دانشگاه فردوسی مشهد
دانشکده مهندسی - گروه مهندسی کامپیوتر

پایان نامه کارشناسی ارشد

ارائه رهیافتی جدید برای تولید پیکره موازی انگلیسی–فارسی

تهیه و تنظیم:

سید احمد جکیان طوسی

استاد راهنما:

دکتر محسن کاهانی

استاد مشاور:

دکتر هادی صدوقی یزدی

تابستان 91

تقدیم به:

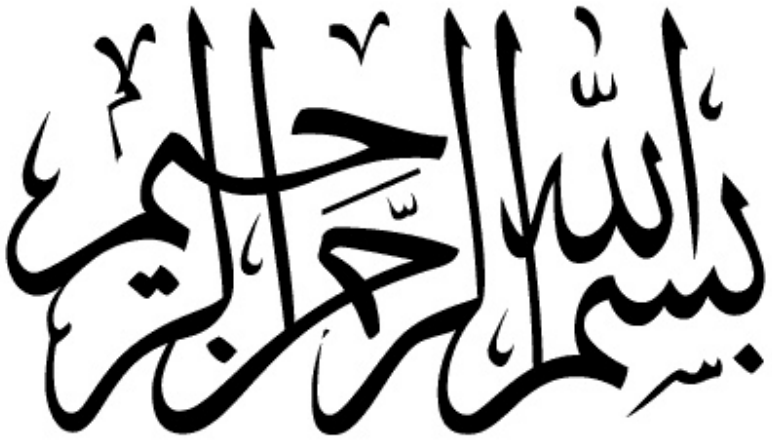
به همه معلمانم، که سرچشمه لطیف علم اند و پیامبران منش فرهنگی ام

به پدرم، که استاد معرفت است و اسوه شکیبائی ام

به مادرم که معدن محبت است و سرمایه آسمانی ام

به همسرم که مروارید بی بدیل عشق است و امید بخش زندگانی ام

و به فرزندان عزیز و عزیزان دلبسته ام


بسم الله الرحمن الرحيم

تعهدنامه

اینجانب سید احمد جکیان طوسی دانشجوی کارشناسی ارشد رشته مهندسی کامپیوتر، گرایش هوش مصنوعی، دانشکده مهندسی، دانشگاه فردوسی مشهد، نویسنده پایان‌نامه *"ارائه رهیافتی جدید برای تولید پیکره های موازی انگلیسی فارسی"* تحت راهنمایی دکتر محسن کاهانی متعهد می‌شوم:

تحقیقات در این پایان‌نامه توسط اینجانب انجام شده و از صحت و اصالت برخوردار است.

در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است.

مطالب مندرج در پایان‌نامه تاکنون توسط خود و یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.

کلیه حقوق معنوی این اثر متعلق به دانشگاه فردوسی مشهد می‌باشد و مقالات مستخرج با نام "دانشگاه فردوسی مشهد" و یا "Ferdowsi University of Mashhad" به چاپ خواهد رسید.

حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از رساله رعایت شده است.

در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است. در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است، اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ امضای دانشجو

دانشگاه فردوسی مشهد
دانشکده مهندسی - گروه مهندسی کامپیوتر

آزمایشگاه تخصصی فناوری وب WTLAB

پایان نامه کارشناسی ارشد

ارائه رهیافتی جدید برای تولید پیکره موازی انگلیسی–فارسی

تهیه و تنظیم:

سید احمد جکیان طوسی

استاد راهنما:

دکتر محسن کاهانی

استاد مشاور:

دکتر هادی صدوقی یزدی

تابستان 91

تقدیر و تشکر

پس از حمد بیکران از لطف و رحمت الهی و پیش از ارایه هر مطلبی، لازم می‌دانم با احترام از کسانی نام ببرم که در طول انجام پژوهش‌های مربوط به پایان نامه دوره کارشناسی ارشد، مرا یاری نمودند و با کمک رسانی مخلصانه خویش امکان غلبه بر موانع و چالش‌های موجود را فراهم کردند.

بیش از همه، شایسته می‌دانم، تشکر ویژه خود را تقدیم به استاد فرهیخته و فرزانه گران قدر، دکتر محسن کاهانی نمایم که در این مدت، شاگرد خویش را دلسوزانه و صبورانه حمایت نموده و با تشویق‌های موثر خود، عامل مهمی در پیشرفت امور علمی ام بوده‌اند.

همچنین از استاد مشاور پروژه، دکتر هادی صدوقی یزدی که سهم بسزایی در طی نمودن صحیح این مسیر داشتند، به طور ویژه، تقدیر و تشکر می‌نمایم.

از استاد فرزانه زبان شناسی، پروفسور نادر جهانگیری نیز به دلیل همراهی موثرشان کمال سپاس و قدردانی را دارم.

