استخراج پیکرهٔ موازی از اسناد قابل مقایسه برای بهبود کیفیت ترجمه در سامانه‌های ترجمهٔ ماشینی

زینب رحیمی، محمدحسین ثمنی و شهرام خدیوی
دانشکدهٔ مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

امروزه با گسترش وسایل ارتباط عمومی و به‌خصوص شبکهٔ جهانی اینترنت، نیاز به عملیات ترجمهٔ خودکار به‌صورت چشم‌گیری افزایش یافته است. یکی از مطرح‌ترین روش‌های ترجمهٔ ماشینی، روش آماری است. پارامترهای سامانهٔ[1] ترجمهٔ ماشینی آماری با استفاده از مجموعهٔ بزرگی از دادگان آموزشی (پیکرهٔ موازی دوزبانه) تخمین زده می‌شود؛ اما در برخی زبان‌ها، هنوز مسألهٔ نیاز پایه‌ای سامانهٔ ترجمه ماشینی آماری یعنی پیکره‌های متنی بزرگ موازی برطرف نشده است. برای رفع این مشکل روشی پیشنهادی جهت بهبود کیفیت پیکره‌های مستخرج از اسناد قابل‌مقایسه و درنتیجه بهبود کیفیت سامانهٔ ترجمه ماشینی ارائه شده است. از آنجایی که در اکثر متون قابل‌مقایسه داده‌های موازی نه به‌صورت جمله، بلکه به‌صورت قطعات زیرجمله‌ای ظاهر می‌شوند، روش پیشنهادی سعی در استخراج قطعات موازی به‌صورت بلوک با استفاده از مجموعه‌ای از ویژگی‌ها دارد که این ویژگی‌ها عبارت‌اند از طول عبارت، امتیاز شباهت لگاریتمی، شیب مسیر ترازبندی در بلوک، پراکندگی شیب قطعات تشکیل‌دهندهٔ بلوک، مربعی‌بودن بلوک و درصد حضور کلمات هم‌ترجمه در بلوک. طبق ارزیابی‌های انجام‌شده روش پیشنهادی کارایی مناسبی دارد؛ و علاوه‌بر اینکه از نظر دقت و بازخوانی از روش‌های موجود استخراج قطعه پیشی گرفته است، دادگان مستخرج از اجرای این روش روی، بخشی از پیکرهٔ قابل‌مقایسه موجود، کارایی سامانهٔ ترجمه ماشینی پایه را برای دادگان آزمون مختلف از ۰/۳۳ تا ۱/۴ واحد بلو افزایش داده است.

واژگان کلیدی: پیکرهٔ قابل‌مقایسه[2]، استخراج قطعات موازی[3]، پیکرهٔ موازی[4]، ترجمهٔ ماشینی[5]

۱- مقدّمه

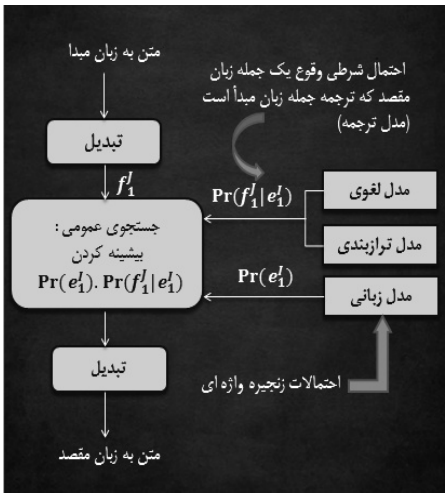
با توجه به حجم روزافزون اطلاعات و اهمیت پردازش خودکار گفتار و اسناد توسط رایانه، می‌توان به اهمیت مدل‌سازی زبان و انجام عملیات ترجمه توسط ماشین پی برد. در زمینهٔ ترجمه ماشینی دو روی‌کرد اصلی مبتنی بر قانون و روش آماری وجود دارند که روش مورد نظر ما در این پژوهش روش آماری است؛ یعنی استفاده از مدل‌سازی آماری برای ترجمهٔ یک متن از یک زبان به زبانی دیگر. روی‌کرد آماری ترجمهٔ ماشینی یکی از روش‌های اصلی در میان روش‌های مبتنی بر پیکره نیز هست؛ چون سامانهٔ طراحی‌شده، تمام اطلاعات مورد نیاز خود را از یک پیکرهٔ موازی دوزبانهٔ متشکل از مجموعهٔ بسیار بزرگی از جملات هم‌ترجمه استخراج می‌کند.

مبنای عملکرد یک مترجم آماری، نظریهٔ تصمیم آماری است. نظریهٔ تصمیم آماری روش شناخته‌شده‌ای برای ساخت یک سامانهٔ تصمیم‌گیری مرکب، از چندین منبع اطلاعاتی موجود، با هدف حداقل‌سازی خطای تصمیم‌گیری است. در ترجمهٔ ماشینی آماری، بر آنیم تا احتمال ترجمهٔ $\Pr(\mathbf{f}_1^J|\mathbf{e}_1^I)$ را به‌نحوی مدل کنیم که رابطهٔ بین رشتهٔ زبان مبدأ $\mathbf{f}_1^J$ و رشتهٔ زبان مقصد $\mathbf{e}_1^I$ را بیان کند. شکل (۱-۱) روند کلی ترجمه را در سامانهٔ آماری نشان می‌دهد که در آن I و J به‌ترتیب تعداد کلمات رشتهٔ زبان مبدأ و مقصد هستند. پارامترهای چنین سامانه‌ای با استفاده از مجموعهٔ بزرگی از دادگان آموزشی (پیکرهٔ موازی دوزبانه) تخمین زده می‌شود.

[1] System
[2] Comparable corpus
[3] Parallel fragments extraction
[4] Parallel corpora
[5] Machine translation

این روش مدل‌هایی را با داده‌های آموزشی بسیار زیاد آموزش می‌دهد و بنابراین قدرت محاسباتی زیادی احتیاج دارد. پارامترهای آماری مورد نیاز برای مدل آماری به‌صورت خودکار از توزیع کلمات در دادگان متنی بزرگ استخراج و از جفت جملات موازی ورودی و خروجی یاد گرفته می‌شوند. درواقع یک مدل آماری، با توجه به ویژگی‌های متفاوت هم‌وقوعی و غیرهم‌وقوعی یک کلمهٔ خاص با دیگر کلمات در زبان دیگر و با توجه به یک پیکرهٔ آموزشی دوزبانه یاد گرفته می‌شود.

در سال‌های اخیر برای برخی زبان‌ها از روش آماری استفاده شده است؛ اما در برخی زبان‌ها هنوز مسأله نیاز پایه‌ای سامانه ترجمهٔ ماشینی آماری یعنی پیکره‌های متنی بزرگ موازی برطرف نشده است.



(شکل ۱-۱): روند کلی ترجمه در سامانهٔ ترجمه ماشینی آماری

روش آماری به‌کمک سه منبع، داده‌های آموزشی خود را به‌دست می‌آورد. این سه منبع پیکره‌های متنی موازی، غیرموازی و قابل‌مقایسه بوده که جزء تفکیک‌ناپذیری از سامانه‌های ترجمهٔ آماری پیکره‌های موازی هستند. پیکره‌های موازی منبعی ایده‌آل برای ترجمه هستند؛ اما به‌خاطر کمبود این متون از اسناد قابل‌مقایسه برای استخراج متون موازی استفاده می‌کنیم؛ حتی با وجود اینکه این متون برای زبان‌های خاصی موجود هستند، ممکن است در دامنهٔ خاصی موجود نباشند و استفاده از متون موازی با زمینهٔ متفاوت برای آموزش سامانهٔ ترجمه، از کارایی سامانهٔ ترجمه ماشینی به‌شدت خواهد کاست؛ پس به‌ناچار از متون غیرموازی و قابل‌مقایسه در کنار آنها استفاده می‌کنیم. از آنجایی که پیکره‌های موازی در آموزش سامانه‌های ترجمهٔ ماشینی نقش اساسی دارند، بر آنیم تا این پیکره‌ها را با استفاده از اسناد قابل‌مقایسه ایجاد کنیم.

منظور از پیکره‌های قابل‌مقایسه، اسنادی با موضوعات مشابه در دو زبان متفاوت است که با توجه به رشد و گسترش روزافزون وسایل ارتباط عمومی و اطلاع‌رسانی، به‌خصوص اینترنت، در حجم زیادی موجود هستند. به‌عنوان مثال، اخبار منتشرشده از وقایع مشترک در زبان‌های مختلف، نمونه‌ای از این نوع پیکره‌هاست. مزیت این‌گونه منابع دردسترس بودن آنهاست که موجب شده تا ایجاد و استفاده از پیکره‌های قابل‌مقایسه به‌عنوان زمینهٔ ارزشمندی در بازیابی اطلاعات بین زبانی به‌خصوص در وب مطرح شود.

برای تهیهٔ پیکرهٔ موازی دو رویکرد عمدهٔ روش دستی و خودکار وجود دارد. در این راستا ابتدا باید متونی که با نیازها مطابقت داشته باشند، را جمع‌آوری کرد؛ گاهی لازم است داده‌های اولیه جمع‌آوری شده و به‌صورت دستی وارد رایانه شوند. اگر داده‌های مورد نیاز آماده نباشند، باید متون نوشتاری را به متون الکترونیکی تبدیل کرد. این کار بسیار زمان‌بر است و انجام آن به‌صورت دستی و یا با استفاده از نرم‌افزار OCR به‌صرفه نیست. OCR با پویش‌کردن و تشخیص کلمات این کار را انجام می‌دهد. از سوی دیگر یک‌سری داده‌های از قبل منتشرشده، از قبیل متن‌های اداری و خبری موجود هستند و این‌گونه داده‌ها مانند صفحات وب، صفحات شخصی افراد و یا شرکت‌ها، فهرست سخنرانی‌ها در اینترنت در حجم زیاد قابل استفاده هستند. بدیهی است که استفاده از روش‌های خودکار برای انجام فرآیند ایجاد پیکرهٔ موازی یا قابل‌مقایسه و در گام بعدی یعنی استخراج دادگان موازی، مناسب‌ترین روش است.

پس از جمع‌آوری متون، برای یافتن جملات موازی (هم‌ترجمه) در آنها جستجو انجام می‌گیرد و جملات موازی متناظر به یکدیگر تراز می‌شوند. روش‌های مختلفی برای این ترازبندی وجود دارد. یک روش خوب باید سریع و دقیق باشد، به زبان وابسته نباشد و احتیاج به دانش خاصی در مورد دو زبان نداشته باشد. اگر سندها ترازبندی نشده باشند و ترازبندی در حد سند موجود نباشد، نیاز به توسعهٔ روش‌هایی است که این ترازبندی را انجام دهند.

روش‌های ترازبندی در سطوح مختلفی مانند کلمه، قطعه، جمله، بند و سند (برای متون قابل‌مقایسه) انجام می‌شوند. در ترازبندی براساس کلمه، از مدل‌های آماری استفاده می‌شود. در ترازبندی در سطح قطعات، متن به قطعاتی تقسیم می‌شود که اگر با روش احتمالاتی، ترجمه

یکدیگر تشخیص داده شدند، جفت‌قطعه نامیده می‌شوند. این کار (ترازبندی زیر جملات (قطعات)، به جای ترازبندی جملات) که در اسناد موازی انجام می‌شود، حافظه مورد استفاده را کاهش می‌دهد و کیفیت فرآیند ترازبندی لغات را بهبود می‌بخشد. بندها با استفاده از شباهت لغوی یا خوشه‌بندی مبتنی بر محتوا ترازبندی و به‌طور کلی در ترازکردن اسناد نیز از کلمات مشترک آنها استفاده می‌شود.

