربیر و دانشگاه پتسدام هر دو در کشور آلمان مشغول به تحصیل و پژوهش بودند. ایشان در سال ۱۳۸۴ موفق به اخذ درجه دکتری علوم کامپیوتر گردید. محمد قاسم‌زاده از سال ۱۳۷۵ تاکنون به عنوان عضو هیأت علمی در دانشگاه یزد مشغول به تدریس و تحقیق می‌باشند. زمینه‌های علمی مورد علاقه ایشان عبارتند از: طراحی و تحلیل الگوریتم‌ها، پردازش زبان طبیعی، سیستم‌های هوشمند و محاسبات نرم.

امین نظارات در سال ۱۳۸۱ مدرک کارشناسی علوم کامپیوتر خود را از دانشگاه شهید باهنر کرمان و در سال ۱۳۹۰ مدرک کارشناسی ارشد مهندسی فناوری اطلاعات خود را از دانشگاه شیراز دریافت نمود. از سال ۱۳۹۱ الی سال ۱۳۹۵ نامبرده جهت انجام دوره دکتری مهندسی کامپیوتر (نرم‌افزار) در دانشگاه شیراز مشغول تحصیل و پژوهش بودند. ایشان در سال ۱۳۹۵ موفق به اخذ درجه دکتری مهندسی کامپیوتر گردید. دکتر نظارات از سال ۱۳۹۶ تا ۱۳۹۸ عضو هیأت علمی دانشگاه پیام نور یزد بودند. در حال حاضر ایشان در حال گذراندن دوره پسادکتری در دانشگاه ماساریک در جمهوری چک می‌باشند. زمینه‌های علمی مورد علاقه نامبرده عبارتند از: بازیابی هوشمند اطلاعات، داده‌های حجیم، محاسبات ابری و محاسبات با کارایی بالا.