بدون تردید اگر همکاری سایر دوستان و همکارانم در آزمایشگاه فناوری وب نبود، انجام کارهای مربوط به جمع آوری داده و پیاده سازی سیستم با مشکلات بسیار همراه می‌شد به همین منظور با تقدیم سپاس به مجموعه توانمند آزمایشگاه فناوری وب شایسته می‌دانم به طور ویژه از زحمات دوستان عزیزم آقایان مهندس، رضا سعیدی، نیما صالحی، احمد استیری، امین الله یار و بهداد بهدی به جهت همکاری هایی که در مقاطع گوناگون با اینجانب داشته‌اند، صمیمانه تشکر نمایم.

در نهایت با امید به بهبود هر چه بیشتر فرآیندهای تحقیقاتی، توفیقات روزافزون را برای همه فعالان عرصه رشد و شکوفایی علمی، در دانشکده مهندسی دانشگاه فردوسی مشهد آرزو می‌کنم.

سید احمد طوسی

چکیده:

در این پژوهش، برای اولین بار مدلی ترکیبی برای تراز بندی جملات، جهت ساخت پیکره‌های موازی انگلیسی-فارسی ارائه شده است. در حالت کلی چارچوب روش پیشنهادی، غیر وابسته به زبان‌های مبدأ و مقصد بوده و از آن می‌توان برای تولید پیکره‌های موازی، برای هر جفت زبان دیگر، نیز استفاده کرد. نتایج بدست آمده از پیاده‌سازی‌ها نشان می‌دهد که بکار بردن ویژگی‌های زبانی و غیر زبانی ، عملکرد سیستم را تا حد قابل قبولی بهبود خواهد بخشید. همچنین در این تحقیق میزان تأثیر استفاده از شباهت‌های طولی، ترجمه تحت‌اللفظی و نقش دستوری کلمات، به صورت مستقل یا ترکیبی، برای عملیات ترازبندی مورد بررسی قرار گرفته است. بکارگیری طبقه بندهای سلسله‌مراتبی در تشخیص نوع تراز بندی به عنوان یکی از شاخص‌های اصلی سیستم پیشنهادی محسوب می‌شود که موجب بالا رفتن دقت وسرعت عملیات تولید پیکره نسبت به سایر مدل‌ها می‌شود. از ویژگی‌های دیگر این روش، قابلیت توسعه‌پذیری آن است. به این ترتیب که می‌توان با گنجاندن خصوصیاتی (در بدنه بردارهای ورودی) که ممکن است در آینده برای تشخیص بهتر نوع تراز بندی مورد توجه قرار گیرند کیفیت سیستم ترازبندی را ارتقاء بخشید. با این حال، چالش اساسی این رهیافت و بسیاری از روش‌های پیشین، وجود سلایق متنوع در ترجمه متون است. این مسئله در برخی موارد موجب تولید جملاتی می‌گردد که مشابهت آن‌ها با متن اصلی در حد معنا بوده و تنها قابل درک و تشخیص برای انسان می‌باشد. در چنین وضعیتی کار استخراج جفت عبارات معادل، بسیار دشوار است. علاوه براین در مورد متونی که مقید به رعایت ساختار دستوری زبان فارسی و انگلیسی نمی‌باشند عملیات تشخیص جملات هم تراز ممکن است به خوبی انجام نپذیرد.

کلمات کلیدی:

تراز بندی جملات، پیکره موازی، ترجمه آماری

فهرست مطالب

فهرست نمودارها و شکل‌ها

فهرست جدول‌ها

فصل اول:

مقدمه

1- مقدمه

بنابر آمار منتشر شده از سوی یکی از نشریات معتبر مربوط به زبان‌شناسی دنیا [VER04]، ترجمه ماشینی به عنوان یک نامزد جدی در تصاحب اولین جایگاه در میان 10 فناوری برتر دنیا مطرح می‌باشد. این فناوری‌های نوین مدعی‌اند که می‌توانند دنیا را به طور اساسی دچار تحول نمایند. در این فهرست از فناوری‌های نانو ، بیوتکنولوژی ، تکنولوژی اطلاعات ، علوم شناختی ، روباتیک و هوش مصنوعی نیز به عنوان دیگر فناوری‌های برتر نام برده شده است.