گفتیم برای تهیهٔ پیکرهٔ موازی دو رویکرد عمدهٔ دستی و خودکار وجود دارد؛ همچنین به معرفی پیکره های قابل‌مقایسه و حجم قابل‌توجه آن در وب پرداختیم. در اینجا سعی شده تا روشی برای بهبود کیفیت پیکره های مستخرج از اسناد قابل مقایسه و درنتیجه بهبود کیفیت سامانهٔ ترجمهٔ ماشینی ارائه شود. از آنجایی که در اکثر متون قابل مقایسه داده های موازی، نه به‌صورت جمله، بلکه به‌صورت قطعات زیرجمله‌ای ظاهر می‌شوند، روش پیشنهادی سعی در استخراج قطعات موازی به‌صورت بلوک با استفاده از مجموعه‌ای از ویژگی‌ها دارد.

در ادامه در بخش دوم به معرفی اهم کارهای انجام‌شده در این زمینه پرداخته می‌شود. در بخش سوم به معرفی روش پایه پیاده‌سازی‌شده پرداخته شده و روش جدید پیشنهادی معرفی و جزییات آن بیان می شود. در بخش ارزیابی به ارزیابی روش‌های پیاده‌سازی‌شده و تحلیل نتایج پرداخته می‌شود و درنهایت جمع‌بندی و نتیجه‌گیری انجام می‌شود.

۲- کارهای انجام‌شده

برای استخراج پیکرهٔ موازی از اسناد موازی، روش‌های مشهوری مانند روش (گیل و چرچ، ۱۹۹۱) وجود دارند که در آن ترازبندی با طول جمله شباهت لغوی (مدل آی‌بی‌ام۱[1]) انجام می‌شود؛ روش (مور، ۲۰۰۲) که از تطابق کلمات و طول جملات برای ترازبندی و روش (ملامد، ۱۹۹۹) که برای پیداکردن ترازبندی‌های بین دو متن، از خصوصیات هندسی آنها استفاده می‌کنند؛ اما در زمینهٔ پیکره‌های قابل‌مقایسه مطالعات محدودتر است. بخش عمده‌ای از کارهایی که با پیکره‌های قابل‌مقایسه انجام شده‌اند روی استخراج ترجمهٔ کلمه متمرکز هستند.
(فونگ و یی، ۱۹۹۸؛ دیاب و فینچ، ۲۰۰۰؛ کوهن و نایت، ۲۰۰۴؛ گاوسیر و همکاران، ۲۰۰۴)

از دیگر فعالیت‌های پژوهشی مربوط به این زمینه کار *رزنیک* و *اسمیت* در سال ۲۰۰۳ (رزنیک و اسمیت، ۲۰۰۳) است که سامانه آنها برای کشف جفت‌اسناد موازی در وب طراحی شده است.

اولین تلاش در زمینهٔ کاوش پیکره‌های قابل‌مقایسه برای استخراج جملات موازی توسط *ژائو* و *وگل* در سال ۲۰۰۲ انجام شد که از گسترش الگوریتم‌های طراحی‌شده برای ترازبندی جملات متون موازی استفاده کرده و با استفاده از برنامه‌نویسی پویا اسنادی را که فرض می‌شود، مشابه‌اند ترازبندی کردند. این روش‌ها تنها برای "اسناد موازی و نوفه‌ای" مناسب است؛ یعنی برای اسنادی که به‌نسبه مشابه هستند، هم در محتوا و هم ترتیب جملات قابل اجرا می‌باشد.

در کار (بارزیلای و الحداد، ۲۰۰۳) ترازبندی جملات با استفاده از ویژگی شباهت ترجمه انجام شده است. به این صورت که برای ترازکردن جملات، شباهت لغوی آنها را در نظر گرفته است. تحقیق گزارش‌شده توسط *فونگ* و *چونگ* (فونگ و چونگ، ۲۰۰۴) و (فونگ و وو، ۲۰۰۵) نیز برای اسناد غیرموازی عمل می‌کند. آنها نیز هر سند مبدأ را با چند سند مقصد متناظر کرده و تمام جفت‌جمله‌های ممکن را بررسی می‌کنند؛ اما فهرست ثابتی از جفت‌اسناد وجود ندارد. پس از یک دور استخراج جملات، فهرست با اسناد بیشتر کامل شده و سامانه به‌صورت تکراری اجرا می‌شود. *مانتینیو* و *مارکو* در پژوهش سال ۲۰۰۵ خود (مانتینیو و مارکو، ۲۰۰۵) استفاده از بیش‌ترین بی‌نظمی را برای انتخاب جملات هم‌تراز استفاده کردند؛ بدین شکل که جفت‌جملات را جدا از مضمون جمله تجزیه و تحلیل کرده و آنها را به‌عنوان موازی و غیر موازی طبقه‌بندی کردند. آنها هر سند مبدأ را با چندین سند مقصد مطابقت داده و همهٔ جفت‌جملات ممکن را از هر جفت سند طبقه‌بندی کردند. این کار آنها را قادر می‌سازد که جملات از اسناد به‌نسبه غیرمشابه پیدا کنند و هر مقدار از تغییر ترتیب جملات قابل‌تحمل باشد.

در روش اکس ما (اکس ما، ۲۰۰۶) یک مدل رتبه‌بندی‌کنندهٔ آموزش می‌بیند که برای هر جمله در سند مبدأ یک جملهٔ موازی در سند مقصد انتخاب می‌کند یا تهی را به آن نسبت می‌دهد. این فرمول‌سازی از مشکل عدم توازن طبقه‌بندی‌کنندهٔ دودویی جلوگیری می‌کند و در هر دو روش طبقه‌بندی‌کنندهٔ دودویی و رویکرد رتبه‌بندی‌کننده، از طبقه‌بندی‌کنندهٔ بیش‌ترین بی‌نظمی استفاده می‌شود.

[1] IBM 1

روش (تیلمن، ۲۰۰۹) درواقع توسعه‌ای برای روش ارائه‌شده توسط مانتینیو است که در آن برای جملهٔ داده‌شده، یک طبقه‌بندی‌کنندهٔ بیش‌ترین آنتروپی به مجموعه بزرگی از ترجمه‌های نامزد اعمال می‌شود. یک الگوریتم جستجوی شعاعی برای این کار استفاده می‌شود که جملات مقصدی که خارج از شعاع[1] می‌افتند هر چه سریع‌تر غیرموازی تشخیص داده شوند. در روش (عبدالرئوف و همکاران، ۲۰۰۹) آزمایش‌ها با استفاده از فیلتر جدید نرخ خطای ترجمهٔ امتیازی در کنار فیلترهای نرخ خطای ترجمه و نرخ خطای کلمه انجام شده است و در آن از یک سامانهٔ ترجمهٔ ماشینی آماری استفاده شده که از دادگان موازی کمی برای ترجمه سمت مبدأ استفاده می‌کند و متون مقصد در قسمت مدل زبانی سامانه استفاده می‌شوند. پس از آن از تکنیک‌های بازیابی اطلاعات و فیلترهای ساده برای ایجاد دادهٔ موازی از پیکرهٔ خبری قابل مقایسه استفاده می‌شود. روش (تیلمن و ژو، ۲۰۰۹) روش ساده‌ای است؛ به این صورت که مجموعهٔ بزرگی از جملات نامزد در سطح جمله به‌طور مستقیم امتیازدهی می‌شوند. در این روش استخراج در سطح جملات، وابسته به پیاده‌سازی کارآمدی از تابع امتیازدهی متقارن ساده است که منجر به افزایش سرعت محاسبه با عامل ۳۰ شده است.

در روش (لی و همکاران، ۲۰۱۰) یک مدل ترکیبی بدون ناظر معرفی می‌شود که ویژگی‌های آماری، لغوی، زبانی، مفهومی و زمانی را در یک چارچوب مبنی بر حداکثرسازی انتظار معمولی را ترکیب می‌کند تا واژگان دوزبانه را از پیکرهٔ قابل مقایسه استخراج کند.

یکی از محدودیت‌های همهٔ این روش‌ها این است که آنها برای پیداکردن جملات کامل طراحی شده‌اند. روش ما این است که چون جملات به‌طور کامل موازی در اسناد قابل‌مقایسه کم هستند، زیرجملات موازی را تشخیص دهیم. کارهای زیادی در این زمینه انجام نشده است. کار اصلی انجام‌شده در زمینهٔ استخراج قطعهٔ سامانهٔ طراحی‌شده توسط مانتینیو و مارکو در سال ۲۰۰۶ است (مانتینیو و مارکو، ۲۰۰۶) که با تجزیه و تحلیل جفت‌جملات به‌طور بالقوه مشابه، با استفاده از روشی الهام‌گرفته از پردازش سیگنال، تشخیص می‌دهد که کدام بخش از جملهٔ مبدأ به کدام بخش از جملهٔ مقصد مرتبط است.

کاردنگ و همکاران (دنگ و همکاران، ۲۰۰۶) نیز به قطعات زیرجمله می‌پردازد؛ اما آنها قطعات موازی را از جفت‌جمله‌های موازی استخراج می‌کنند؛ درحالی‌که ما آنها را از جفت‌جملات قابل‌مقایسه و یا غیر موازی به‌دست می‌آوریم. در روش (هواویتارانا، ۲۰۱۰) از سه روش ترازبندی عبارات برای یافتن جفت عبارات استفاده می‌شود: الگوریتم استاندارد استخراج عبارات (ابزار موزس و با استفاده از مکاشفه)، استفاده تنها از ویژگی‌های نحوی که وابسته به مسیر ویتربی نیست و طبقه‌کنندهٔ دودویی با استفاده از روش بی‌نظمی بیشینه.

در زمینهٔ استخراج متون موازی از اسناد قابل‌مقایسه برای زبان فارسی کارهای شاخص زیادی انجام یا منتشر نشده است. یکی از محدود کارهای انجام‌شده، توسعهٔ پیکرهٔ تطبیقی UTPECC (هاشمی و همکاران، ۲۰۱۰) است که از تکنیک‌های بازیابی اطلاعات استفاده کرده و با تکیه بر تاریخ انتشار اخبار و شباهت محتوای اسناد عمل می‌کند. کار دیگر مرتبط (محمدی و قاسم آقایی، ۲۰۱۰) است که در این مقاله یک پیکرهٔ متنی موازی و ترازشده برای جفت زبان های فارسی و انگلیسی با کاوش در محتویات ویکی‌پدیا ارائه می‌شود. در این پژوهش روشی برای ترازبندی مبتنی بر لغت‌نامه دوزبانه در سطح جمله ارائه شده که از روشی بهبودیافته مبتنی بر ابرپیوند استفاده می‌کند.

با توجه به اطلاعات بالقوه پیکره‌های قابل‌مقایسه (حداقل به‌صورت عبارات زیرجمله‌ای) و اینکه کار زیادی در این زمینه برای زبان فارسی انجام نشده است روش پیشنهادی ارائه شد تا در این زمینه راه‌گشا باشد.

۳- معرفی روش پایه و پیشنهادی

اسناد قابل‌مقایسه موضوع یکسانی دارند؛ اما ترجمهٔ دقیق هم نیستند. از اسناد قابل‌مقایسه اطلاعات مختلفی استخراج می‌شود که شامل استخراج اسناد موازی، جملات موازی و زیرجملات موازی است. به‌طور معمول پیکره‌های خیلی غیرموازی دارای زوج‌جملات خوب نیستند؛ یا دارای تعداد اندکی از جملات مناسب هستند و اکثر دادهٔ موازی آنها در سطح زیرجمله موجود است. در این بخش روش پایهٔ استفاده‌شده برای پروژه و روش جدید پیشنهادی معرفی می‌شود.

۳-۱- ترازبندی اسناد

برای برخی اسناد مانند اسناد استخراج‌شده از سایت khamenei.ir به‌دلیل عدم ارتباط میان اسناد فارسی و انگلیسی و نبود رابطهٔ معین، لازم است تا ابتدا ترازبندی در

[1] beam

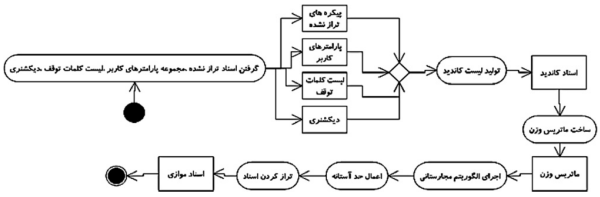
سطح اسناد صورت گیرد. برای این منظور از روشی مبتنی‌بر الگوریتم مجارستانی[1] استفاده شده و از طول و ارتباط معنایی کلمات به‌عنوان ویژگی استفاده شده است.

الگوریتم مجارستانی یک الگوریتم بهینه‌سازی ترکیبی است که مسألهٔ تخصیص را در زمان چندجمله‌ای حل می‌کند. ورودی الگوریتم یک ماتریس غیر منفی است که درایه‌های سطر iام و ستون jام آن در حالت کلی نمایندهٔ هزینهٔ تخصیص کار jام به فرد iام است. هدف پیداکردن تخصیص مناسب کارها با بهینه‌ترین هزینه به افراد است. در مورد این ابزار ما با استفاده از این الگوریتم، با دادن ماتریس وزن به‌عنوان ورودی، مناسب‌ترین اسناد را به هم نگاشت می‌کنیم. پیاده‌سازی این الگوریتم جزییات زیادی دارد، اما گام‌های آن به‌طور خلاصه برای یک ماتریس m*m به شرح زیر است:

- کوچک‌ترین عنصر هر سطر یا ستون را از تمام عناصر آن سطر یا ستون کم می‌کنیم.
- با کمترین تعداد خط صفرهای جدول را می‌پوشانیم. اگر تعداد خطوط پوششی برابر m شد ماتریس بهینه است.
- در غیر این صورت کوچک‌ترین عددی را که روی آن خط کشیده نشده از اعدادی که روی آنها خط کشیده نشده کم و به محل تقاطع خطوط مرحله دو اضافه می‌کنیم و دوباره صفرهای جدید را با خط پوششی، می‌پوشانیم تا به m خط برسیم.
- مراحل ۲ و ۳ را تا رسیدن به m خط ادامه می‌دهیم. (کوهن، ۱۹۹۵)

همچنین در برنامه از دو فایل لغت‌نامهٔ فارسی به انگلیسی و بالعکس استفاده می‌شود.

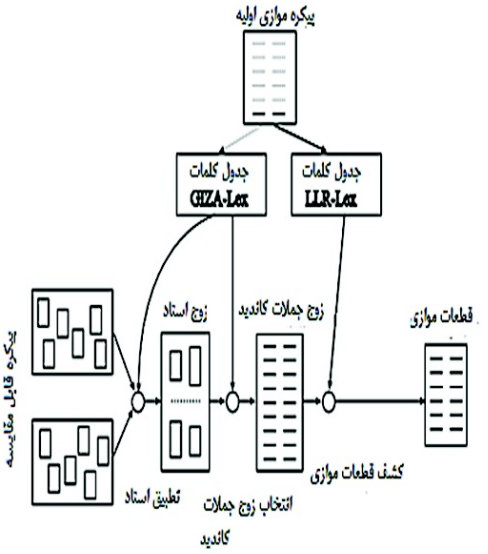
نمودار جعبه‌ای نحوه عملکرد این برنامه به‌صورت زیر است:



(شکل۳-۱): نمودار فرآیند ابزار Document Aligner

۳-۲- روش پایه

به‌عنوان روش پایه از مهم‌ترین کار انجام‌شده در زمینه استخراج قطعات زیرجمله‌ای موازی یعنی روش (مانتینیو و مارکو، ۲۰۰۶) استفاده می‌شود. در این قسمت جزییات و فرآیند پیاده‌سازی آن بیان می‌شود. فرآیند کلی برنامه در نمودار جعبه‌ای زیر مشخص است:



(شکل ۳-۲): نمودار جعبه‌ای روش پایه (مانتینیو و مارکو، ۲۰۰۶)

- در نخستین مرحله زوج اسناد مشابه با استفاده از ابزار بازیابی اطلاعات Lemur تشخیص داده می‌شود. به این صورت که سند مبدأ کلمه‌به‌کلمه ترجمه‌شده و به یک پرس‌وجو تبدیل می‌شود که بر روی اسناد زبان مقصد، اجرا می‌شود و بیست نتیجهٔ بالاتر بازیابی می‌شوند. زوج‌جملات انتخاب‌شده از اسناد جفت‌شده به‌عنوان ورودی مرحلهٔ بعدی استفاده می‌شوند.
- در مرحلهٔ دوم زوج‌جملات این زوج اسناد به فیلتر انتخاب نامزد ارسال می‌شوند. در این مرحله زوج‌جمله‌هایی که تعداد کمی کلمهٔ ترجمه‌شده دارند، حذف شده و روش کشف قطعات برای زوج جملات باقی‌مانده، اعمال می‌شود که نتیجهٔ آن خروجی سامانه را تشکیل می‌دهد.
- در این روش از دو فرهنگ‌لغت دوزبانه استفاده می‌شود: جدول کلمات جیزا و جدول شباهت لگاریتمی. جدول جیزا از اجرای نرم‌افزار جیزا++[2] (پیاده‌سازی مدل‌های ترازبندی کلمات آی‌بی‌ام) روی یک پیکرهٔ موازی به‌دست می‌آید که یکی از ویژگی‌های آن این است که هر کلمه

[1] hungrian

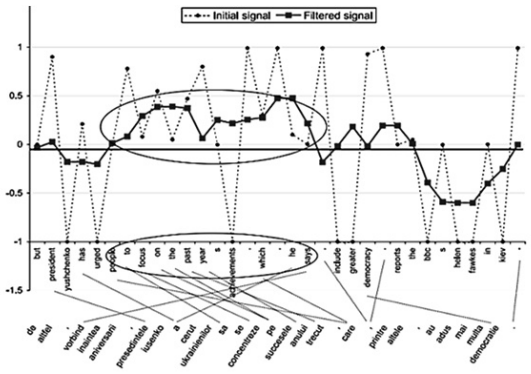
[2] Giza++

ممکن است، چندین معنا داشته باشد. جدول شباهت لگاریتمی شامل امتیازات جفت کلمات بر اساس نسبت خطی لگاریتمی[1] یا همان LLR آنها است که با استفاده از فرمول ذکرشده در (مور، ۲۰۰۲) و روند زیر محاسبه می‌شود:

- روی پیکرهٔ موازی ترازشده، کلمات جیزا++ را در دو جهت اجرا می‌کنیم و هم‌ترازی‌ها را با استفاده از روش‌های مکاشفه‌ای بهبودیافته متقارن می‌کنیم.
- همهٔ امتیازات LLR را محاسبه می‌کنیم. برای هر جفت کلمه که حداقل یک اتصال در پیکرهٔ هم‌ترازشده دارد، یک امتیاز LLR وجود دارد.
- همه LLR(e,f)ها را به دسته‌های مثبت (p(e,f)>p(e).p(f)) و منفی (ارتباط منفی) تقسیم می‌کنیم.
- برای هر f عامل نرمال‌سازی $\sum_e LLR^+(e,f)$ یا $\sum_e LLR^-(e,f)$ را محاسبه می‌کنیم.
- همه عبارات $LLR^+(e,f)$ را به عامل‌های نرمال‌سازی مربوطه تقسیم می‌کنیم تا $p^+(e|f)$ را به دست آوریم.
- همهٔ عبارات $LLR^-(e,f)$ را به عامل‌های نرمال‌سازی مربوطه تقسیم می‌کنیم تا $p^-(e|f)$ را به‌دست آوریم.

دلیل استفاده از مفهوم دسته‌های مثبت و منفی این است که با توجه به فرهنگ‌لغت احتمالاتی و رابطهٔ بالا تشخیص داده شود که با احتمال زیاد یک‌جفت کلمه، ترجمهٔ هم هستند یا خیر، اگر باشند مثبت و در غیر این صورت منفی دسته‌بندی می‌شوند.

پس از محاسبهٔ جدول شباهت لگاریتمی، برای هر جفت‌جملهٔ نامزد به‌صورت حریصانه با توجه به امتیاز LLR ترازبندی صورت گرفته و مقدار آن در نمودار قرار می‌گیرد و سیگنال تشکیل می‌شود. قرارگیری مقادیر در نمودار به این صورت است که برای هر جفت‌کلمه اگر p^+ وجود داشته باشد، یعنی تخصیص مثبتی بین جفت‌های حاصل از کلمهٔ مبدأ و هر یک از کلمات جملهٔ مقصد وجود خواهد داشت و بالاترین مقدار P^+ها در نمودار قرار می‌گیرد و اگر P^+ وجود نداشته باشد و همهٔ تخصیص‌ها منفی باشد، کمترین مقدار منفی‌ها در نمودار قرار می‌گیرد. این امر به این دلیل است که کلمه‌ای که هیچ یک از تخصیص‌هایش مثبت نیستند، به‌حتم ترجمهٔ بهتری دارد که در این جمله نیست و نباید ترازبندی برایش صورت بگیرد. پس کم‌ترین عدد در بین p^- ها برایش انتخاب می‌شود تا در نرمال‌سازی حذف شود. اگر هم کلمه‌ای در جملهٔ مقصد هیچ‌گونه ترجمه‌ای، حتی با p^- در جمله مقصد نداشت، مقدار آن در نمودار ۱- قرار می‌گیرد. پس از این مرحله بین مقادیر همسایه، میانگین‌گیری صورت می‌گیرد (هر مقدار با دو مقدار مجاورش) و مقادیر نرمال‌سازی می‌شوند. پس از نرمال‌سازی، بخش‌هایی از نمودار که مقادیر متوالی مثبت دارند، به‌عنوان قطعه استخراج می‌شوند. شکل (۳-۳) نمودار سیگنال قبل و بعد از فیلترشدن و قطعه استخراج‌شده را نشان می‌دهد.



(شکل۳-۳): نمودار سیگنال برای یک جفت‌جمله (مانتینیو و مارکو، ۲۰۰۶)

۳-۳- روش پیشنهادی

۳-۳-۱- توضیح کلی الگوریتم

روش پیشنهادی، روشی برای استخراج قطعات موازی از پیکرهٔ قابل‌مقایسه است. در این روش نیز همانند روش پایه از دو فرهنگ‌لغت استفاده می‌شود که نخستین‌شان فرهنگ‌لغت ساده برای تعیین جملات نامزد و دیگری فرهنگ‌لغت مبتنی بر LLR است؛ که در قسمت تعیین قطعات استفاده می‌شود. فرهنگ‌لغت مبتنی بر LLR با استفاده از روش ذکرشده در روش پایه ایجاد می‌شود.