تاکنون دانشمندان و نظریه‌پردازان علوم مربوط به زبان شناسی و فنون مهندسی رایانه، راهکارهای مشترک و متنوعی را برای حل مسئله ترجمه ارایه نموده‌اند. با ظهور پدیده اینترنت و برداشته شدن موانع ارتباطی میان ملت‌ها، این موضوع بیش از پیش، مورد تقاضای ملل گوناگون، که دارای گویش‌ها و فرهنگ‌های مختلف هستند، قرار گرفته است. با اینکه بین‌المللی کردن زبان انگلیسی، به عنوان یک راهکار اولیه، برای تسهیل ارتباطات انسان‌ها در سراسر دنیا، مورد قبول جامعه جهانی است، اما وفاداری ملت‌ها به حفظ فرهنگ گفتاری و نوشتاری خود مانع از رفع کامل نیازشان به سیستم‌های مترجم می‌باشد. از این رو از دیرباز، با وجود دوران رخوت و سستی در طراحی این سیستم‌ها باز هم این موضوع فراموش نشده و در مقاطع گوناگون به عنوان یکی از معمای علمی ممتاز مورد توجه محققان بوده است.

1-1- انگیزه

یک مترجم ایده‌آل، در حقیقت یک انتقال دهنده دانش است که این کار را با تبدیل معنایی عبارات مبدأ به جملاتی در زبان مقصد انجام می‌دهد. به بیان ساده تر یک مترجم خوب مترجمی است که بتواند منظور نویسنده یا گوینده را که در قالب زبان خاص بیان شده است به صورت جملاتی به زبان خاص دیگر که متعلق به خواننده یا شنونده است تبدیل نماید. با وجود سادگی این تعریف، پیاده‌سازی چنین سیستمی نه تنها ساده نیست بلکه بسیار پیچیده و نیازمند دقت در ابعاد مختلف می‌باشد. از این رو، انگیزه بسیاری از محققان فنون ترجمه و علوم زبان شناسی رایانه ای یافتن راهی است که از آن طریق بتوان به ترجمه ایده آل برای متون مختلف دست پیدا کرد.

یکی از راه‌های ساخت مترجم ماشینی، استفاده از دانش آماری نهفته در متون ترجمه شده‌ای است که داده های آن در قالب پیکره های موازی برای زبان مبدأ و مقصد در اختیار می‌باشند. [RAM04]. به عبارت دیگر، ایده اصلی این روش –که با عنوان ترجمه ماشینی آماری شناخته می‌شود- انجام عملیات ترجمه بر اساس یافتن شبیه‌ترین عبارات (موجود در) پیکره موازی به عبارت ورودی می‌باشد. صرف نظر از تکنیکی که برای پیدا کردن زیر عبارات مشابه مورداستفاده قرار می‌گیرد به نظر می‌رسد مهم‌ترین چالش موجود در این تکنیک، دسترسی به پیکره موازی به حد کافی بزرگ برای دو زبان مبدأ و مقصد می‌باشد.

تا کنون روش‌های مختلفی برای جمع آوری متون موازی از روی نوشته های دو زبانه پیشنهاد شده است. ولی به این دلیل که همواره مترجم انسانی به دلایلی ترجمه خود را جمله به جمله انجام نمی‌دهد، مسئله دیگری با عنوان تراز بندی مطرح می‌شود. به این معنا که ممکن است مثلاً مترجم یک جمله زبان مبدأ را به دو یا چند جمله در زبان مقصد و یا بالعکس ترجمه نماید. از این رو در سال‌های اخیر، مقالات متعددی با عنوان تراز بندی جملات به منظور تولید پیکره های موازی ارایه شده‌اند که با دیدگاه‌های متفاوت به دنبال تشخیص نوع تراز بندی بکار رفته در ترجمه می‌باشند.

به عنوان نمونه، در برخی از روشهای ترازبندی جملات، بدون توجه به ویژگی‌های زبانی و معنایی دو زبان، تنها بر اساس تعداد کاراکترهای تشکیل دهنده جمله/جمله‌ها، عمل ترازبندی انجام می‌پذیرد. ایده اساسی

این رویکرد را می‌توان در یک عبارت چنین خلاصه نمود که معمولاً عبارات با طول کم به عبارات با طول کم و عبارات با طول بیشتر به عبارات با طول بیشتر ترجمه می‌شوند.[GAL93]

بررسی‌ها نشان می‌دهد که در مورد برخی از جفت زبان‌ها، نقش دستوری کلمات در هنگام ترجمه چندان دست‌خوش تغییر نمی‌شوند. یعنی واژه‌هایی که با نقش فعل در جمله مبدأ بکار رفته‌اند در عبارت ترجمه نیز به صورت فعل ظاهر می‌شوند. از این رو ابزارهایی برای تولید برچسب‌های زبانی چه از نوع معنایی و چه از نوع دستوری مورد توجه بخشی دیگری از محققان قرار گرفت.[CHE95]

در تعداد دیگری از روش‌های مطرح شده، [MEY98] از فرهنگ واژگان برای انجام عمل تراز بندی بهره گرفته شده است. در واقع لغت نامه نقش یک مترجم واژه به واژه را در این روش ایفا می‌کند که کار یافتن شبیه‌ترین عبارت/عبارات را در دو متن موازی انجام می‌دهد.