در مرحلهٔ نخست بر روی متون ورودی پیش‌پردازش انجام می‌شود. در بخش پیش‌پردازش، زیربخش‌های توکن‌بندی، نرمال‌سازی و تعیین مرز جملات صورت می‌گیرد. بدین شکل که نویسه‌های متن ورودی یکسان‌سازی، فواصل و خطوط اضافی حذف و در مرحلهٔ بعدی به واحدهای جمله شکسته می‌شوند. در این بخش مرز جمله‌ها با استفاده از علایم جداکنندهٔ جمله شامل («.»، «!»، «؟» و «?») و مرز کلمه‌ها با استفاده از علایم فضای خالی،

[1] Log Linear Ratio

«،»، «.»، «<>»، «><»، «[]»، «-»، «»، «/» و خط جدید مشخص می‌شود.

پس از این مرحله مجموعه جملات نامزد مشخص می‌شوند. بدین صورت که امتیاز تمامی جفت‌جمله‌ها با استفاده از فراوانی جفت‌کلمات هم‌ترجمه موجود در فرهنگ واژگان به‌دست می‌آید و بالاترین امتیازات به‌عنوان جفت‌جمله‌های نامزد انتخاب می‌شوند. شکل (۳-۴) فرآیند بخش انتخاب جملات نامزد را نشان می‌دهد.

در روش پیشنهادی برای درنظرگرفتن حالت‌های مختلف ترازبندی، از ماتریس امتیازات استفاده می‌کنیم. به این صورت که برای هر جفت‌جملهٔ نامزد، ماتریس امتیازات LLR متناظر جفت کلمات ایجاد می‌گردد؛ سپس به‌صورت نزولی از نقطهٔ دارای بیش‌ترین امتیاز، به‌عنوان نقطهٔ شروع استخراج قطعهٔ آغازشده و پررنگ‌ترین نقاط یعنی نقاط دارای بالاترین امتیاز تا وقتی که از حد آستانه‌ای افت نکرده‌اند، برای استخراج قطعه، یکی پس از دیگری پردازش می‌شوند. این حد آستانه به‌صورت تجربی به‌دست می‌آید؛ چون با پایین‌آوردن آن کیفیت قطعه‌ها پایین می‌آید؛ پس تا زمانی که کیفیت قطعات مطلوب باشد، حد آستانه را پایین می‌آوریم. برای هر نقطهٔ پررنگ درواقع خوشه‌بندی صورت می‌گیرد و هر خوشه نمایان‌گر یک قطعه است.

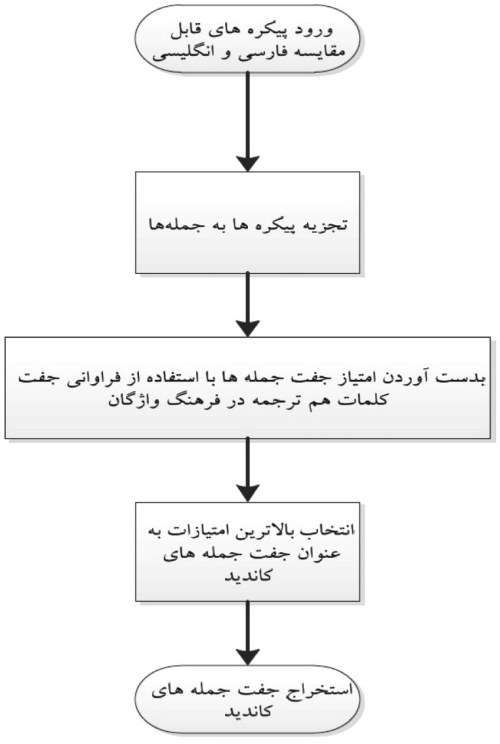
فرآیند پیداکردن قطعه برای هر نقطهٔ پررنگ به شرح زیر است:

- امتیاز قطعه با توجه به ویژگی‌ها و الگوها محاسبه می‌شود. الگوها با توجه به نوع زبان‌های استفاده‌کننده از برنامه (فعلاً فارسی-انگلیسی) و ساختار معمول قطعات در این نوع زبان‌ها که از پیکرهٔ آموزشی یاد گرفته‌شده‌اند، انتخاب شدند. به‌عنوان مثال ساختار قطری با شیب ۱ یا ۱- یکی از ساختارهای بسیار رایج قطعات در زبان فارسی و انگلیسی است. بدین معنی که کلمات یا به‌ترتیب (شیب ۱) و یا به‌طور معکوس (شیب ۱-) با هم متناظرند. ویژگی‌های مورد استفاده نیز در ادامه توضیح داده خواهند شد.
- در همسایگی قطعه (بار نخست قطعه تنها یک نقطه است و همسایگی آن هشت نقطهٔ مجاور آن است) جستجو می‌کنیم و امتیاز کل قطعه را درصورت افزایش هر کلمهٔ همسایه محاسبه کرده و کلمه‌ای که منجر به بالاترین امتیاز شده انتخاب و به مسیر افزوده می‌شود.
- کلمهٔ انتخاب‌شده به کلمه‌ای که x یا y آن کم‌ترین فاصله را با آن دارد، وصل می‌شود (والدش انتخاب می‌شود). این کار به‌دلیل حفظ ساختار قطعه صورت می‌گیرد.
- امتیاز به‌روزرسانی می‌شود.

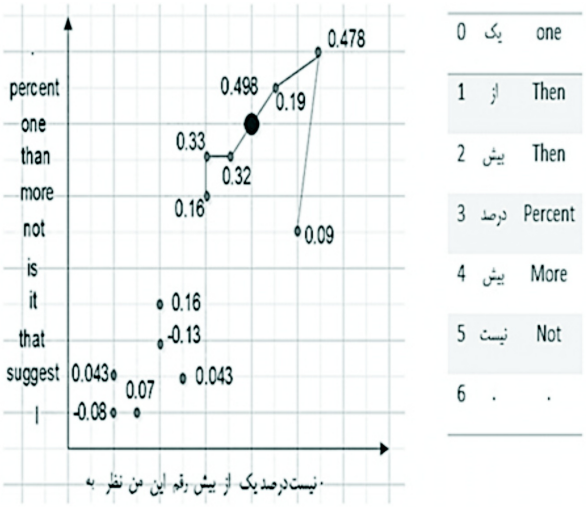
اگر امتیاز بلوک از حد خاصی افت کند، این جستجو خاتمه پیدا می‌کند. همچنین هر قطعه اگر انتخاب شود، کلمات آن در قطعات بعدی لحاظ نخواهند شد (قطعات هم‌پوشانی ندارند). شکل (۳-۵) مثالی از روند اعمال الگوریتم به یک جمله و گسترش بلوک مسیر حول یک نقطهٔ پررنگ را نشان می‌دهد.

شکل (۳-۶) نمودار جعبه‌ای الگوریتم را نشان می‌دهد.

در شکل (۳-۵) مثالی از نحوهٔ عملکرد الگوریتم پیشنهادی آورده شده است. در این شکل یک جفت‌جملهٔ نامزد از پیکرهٔ قابل‌مقایسه انتخاب شده است که روی دو محور قرار گرفته‌اند. همان‌طور که در شکل مشخص است، جفت‌کلمات (یک و one) در مرحلهٔ نخست بالاترین امتیاز LLR را داشته‌اند و بلوک استخراج قطعه حول آن نقطه تشکیل شده است. در هر مرحله مطابق الگوریتم با توجه به امتیازات نقاط همسایهٔ بعدی انتخاب شده و مسیر در بلوک گسترده می‌شود تا درنهایت با گذشتن از حد آستانه جستجو خاتمه یابد.



(شکل ۳-۴): نمودار جعبه‌ای به‌دست آوردن جملات نامزد



not more than one percent . ⟺ بیش از یک درصد نیست .

(شکل ۳-۵): مثالی از شیوهٔ اعمال الگوریتم پیشنهادی

۳-۳-۲- ویژگی‌های استفاده‌شده در روش پیشنهادی

۱- طول قطعه (F1): هر چه طول قطعه بیشتر باشد بهتر است و امتیازها به نسبت طول مقایسه می‌شوند.

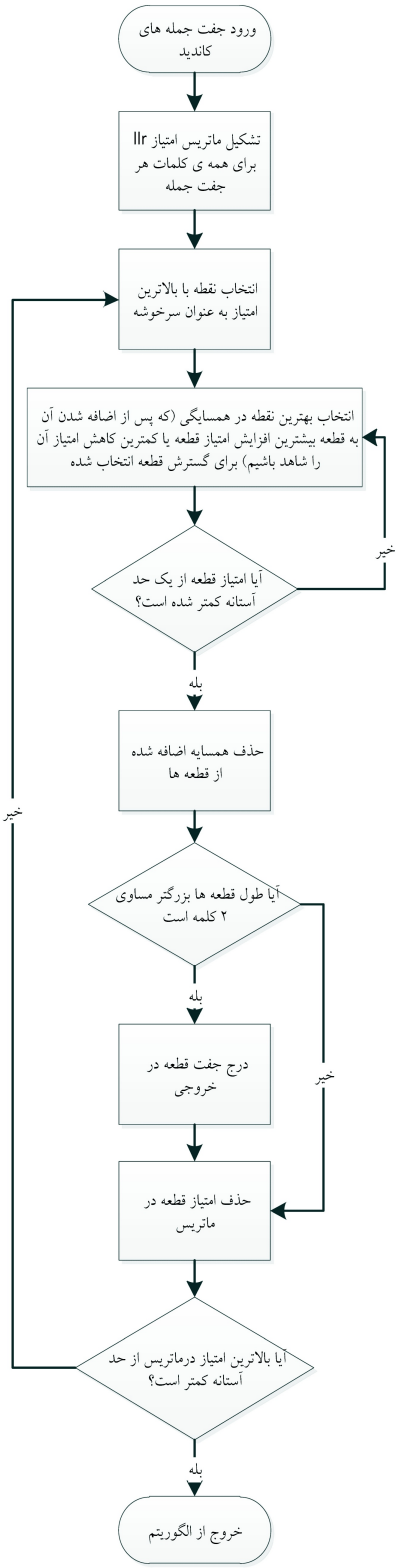
F1=Path Length (۱)
Path Length = تعداد تخصیص‌ها

۲- میانگین LLRها (F2): LLR عامل نخستین است، به هم‌ترجمه‌بودن کلمات قطعه اشاره دارد؛ پس مقدار آن هرچه بیشتر باشد، بهتر است.

$$F2 = \frac{\sum_{i=1}^{N} LLR_i}{N} \quad (۲)$$

۲- طول مسیر در قطعهٔ (F3): بین کم‌ترین و بیش‌ترین اندیس انتخاب‌شده از ماتریس، هرچه تعداد کلمات دارای معادل بیشتر باشد، بهتر است و امتیاز بالاتری به آن تخصیص می‌یابد.

$$F3 = \frac{I}{X} \quad (۳)$$



(شکل ۳-۶): روندنمای الگوریتم پیشنهادی

I: تعداد تخصیص‌ها

X: حداکثر طول بین دو جملهٔ فارسی و انگلیسی

۳-میزان شیب اسکلت قطعه در بلوک (F4): گفتیم که برای هر قطعه اسکلت مسیر تشکیل می‌شود. هرچه قدرمطلق شیب این مسیر به یک نزدیک‌تر باشد، مسیر به بهینه‌ترین الگوی قطعات فارسی-انگلیسی یعنی همان حالت قطری نزدیک‌تر است. در قطعات هم‌ترجمهٔ انگلیسی-فارسی بیش‌ترین الگوی رخداده، حالت قطری با شیب ۱ یا ۱- است. (به‌عنوان مثال ترازبندی کلمات دو عبارت "دوست خوب" و " friend"good در نمودار کلمات آنها که محور افقی عبارت انگلیسی و محور عمودی عبارت فارسی را نشان بدهد، به‌صورت یک نمودار خطی با شیب ۱- است). برای محاسبهٔ این ویژگی، قدرمطلق شیب خطوط بین هر دو نقطه در مسیر قطعه محاسبه می‌شود.

- اگر کوچک‌تر از ۱ است: خودش
- اگر بزرگ‌تر از ۱ است: معکوسش

سپس از مقادیر حاصله میانگین می‌گیریم. هرچه این مقدار به یک نزدیک‌تر باشد، بهتر است. (مقادیر افقی و عمودی نیز حذف می‌شوند)

۴-پراکندگی شیب در خطوط بین نقاط تشکیل‌دهندهٔ مسیر در بلوک قطعه (F5): شکستگی مسیر را بررسی می‌کند. هرچه شکستگی کمتر باشد، مسیر به حالت قطری نزدیک‌تر است (شکستگی بیشتر از یک به‌طور معمول به قطعات مناسبی منجر نمی‌شود).

- جدول توزیع شیب‌ها به دست می‌آید (شیب‌ها و تعداد تکرارشان).
- میزان شیب، عبارت از تعداد درایه‌های جدول تقسیم بر تعداد یال‌های مسیر است.

۵-مربعی‌بودن بلوک قطعه (F6): هرچه بلوک مربعی‌تر باشد، یعنی طول قطعهٔ فارسی و انگلیسی معادل به هم نزدیک‌تر است.

$$F6 = \left|\frac{(maxx-maxy)}{(maxy-miny)}\right| \quad (۴)$$

Maxx: بیشینهٔ اندیس‌ها در قطعهٔ فارسی

Minx: کمینهٔ اندیس‌ها در قطعهٔ فارسی

(Maxy و Miny نیز به همین ترتیب برای قطعه انگلیسی)

اگر مقدار A بزرگ‌تر از یک بود، خودش و اگر کمتر از یک بود معکوسش به‌عنوان امتیاز مربعی‌بودن بلوک در نظر گرفته می‌شود.

این امتیازات طوری تراز شده‌اند که همگی بین ۰ و ۱۰ قرار گیرند و با استفاده از ترکیب خطی به‌صورت وزن‌دار ترکیب‌شده‌اند و امتیاز نهایی بلوک را شکل می‌دهند.

$$Score = \frac{\sum_{i=1}^{6} w_i \cdot f_i}{\sum_{i=1}^{6} w_i} \quad (۵)$$

۳-۳-۳- وزن‌دهی ویژگی‌ها

برای تعیین وزن ویژگی‌ها، یک پیکرهٔ طلایی شامل هشتصد جمله از پیکره‌های central asia، Tehran avenue و wikipedia و ۱۶۳۰ قطعهٔ مشخص‌شده در آنها به‌عنوان مرجع ساخته شد و با استفاده از روش تبرید شبیه‌سازی‌شده[1] (یا تبرید تدریجی فلزات) بهینه‌ترین وزن‌ها تخمین زده شد.

تبرید شبیه‌سازی‌شده، یک روش بهینه‌سازی فراابتکاری ساده و مفید در حل مسائل بهینه‌سازی است. تکنیک تبرید تدریجی، به‌وسیله متالورژیست‌ها برای رسیدن به حالتی که در آن ماده جامد، به‌خوبی مرتب و انرژی آن کمینه شده باشد، استفاده می‌شود. این تکنیک شامل قراردادن ماده در دمای بالا و سپس کم‌کردن تدریجی این دماست. (کرکپاتریک و همکاران، ۱۹۸۳)

در روش شبیه‌سازی تبریدی، هر نقطه در فضای جستجو مشابه یک حالت از یک سامانهٔ فیزیکی است؛ و انرژی داخلی سامانه در آن حالت باید کمینه شود. در این روش، هدف انتقال سامانه از حالت نخستین دلخواه، به حالتی است که سامانه در آن کمترین انرژی را داشته باشد.

برای حل یک مسألهٔ بهینه‌سازی، الگوریتم ابتدا از یک جواب نخستین شروع و سپس در یک حلقه تکرار به جواب‌های همسایه حرکت می‌کند. اگر جواب همسایه بهتر از جواب فعلی باشد، الگوریتم آن را به‌عنوان جواب فعلی قرار می‌دهد (به آن حرکت می‌کند)، در غیر این صورت، الگوریتم آن جواب را با احتمال exp(-ΔE/T) به‌عنوان جواب فعلی می‌پذیرد. در این رابطه ΔE تفاوت بین تابع هدف جواب فعلی و جواب همسایه است و T یک پارامتر به نام دماست. در هر دما، چندین تکرار اجرا و سپس دما به‌آرامی کاهش داده می‌شود. در گام‌های نخستین دما خیلی بالا قرار داده می‌شود تا احتمال بیشتری برای پذیرش جواب‌های بدتر وجود داشته باشد. با کاهش تدریجی دما، در گام‌های پایانی احتمال کم‌تری برای پذیرش جواب‌های بدتر وجود خواهد

[1] Simulated annealing

داشت؛ و بنابراین الگوریتم به سمت یک جواب خوب، همگـرا می‌شود.

در هر مرحله، الگوریتم تبرید شبیه‌سازی‌شده، چند حالت را در همسایگی حالت کنونی در نظر می‌گیرد و به‌طور احتمالی تصمیم می‌گیرد که سامانه را از حالت منتقل کند یا در همین حالت باقی بماند. این احتمالات درنهایت سامانه را به حالت با انرژی کمتر میل می‌دهد.(کرنی، ۱۹۸۵)
نمودار جعبه‌ای این الگوریتم در شکل (۳-۷) آورده شده است.

اعمال الگوریتم تبرید شبیه‌سازی‌شده

برای تعیین بهینه وزن‌های ویژگی‌ها، آنها را بـه‌صـورت یـک بردار شش تایی درنظرگرفتیم که هر بار بـه‌صـورت تصـادفی شش مقدار آن تغییر می‌کنند. به‌صورتی‌که بـه‌احتمـال ۳۳٪ خودش باقی می‌مانند، به احتمال ۳۳٪ ۰/۰۱ افزایش می‌یابد و یا به احتمال ۳۳٪ به میزان ۰/۰۱ کاهش می‌یابد. بـه ایـن صورت به نقطهٔ همسایه حرکت می‌کند.

در این مرحلـه برنامـه فراخـوانی مـی‌شـود و دقـت و بازخوانی با مقایسهٔ با پیکرهٔ طلایی مرجع محاسبه مـی‌شـود. هزینه نیز در پیاده‌سازی انجام‌شده به‌عنوان معکـوس امتیـاز در نظر گرفته مـی‌شـود و ΔE برابـر بـا هزینـه نقطـه جدیـد منهای هزینه نقطه قبلی است.

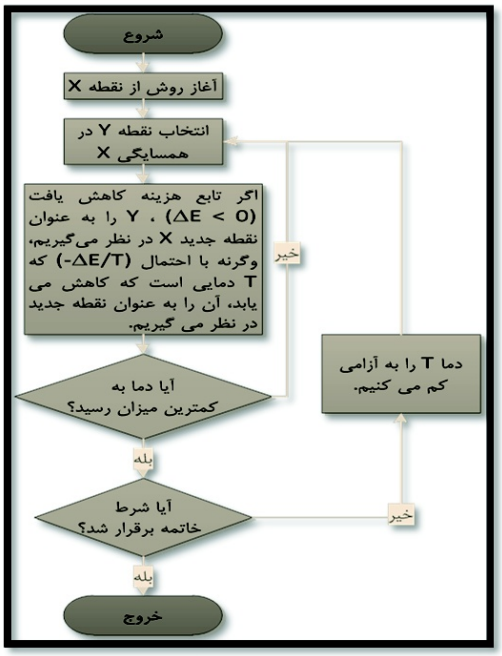
اگر این مقدار کمتر از صـفر بـود، همسـایه بـه‌عنـوان نقطه جدید در نظر گرفته می‌شود و اگر نبود به‌احتمال P که در الگوریتم ذکر شد، نقطهٔ همسـایه بـه‌عنـوان نقطـه جدیـد قبول می‌شود. در هر بار اجرا دما یک درجه کاهش می‌یابد و زمانی که دما صفر شود، جواب برگردانده می‌شود.

۳-۳-۴- نمونه‌ای از خروجی سامانه

در جدول زیر تعدادی از عبارات استخراج‌شده توسط سـامانهٔ پیشنهادی آورده شده است.

۴- ارزیابی

هدف از این بخش ارزیابی سامانهٔ پیشـنهادی در مقایسـه بـا سامانهٔ پایه با استفاده از معیارهای ارزیـابی اسـت. همچنـین بررسی این موضوع که پیکرهٔ موازی استخراج‌شده توسط ابزار پیشنهادی چه تأثیری بر کیفیت سامانهٔ ترجمهٔ ماشینی دارد.



(شکل ۳-۷): دیاگرام بلوکی الگوریتم تبرید شبیه‌سازی‌شده (کرنی، ۱۹۸۵)

(جدول ۳-۱): نمونه‌ای از قطعات استخراج‌شده توسط سامانهٔ پیشنهادی

announced that the government will -> اعلام کرد که دولت
is that many of these -> این است که برخی از این
concert that night was -> که قرار بود همان شب کنسرت supposed
of love and fear and -> از عشق و ترس و
foreign trade -> تجارت خارجی
believes the -> معتقد است
the first step in -> اولین گام در
presidential election -> انتخابات ریاست جمهوری
these works are either -> این آثار یا
limitations in the use -> محدودیت در استفاده
young soldier -> سرباز جوان
is that the index -> این است که فهرست
the country's economic -> و توسعه اقتصادی این کشور development
journalist and analyst -> روزنامه نگار و تحلیلگر
corrupt security forces -> نیروهای امنیتی فاسد
really effective -> واقعاً مؤثر

۴-۱- معیارهای ارزیابی

مشهورترین معیارهای ارزیابی در حوزهٔ ترجمهٔ ماشینی عبارتند از دقت، بازخوانی، معیار F و بلو که ما از این معیارها در ارزیابی سامانه استفاده خواهیم کرد.

۴-۱-۱- معیار F [1]

در حوزهٔ پردازش زبان طبیعی، عبارت معیار F به ترکیبی از دقت و بازخوانی [2] اشاره دارد. این معیار به‌طور عمومی برای ارزیابی سامانه‌های استخراج اطلاعات استفاده می‌شوند. فرض کنید مجموعهٔ نامزد با **Y** و مجموعه مراجع را با **X** نمایش دهیم. در این صورت سه مقدار دقت، بازخوانی و معیار **F** به‌صورت زیر تعریف می‌شوند:

$$recall(Y|X)=\frac{|X\cap Y|}{X} \quad (۶)$$

$$precision(Y|X)=\frac{|X\cap Y|}{Y} \quad (۷)$$

$$F\text{-}measure=\frac{2*precision*recall}{precision+recall} \quad (۸)$$

۴-۱-۲- معیار بلو [3]

معیار بلو برای ارزیابی خودکار ترجمهٔ ماشینی از مقایسهٔ خروجی سامانه با ترجمه‌های صحیح استفاده می‌کند (پاپیننی و روکوس، ۲۰۰۱). بنابراین، به یک پیکرهٔ آزمایشی که حداقل شامل یک ترجمهٔ دستی از هر جمله است نیاز دارد. در زمان آزمایش، هر جملهٔ آزمایشی به سامانهٔ ترجمهٔ ماشینی داده می‌شود و خروجی با مقایسه با ترجمهٔ درست امتیازدهی می‌شود. این امتیاز را امتیاز بلو می‌نامند. جملهٔ خروجی را جملهٔ نامزد [4] و ترجمه‌های درست را مراجع [5] می‌نامند. امتیاز بلو به‌کمک دو عامل دقت [6] و طول جملهٔ نامزد محاسبه می‌شود.

۴-۲- شرایط آزمون

۴-۲-۱- پارامترهای سامانهٔ پیشنهادی

برای آموزش نخستین سامانه و ساخت (جدول کلمات LLR، از فرهنگ لغت احتمالاتی مستخرج از پیکرهٔ PEN استفاده شده است که یک پیکرهٔ موازی خبری است.

(جدول ۴-۱): مشخصات پیکرهٔ موازی و فرهنگ لغت‌های استفاده‌شده

پیکرهٔ موازی اولیه	**PEN** دارای ۳۰۴۷۹ جفت‌جمله موازی
فرهنگ لغت ساده	دارای ۷۱۰۹۰ درایه
(جدول کلمات LLR	دارای ۲۲۹۷۴۷ درایه

همچنین وزن‌های تخمین‌زده‌شده در جدول زیر آورده شده‌اند:

(جدول ۴-۲): وزن‌های تخمین‌زده‌شده برای ویژگی‌ها

وزن‌های تخمینی	ویژگی‌ها
۰/۱۳	طول قطعه
۰/۱۵	میانگین LLR
۰/۴۵	پراکندگی شیب مسیر در بلوک قطعه
۰/۹۲	درصد وجود کلمات هم ترجمه در قطعه
۰/۵۱	مربعی بودن بلوک قطعه
۰/۴۴	میزان شیب مسیر در بلوک قطعه

برای آزمون سامانه حد آستانهٔ مربوط به بخش امتیاز سامانه که با افت‌کردن امتیاز از آن، جستجو برای قطعهٔ پایان می‌یابد، با توجه به محدودهٔ امتیازات (۰-۱۰۰) و وزن‌های به دست آمده، به‌صورت تجربی با مشاهدهٔ کیفیت قطعات حاصل، پنج قرار داده شد.

محاسبهٔ دقت، بازخوانی و معیار F بر مبنای کلمات صورت گرفت. به این صورت که دقت هر قطعه عبارت از تعداد کلمات استخراج‌شدهٔ صحیح نسبت به تعداد کلمات استخراج‌شده و بازخوانی عبارت از تعداد کلمات استخراج‌شدهٔ صحیح نسبت به تعداد کلماتی که مطابق مرجع باید استخراج می‌شده، است. به‌عنوان مثال:

مرجع: گل سرخ تازه => **fresh red flower**

استخراج‌شده توسط سامانه: گل سرخ => **red flower**

دقت = ۱۰۰٪

بازخوانی = ۶۶٪

1 F-measure
2 Recall
3 BLEU
4 Candidate Sentence
5 References
6 Precision

۴-۲-۲- پیکره‌های آزمون

برای آزمایش سامانه یک پیکرهٔ طلایی آزمون شامل ۱۸۰۰ زوج جمله و قطعات مرجع معین در آن با شرایط ذکرشده در جدول (۴-۳) به سامانه داده شد.

لازم به ذکر است، جملات انتخاب‌شده برای بخش آزمون از دو پیکرهٔ central asia و Tehran avenue با جملات موجود در پیکره تنظیم وزن‌های سامانه، هم‌پوشانی نداشتند.

(جدول ۴-۳): مشخصات دادگان ارزیابی

نام پیکره	تعداد زوج جملات	موضوع پیکره	کیفیت ترجمه پیکره	تعداد قطعات مرجع
Khameneii.ir	۵۰۰ زوج جمله	سیاسی	ضعیف- متون مشترک کم	۴۷۵
Tehran avenue	۵۰۰ زوج جمله	هنری- عمومی	متوسط- قابل‌مقایسه	۶۰۲
Central asia	۵۰۰ زوج جمله	سیاسی- خبری	خوب-نیمه موازی	۶۵۲
Books	۳۰۰ زوج جمله	عمومی	خوب-قایل مقایسه	۲۷۸

۴-۳- روش ارزیابی

برای ارزیابی، سامانه را با توجه به نوع و کیفیت پیکرهٔ ورودی، با ۳ معیار دقت، بازخوانی و معیار F در دو سامانهٔ پایه و سامانهٔ پیشنهادی بررسی می‌کنیم. همچنین تأثیر دادگان استخراج‌شده توسط سامانهٔ پیشنهادی را در سامانهٔ ترجمهٔ ماشینی پایه با استفاده از معیار بلو می‌سنجیم.

۴-۴- مقایسه و ارزیابی سامانهٔ پایه و سامانهٔ پیشنهادی

۴-۴-۱- بررسی عملکرد سامانهٔ پیشنهادی

- **ارزیابی بخش ترازبندی اسناد**

برای ارزیابی این برنامه ۶۷۲ سند انگلیسی و ۵۵۰۰ سند فارسی از مجموعه اسناد سایت khameneii.ir در نظر گرفته شد و اسناد با توجه به مدل ترجمه، مدل طول، هر دو ویژگی و همچنین هر دو ویژگی به‌علاوه پیشوند تاریخ ترازبندی شدند. جدول (۴-۴) تعداد زوج‌اسناد ترازشدهٔ خروجی از اسناد ورودی در هر آزمایش را نشان می‌دهد.

(جدول ۴-۴): تعداد اسناد تراز شده

ویژگی‌های استفاده‌شده	تعداد اسناد تراز شده
طول	۶۷۲
ترجمه	۶۱۲
طول+ترجمه	۵۸۲
طول+ترجمه+تاریخ	۴۷۶

همان‌طور که نتایج نشان می‌دهد، استفاده از تنها یک ویژگی طول یا ترجمه برای ترازبندی اسناد کافی نیست. استفادهٔ تنها از مدل طول، نخستین سندی را که از نظر طول با سند مبدأ با توجه به نسبت طول تطابق داشته باشد، را به آن تخصیص می‌دهد و استفادهٔ تنها از مدل ترجمه، ممکن است اسنادی را با اختلاف طول زیاد فقط به‌دلیل داشتن عبارات مشترک به هم تخصیص دهد. به این دلیل تعداد اسناد خروجی در آزمایش استفاده از این دو ویژگی به‌تنهایی بالا، اما دقت آنها پایین است. استفاده توأم از این دو ویژگی منجر به نتایج منطقی می‌شود؛ اما چون مجموعه نامزد بزرگ است، زمان رویه بسیار طولانی خواهد شد. استفاده از ویژگی تاریخ، مجموعهٔ نامزد را هرس می‌کند و در نتایج در زمان کوتاهی نتایج دقیقی ارائه می‌کند که در جدول (۴-۵) قابل مشاهده است. واضح است که ویژگی تاریخ تنها می‌تواند به‌عنوان یک ویژگی جانبی استفاده شود و استفاده از آن بدون دو ویژگی اصلی طول و ترجمه بی‌معنی است.

برای محاسبهٔ دقت، بازخوانی و معیار F، ۱۰۰ تا از جفت اسناد خروجی به‌صورت کاملاً تصادفی انتخاب و صحت تخصیص آنها بررسی شد. در مورد بازخوانی پس از انجام بررسی‌ها مشخص شد ۹۱ مورد از جفت‌اسناد به‌طور واقعی موازی بوده‌اند.

(جدول ۴-۵): ارزیابی Document Aligner

معیار ارزیابی	معیار F	بازخوانی	دقت
طول	۴۴	۴۶	۴۲
ترجمه	۶۶	۷۰	۶۴
طول+ترجمه	۷۹	۸۳	۷۶
طول+ترجمه+تاریخ	۸۸	۹۳	۸۴

- **ارزیابی بخش اصلی سامانهٔ پیشنهادی**

در این بخش عملکرد سامانهٔ پیشنهادی ارزیابی می‌شود. این ارزیابی در دو بخش صورت می‌گیرد:

۱- پیکرهٔ استخراج‌شده توسط برنامه چقدر خوب است.

۲- روی سامانهٔ ترجمهٔ ماشینی تا چه اندازه تأثیرگذار است.

در این راستا ابتدا با استفاده از سه معیار دقت، بازخوانی و معیار F برای هر یک از چهار گروه داده آزمون سنجیده شده است؛ سپس سامانهٔ پیشنهادی روی سامانهٔ ترجمهٔ ماشینی پایه آزمایش شده است.

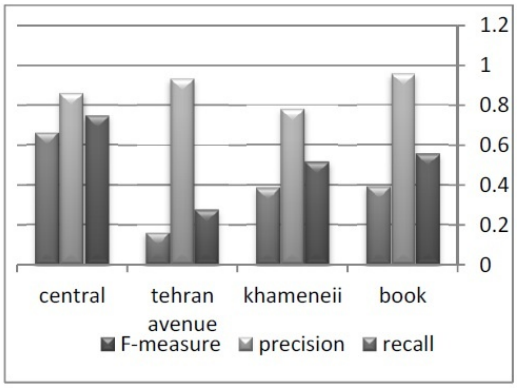
پیکرهٔ khameneii.ir یک پیکرهٔ نوفه‌ای با کیفیت ترجمهٔ نامطلوب است که حوزهٔ آن سیاسی محسوب می‌شود. برای این پیکره به‌دلیل سخنرانی‌بودن اغلب متون و ثقیل‌بودن آنها، اغلب کلمات در جدول کلمات LLR یافت نشده و بازخوانی سامانه پایین است؛ اما قطعه‌هایی که استخراج شده‌اند کیفیت تاحدودی خوب دارند. به‌دلیل محاوره‌ای‌بودن متون، قطعات زیاد از ساختار معمول زبان پیروی نکرده و باعث کاهش دقت الگوریتم نسبت به سایر پیکره‌ها شده است.

پیکرهٔ Tehran Avenue یک پیکره به‌طور کامل قابل‌مقایسه در حوزهٔ عمومی است که کیفیت ترجمهٔ متوسطی داشته و دارای اسامی خاص بسیاری است که به‌طور طبیعی در جدول LLR موجود نیستند؛ درنتیجه بازخوانی مربوط به این پیکره بسیار پایین است. از سوی دیگر کلماتی که در جدول موجود بوده‌اند با دقت زیادی قطعه تولید کرده‌اند که این نشان‌دهندهٔ دقت الگوریتم است.

پیکرهٔ central asia پیکره‌ای نیمه‌موازی، با کیفیت ترجمهٔ بالا و در حوزهٔ خبری است. به‌دلیل اینکه این حوزه با حوزهٔ فرهنگ لغت استفاده‌شده برای ساخت جدول LLR به‌طور کامل مطابق است، مشاهده می‌شود که بازخوانی الگوریتم برای این پیکره، بالا و دقت آن بسیار بالا و در کل کارایی سامانه مطلوب است.

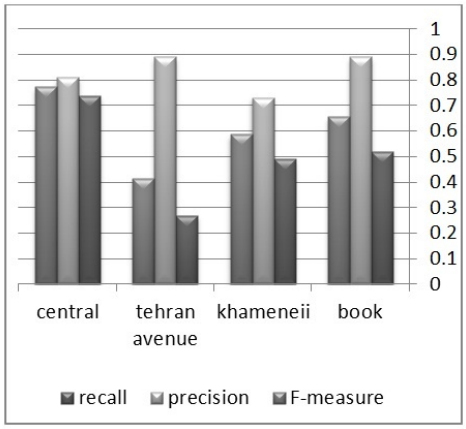
پیکرهٔ Book شامل متون عمومی و اسامی خاص زیاد است و کیفیت ترجمه متوسط است. در جملات ورودی این پیکره جملات کوتاه و محاوره‌ای نیز مشاهده می‌شود. به‌دلیل عدم اشتراک حوزهٔ این پیکره با پیکرهٔ استفاده‌شده برای ساخت جدول LLR بازخوانی برنامه برای این پیکره نیز تاحدودی پایین است. همچنین کوتاه‌بودن جملات به‌دلیل عدم امکان مانور الگوریتم برای استخراج قطعه‌ها مزید بر علت است، اما دقت برنامه به‌دلیل ترجمهٔ تاحدودی خوب کتاب‌ها برای لغات موجود در جدول LLR بالاست.

در شکل (۴-۱) عملکرد الگوریتم پیشنهادی به‌صورت کلی و به‌صورت مقایسه‌ای برای پیکره‌های مختلف آورده شده است. با توجه به نمودار، واضح است، مهم‌ترین نکته برای افزایش کارایی برنامه، تطابق حوزهٔ پیکرهٔ آزمون با حوزهٔ متون آموزشی برای ساخت جدول LLR است، که باعث می‌شود تعداد عبارات استخراج‌شده و درنتیجه بازخوانی برنامه بالا برود. همچنین ساختار جملات ورودی در دقت برنامه مؤثر است. چون الگوهای قطری که به‌عنوان ویژگی در برنامه استفاده شده‌اند، از روی ساختار زبانی برداشته‌شده‌اند. مشاهده می‌شود که دقت کلی برنامه بسیار بالاست؛ که این امر به‌دلیل دقت در انتخاب ویژگی‌های مناسب است. اما نکته‌ای که در این نمودار به چشم می‌آید، بازخوانی پایین نرم‌افزار است. با بررسی دادگان خروجی مشاهده شد تعداد زیادی از قطعه‌های استخراجی و مرجع برای پیکره‌های Tehran avenue و book طول دو دارند. این شبهه ایجاد شد که شاید الگوریتم از نظر ساختاری به داده‌های با طول کم متمایل شده باشد. در این راستا قطعه‌های با طول ۲ از مرجع حذف و بهینه‌سازی پارامترها دوباره انجام شد. این‌بار وزن ویژگی طول حدود ۰/۲ افزایش یافت. دقت و بازخوانی دوباره محاسبه شد.



(شکل ۴-۱): ارزیابی سامانه پیشنهادی روی کل پیکره‌ها

مشاهده می‌شود که این‌بار بازخوانی برنامه بالاتر رفته است؛ اما از دقت الگوریتم کاسته شده است. این امر به این دلیل است که قطعه‌ای استخراجی به‌سمت طول بیشتر هدایت شده‌اند که این کار با افزایش تعداد کلمات استخراجی، احتمال رخداد کلمات اشتباه را نیز بالا می‌برد.



(شکل ۴-۲): ارزیابی مجدد سامانهٔ پیشنهادی روی کل پیکره‌ها

۴-۴-۲- بررسی تأثیر دادگان استخراج‌شده بر روی سامانهٔ ترجمه ماشینی

در بخش دوم، ارزیابی تأثیر دادگان استخراج‌شده را بر روی سامانهٔ ترجمهٔ ماشینی بررسی می‌کنیم. الگوریتم پیشنهادی بر روی پیکره‌های قابل‌مقایسه Tehran avenue، khameneii.ir و book اعمال شد که پیش از این با استفاده از روش‌های موجود به‌دلیل کیفیت پایین قابل‌استفاده برای آموزش سامانه نبودند و از آنها ۹۸۳۴ قطعه شامل ۲۵۷۷۵ کلمهٔ انگلیسی و ۲۶۰۲۹ کلمهٔ فارسی استخراج شد.

برای ارزیابی، از سامانهٔ ترجمهٔ ماشینی موزس[1] استفاده شده است. موزس یک سامانهٔ ترجمهٔ ماشینی آماری است که در آن می‌توانیم مدل‌های ترجمه را برای هر جفت زبان دلخواه بسازیم. کافی است مجموعه‌ای از متون آموزشی دوزبانه داشته باشیم تا پس از ساخت مدل، موزس با استفاده از یک الگوریتم جستجوی بهینه محتمل‌ترین ترجمه را ارائه کند.

برای ارزیابی سامانهٔ ترجمهٔ ماشینی را یک‌بار با پیکرهٔ PEN آموزش داده و بار دیگر قطعات موازی استخراج‌شده از پیکرهٔ قابل‌مقایسه را به آن افزودیم. برای آزمون هم یک‌بار از ۱۰۰۰ جمله از پیکرهٔ PEN (مجموعه آزمون یک که البته در آموزش استفاده نشده بودند) و یک‌بار از همین تعداد جمله منتخب از پیکرهٔ khameneii.ir، Wikipedia و central asia (مجموعه آزمون ۲) استفاده کردیم. نتایج حاصل از ارزیابی بلوی سامانه عبارتند از:

(جدول ۴-۶): نتایج ارزیابی الگوریتم پیشنهادی روی سامانهٔ ترجمه

	سامانه پایه	سامانه با داده‌های افزوده
مجموعه آزمون ۱	۱۸.۱۶	۱۸.۴۶
مجموعه آزمون ۲	۱۴.۳۶	۱۵.۷۵

همان‌طورکه مشاهده می‌شود، افزودن دادگان استخراج‌شده توسط الگوریتم، اگرچه دارای حجم ناچیزی به نسبت داده‌های پیکرهٔ PEN بودند (نسبت ۱ به ۱۵) در هر دو حالت آزمون در بلوی سامانه بهبود ایجاد کرده است. در حالت اول که مجموعه آزمون از خود پیکره PEN انتخاب شده بود این بهبود کمتر و برابر ۰/۳۳ است؛ اما در حالت دوم که بخشی از مجموعه آزمون از khameneii.ir انتخاب شده بود اگر چه بلوی سامانه کاهش پیدا کرد، اما به‌صورت مقایسه‌ای نسبت به سامانهٔ پایه ۱/۴ واحد بلو افزایش داشته است. این امر به این دلیل است که درون دادگان آموزشی سامانه دو، دادگان استخراج‌شده از khameneii.ir وجود داشته است که به ترجمه بهتر مجموعه آزمون کمک کرده است. دلیل کاهش بلوی کلی سامانه‌ها نیز این است که بیشتر دادهٔ آموزشی از پیکرهٔ PEN بوده است و حجم دادهٔ آموزشی مناسب برای مجموعهٔ آزمون کافی نبوده است. همچنین متون آزمون ساختار متنی سخنرانی و سخت داشته که ترجمهٔ آن برای سامانه ساده نبوده است.

در (جدول ۴-۷) نمونه‌هایی از جملات ترجمه‌شده توسط دو سامانهٔ ترجمهٔ ماشینی مورد ارزیابی (سامانهٔ ترجمه پایه و سامانه ترجمه با دادگان افزودهٔ استخراج‌شده) برای تأیید صحت بهبود آورده شده است.

۴-۴-۳- مقایسهٔ سامانه‌های پایه و پیشنهادی

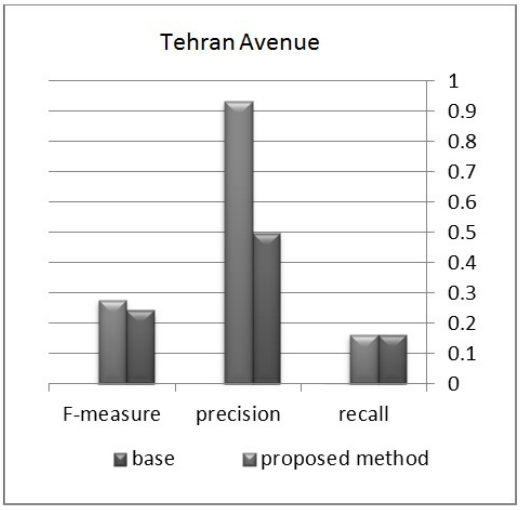
در این بخش به مقایسه کارایی سامانهٔ پایه، سامانهٔ پایه بهبودیافته و سامانهٔ پیشنهادی با استفاده از سه معیار دقت، بازخوانی و معیار F بر روی پیکره‌های آزمون پرداخته می‌شود.

همان‌طورکه در شکل (۴-۳) مشاهده می‌شود، دقت روش پیشنهادی به نسبت روش پایه، تفاوت فاحشی دارد. این امر به‌دلیل استفاده از ویژگی‌های عنوان‌شده و در نظر گرفتن الگوهای مناسب زبانی برای استخراج بلوک‌های قطعه است.

[1] moses

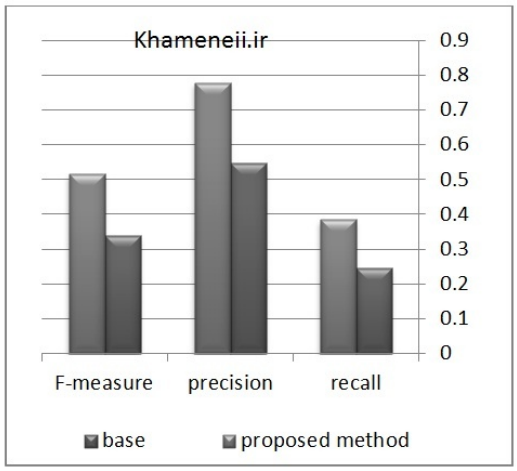
(جدول ۴-۷): مقایسهٔ نمونه‌های ترجمه در دو سامانهٔ ترجمه

۱	جمله مبدا	according to law enforcement sources , five tajik soldiers were wounded and six militants were killed in the ensuing fire fight .
	ترجمه سامانه پایه	به گفته منابع اعمال قانون , پنج سرباز تاجیکستان تن دیگر نیز زخمی شده‌اند و شش شبه‌نظامیان کشته شده در این مبارزه می‌گرفتند.
	ترجمه سامانه با دادگان افزوده	به گفته منابع اعمال قانون , پنج سرباز تاجیکستان زخمی شدند و شش نفر از پیکارجویان کشته‌شده در این مبارزه می کرد .
۲	جمله مبدا	it was so dangerous for me and for the patients .
	ترجمه سامانه پایه	این خطرناک بود برای من و برای بیماران سودمند است .
	ترجمه سامانه با دادگان افزوده	این بود بسیار خطرناک برای من و برای بیماران .



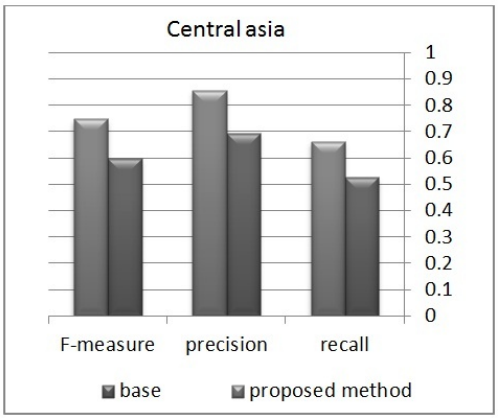
(شکل ۴-۳): مقایسه سامانهٔ پایه و پیشنهادی روی پیکره
Tehran avenue

همچنین عدم انجام ترازبندی حریصانه و درنظرگرفتن ترازبندی چندگانه که با ذات زبان فارسی مطابقت بیش‌تری دارد نیز درنتیجه بهتر مؤثر است. بازخوانی پایین برنامه همان‌طورکه پیش از این هم ذکر شد، به‌دلیل عدم تطابق حوزهٔ متن پیکره با جدول LLR است که استخراج قطعات زیاد را ناممکن ساخته است.



(شکل۴-۴): مقایسهٔ سامانهٔ پایه و پیشنهادی روی پیکرهٔ
Khameneii.ir

در این شکل نیز مشاهده می‌شود که نتایج حاصل از روش پیشنهادی از روش پایه بهتر است. بازخوانی به‌نسبه پایین الگوریتم‌ها به‌دلیل وجود کلمات سنگین و سخنرانی‌گونه و عدم تطابق کامل حوزهٔ متون با جدول LLR است؛ اما دقت برنامه مناسب است. تفاوت دقت الگوریتم پیشنهادی به‌دلیل استفاده از ویژگی‌های دقیق‌تر به نسبت روش پایه است.



(شکل ۴-۵): مقایسه سامانه پایه و پیشنهادی روی پیکره
Central Asia

همان‌طورکه در نمودار مشاهده می‌شود، مشابه پیکره‌های آزمون شکل (۴-۵) دقت برنامه بالاست که دلیل آن پیش‌تر توضیح داده شد؛ اما نکته‌ای که در اینجا مشاهده می‌شود، این است که بازخوانی برنامه و درنتیجه کارایی آن نیز بسیار بالاست که این امر به‌دلیل تطابق حوزهٔ متون

آزمـون بـا پیکـرۀ سـازنده جـدول LLR اسـت و کلمـات، هم‌پوشانی بیش‌تری دارند. نکتۀ دیگر اینکه از بخش دیگـری از این پیکره در بخش تعیین وزن ویژگی‌ها استفاده شده بود. درنتیجـه سـاختار متنـی ایـن پیکـره مناسـب بـا وزن‌هـای ویژگی‌ها بوده و نتایج بسیار مطلوب است.

۵- نتیجه‌گیری

ما در این نوشتار به معرفی روشی جدید برای استخراج قطعه از پیکـرۀ قابـل‌مقایسـه پـرداختیم. بـرای تحقـق ایـن امـر الگوریتمی طراحی و پیاده‌سازی شد که با استفاده از جـدول کلمات LLR و الگوهای رایج برای ساختار موجود قطعه‌هـای فارسی و انگلیسی و همچنـین ویژگـی‌هـایی از قبیـل طـول قطعه، مربعی‌بودن بلوک قطعه و درصد کلمات هم‌ترجمـه در قطعه که به‌صورت بهینه وزن‌دار شده‌اند، قطعات مـوازی را از متون قابل‌مقایسه استخراج می‌کند. فرآیند استخراج قطعه به این صورت است که برای هـر جفـت‌جملـه نـامزد، مـاتریس امتیازات LLR متناظر جفت‌کلمات ایجـاد مـی‌شـود؛ سـپس به‌صورت نزولی از نقطۀ دارای بیش‌ترین امتیاز، به‌عنوان نقطۀ شروع استخراج قطعه آغاز شده و پررنـگ‌تـرین نقـاط یعنـی نقاط دارای بالاترین امتیاز تا وقتی که از حد آسـتانه‌ای افـت نکرده‌اند، برای استخراج قطعه، یکی پس از دیگری پـردازش می‌شـوند. نتـایج ارزیـابی‌هـا نشـان مـی‌دهـد کـه داده‌هـای استخراج‌شده از یک‌سری متون قابل‌مقایسه موجود، عملکـرد سامانۀ ترجمۀ ماشـینی پایـه را بـا وجـود نـاچیزبودن حجـم دادگان افزوده‌شده به نسبت پیکرۀ پایۀ آموزشـی از ۰/۳۳ تـا ۱/۴واحد بلو بسته به مجموعه آزمون افزایش داده است.

همـان‌طورکـه نتـایج نشـان داد، دقـت برنامـه بسـیار بالاست؛ اما ایده‌آل‌بودن بازخوانی برنامه تا حد زیادی منـوط به تطابق حوزۀ متن آزمـون بـا پیکـرۀ سـازندۀ جـدول LLR است.

در راستای تکمیل و بهبود روش پیشنهادی، کارهـای آینده‌ای از قبیل اعمال روش‌هـای پـیش‌پردازشـی دقیـق‌تـر مانند ریشه‌یابی و دقیق‌ترسـاختن تـوکن‌بنـدی برنامـه بـرای تشـخیص تـاریخ‌هـا، اعـداد و ... بـه دادگـان ورودی، بهبـود فرهنگ لغت ساده با استفاده از پیکرۀ موازی بـزرگ‌تـر بـرای افزایش دقت در بخش انتخاب جملات نامزد، تکمیل جـدول کلمات LLR بـا اسـتفاده از داده‌هـای حـوزه‌هـای مختلـف و پیکرۀ آموزشی بزرگ‌تر و امکان انتخاب موضوع متن و به تبع آن اسـتفاده از فرهنـگ لغـت خـاص آن موضـوع پیشـنهاد می‌شود.

۶- مراجع

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

C. Tillmann, “A beam-search extraction algorithm for comparable data,” Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 225–228, Suntec, Singapore, 4 August.

C. Tillmann, Jian-ming Xu, “A simple sentence-level extraction algorithm for comparable data,” Proceedings of NAACL HLT 2009: Short Papers, pp 93–96, Boulder, Colorado.

Cerny, V. , “A thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm”, Journal of Optimization Theory and Applications, 1985, Vol. 45, pp. 41-51.

D. Munteanu and D. Marcu, “Extracting parallel subsentential fragments from non-parallel corpora,” in Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Ling-istics, 2006 , pp. 81–88.

D. Munteanu and D. Marcu, “Improving machine translation performance by exploiting non-parallel corpora,” Computational Linguistics, 2005, vol. 31, no. 4, pp. 477–504.

D.Melamed, “Bitext mapping via Pattern Recognition”, Computational Linguistics , Published by MIT Press, 1999, pp 107 – 130 Maps and Alig-ment.

Eric Gaussier, Jean-Michel Renders, Irina Mat-veeva, Cyril Goutte, and Herve Dejean, “A geometric view on bilingual lexicon extraction from comparable corpora”, In ACL 2004, pages 527–534.

Harold W. Kuhn, “The Hungarian Method for the assignment problem”, Naval Research Logistics Quarterly, 1995, 2:83–97, Kuhn's original publication.

Kirkpatrick, S. , Gelatt, C. D. , and Vecchi, M. P., (1983), “Optimization by simulated annealing, Science”, Vol. 220, pp. 671-680.

L.Lee, A.Aw, M.Zhang, and H.Li, “EM-based hybrid model for bilingual terminology extraction from comparable corpora”, Coling 2010: Poster Volume, pp. 639–646, Beijing.

Yonggang Deng, Shankar Kumar, and William Byrne, "Segmentation and alignment of parallel text for statistical machine translation". Journal of Natural Language Engineering. to appear. 2006.

Young, Steve, Corpus-Based Methods in Language and Speech Processing, Kluver Academic Publisher, 1997.

زینب رحیمی در سال ۱۳۶۶ در شاهرود متولد شد. تحصیلات تا مقطع دیپلم را در شهر شاهرود سپری و دیپلم متوسطه خود را در سال ۱۳۸۳ دریافت کرد. وی تحصیلات خود را در مقطع کارشناسی در رشتهٔ مهندسی کامپیوتر (نرم‌افزار) در دانشگاه صنعتی امیرکبیر (۱۳۸۹) و کارشناسی ارشد را در رشتهٔ مهندسی فناوری اطلاعات (سامانه‌های چندرسانه‌ای) در دانشگاه صنعتی امیرکبیر (۱۳۹۱) به پایان رساند. ایشان هم‌اکنون دانشجوی مقطع دکترا در رشتهٔ مهندسی کامپیوتر (هوش مصنوعی) در دانشگاه شهید بهشتی تهران هستند. از موضوعات مورد علاقهٔ ایشان می‌توان به پردازش زبان طبیعی، ترجمه ماشینی و هستان‌شناسی اشاره کرد.

نشانی رایانامهٔ ایشان عبارت است از:

rahimi.zeinab@gmail.com

محمدحسین ثمنی در سال ۱۳۶۵ در تهران متولد شد. ایشان تحصیلات خود را در مقطع کارشناسی در رشتهٔ مهندسی فناوری اطلاعات در دانشگاه صنعتی امیرکبیر (۱۳۸۸) و کارشناسی ارشد را در رشتهٔ مهندسی کامپیوتر (معماری شبکه‌های کامپیوتری) در همان دانشگاه (۱۳۹۱) به پایان رساند. ایشان هم‌اکنون پژوهش‌گر گروه امنیت زیرساخت (پژوهشکدهٔ افتا) در پژوهشگاه توسعه فناوری‌های پیشرفته خواجه نصیرالدین طوسی هستند. از موضوعات مورد علاقهٔ ایشان می‌توان به امنیت زیرساخت، هوش مصنوعی، داده‌کاوی و پردازش داده‌های کلان اشاره کرد.

نشانی رایانامهٔ ایشان عبارت است از:

mhsamani@gmail.com

Mona Diab and Steve Finch, "A statistical word-level translation model for comparable corpora". In RIAO 2000.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation, RC22176 (Technical Report). IBM T.J. Watson Research Center, 2001.

Pascale Fung and Lo Yuen Yee, "An IR approach for translating new words from nonparallel, comparable texts". In ACL 1998, pages 414–420.

Pascale Fung and Percy Cheun, "Mining very non-parallel corpora: Parallel sentence and lexicon extraction vie bootstrapping and EM". In EMNLP 2004, pages 57–63.

Pascale Fung and Percy Cheung. Mining Very Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In D. Lin and D. Wu, editors, Proceedings of Empirical Methods on Natural Language Processing (EMNLP'04), 2002, pp. 57–63, Barcelona, Spain.

Philipp Koehn and Kevin Knight, "Estimating word translation probabilities from unrelated mono-lingual corpora using the EM algorithm", In National Conf-erence on Artificial Intelli-gence, 2004, pages 711–715

R.Barzilay, N.Elhadad "Sentence alignment for mono-lingual comparable corpora, "Proceedings of the conference on Empirical methods in natural language processing, 2003, pp. 25 – 32.

R.C.Moore, "Fast and Accurate Sentence Alignment of Bilingual Corpora", Proceedings of the 5th Conference of the Association for Machine Transl-ation in the Americas on Machine Translation, 2002, pp 135 – 144.

S. Abdul-Rauf and H. Schwenk, "Exploiting com-parable corpora with ter and terp," in BUCC 09: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, 2009, pp. 46–54.

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati and Stephan Vogel, "CMU Haitian Creole-English Translation System", 2011, In Proc. WMT. Pp386-392.

W.A.Gale, K.w.church, "A Program for Aligning Sentences in Bilingual Corpora", Proceedings of the 29th annual meeting on Association for Comp-utational Linguistics, 1991, pp 177 – 184.

X.Ma, "Champollion: a robust parallel Text sentence aligner," Proceedin of the fifth international confer-ence on language resource and evaluation (LREC 2006), pp. 167-172.

شهرام خدیوی مدارک کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی کامپیوتر از دانشگاه صنعتی امیرکبیر به‌ترتیب در سال‌های ۱۳۷۵ و ۱۳۷۸ دریافت کرده و همچنین مدرک دکترای خود را در سال ۱۳۸۷ از دانشگاه آخن RWTH در رشتهٔ علوم کامپیوتر دریافت کرده‌اند. ایشان در حال حاضر عضو هیأت علمی و استادیار دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر هستند. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش زبان طبیعی، ترجمهٔ ماشینی آماری و یادگیری ماشین است.

نشانی رایانامهٔ ایشان عبارت است از:

khadivi@aut.ac.ir