عده دیگری از پژوهشگران، با توجه به رشد روزافزون وب و انتشار سریع داده های متنی در آن، تکنیک‌هایی را برای یافتن جفت عبارات موازی از روی وب مطرح نموده‌اند. به عنوان مثال در سایت‌های خبری بین‌المللی یا دایره‌المعارف‌های چند زبانی، معمولاً ابر پیوندها[1] در دو عبارت موازی، به یک محل مشترک اشاره می‌کنند. تا کنون بوسیله این روش ابتکاری، تعداد قابل توجهی عبارات موازی، از درون مخازن وب استخراج شده است.[MOH10]

در این میان برخی روش‌های ترکیبی [ASW05] مطرح شده‌اند که انگیزه اصلی آن‌ها از ترکیب روش‌های گوناگون، پوشش نقاط ضعف برخی از روش‌ها، به وسیله روش‌های دیگر می‌باشد. نتایج بررسی‌ها نشان دهنده این موضوع است که روش‌های ترکیبی نسبتاً بهتر از روش‌های قبلی عمل می‌کنند.

تا کنون عملیات پیاده سازی روش‌هایی که پیش از این به آن‌ها اشاره اجمالی شد بر روی تعداد بسیاری از جفت زبان‌های دنیا صورت گرفته است که نتیجه آن تولید پیکره حجیم موازی بوده است. همانطور که گفته شد استفاده از این پیکره های عظیم در مدل‌های ترجمه آماری سبب ارتقاء کیفیت ترجمه های ماشینی بدست آمده شده است. با این وجود، متأسفانه سهم پارسی‌زبانان در تولید خودکار پیکره انگلیسی

---

[1] Hyperlink

فارسی چندان قابل توجه نبوده است. البته در حال حاضر حرکت‌های مثبت توسط تیم‌های دانشگاهی مرتبط در حال انجام می‌باشد لکن اهمیت موضوع به اندازه‌ای است که مستعد انجام کوشش‌های بیشتر بوده و رسیدگی به این امر را به یک ضرورت قطعی تبدیل نموده است.

انگیزه اصلی پایان نامه حاضر ارایه راهکاری مؤثر، برای تولید خودکار پیکره موازی انگلیسی فارسی می‌باشد. رهیافت کلی تحقیق پیش رو، مبتنی بر یک روش ترکیبی، می‌باشد. چارچوب کلی روش ترازبندی پیشنهاد شده، مستقل از هر دو زبان مبدأ و مقصد می‌باشد. با این حال امکان توجه به خصوصیات زبانی و فرا زبانی به صورت مستقل و یا با هم نیز وجود دارد. نتایج حاصل از کاربرد روش پیشنهادی برای تولید پیکره‌های حجیم، می‌تواند باعث توسعه سیستم‌های مترجم قوی‌تر شود و لذا تأثیر بسزایی در تسهیل فرآیندهای تحقیقاتی ، آموزشی و تجاری خواهد داشت.

1-2- راه حل پیشنهادی

تا کنون روش‌های متنوعی برای تولید پیکره های حجیم موازی ارائه شده است. تعدادی از روش‌های مذکور از معیارهای فرا زبانی[2] یعنی معیارهای غیر وابسته به زبان خاص، برای حل مسئله تراز بندی عبارات استفاده می‌کنند. این ویژگی سبب شده است که از آن بتوان برای تولید هر نوع پیکره دو زبانی استفاده نمود.

در گروه دیگری از روش‌ها به ویژگی‌های زبان‌های مطرح در پیکره، توجه خاص شده است. از این رو نمی‌توان انتظار داشت که این گونه مدل‌ها برای هر جفت زبان موجود در دنیا کاربرد داشته باشند. مزیت کلی این گروه از روش‌ها را می‌توان به دلیل بالا بودن دقت آن‌ها در تشخیص انواع تراز بندی جملات در دو متن موازی دانست.

در مدل پیشنهادی که در این تحقیق به آن پرداخته شده است هم به خصوصیات فرا زبانی و هم به ویژگی‌های زبانی توجه شده است، تا ضمن برخورداری از مزایای روش‌های فرا زبانی، با بهره‌مندی از ویژگی‌های خاص زبانی، امکان تولید پیکره های مطلوب انگلیسی فارسی به وجود آید. بر این اساس

---

[2] Language-independent

می‌توان از این مدل با اطمینان برای ترازبندی جملات مربوط به هر یک از زبان‌های هم خانواده با زبان فارسی از قبیل هندی و عربی در برابر عبارات انگلیسی استفاده نمود.

همان‌طور که در ادامه نیز تشریح خواهد شد استفاده از تئوری‌های یادگیری ماشین از قبیل طبقه‌بندهای درخت تصمیم[3] CART و ماشین بردار پشتیبان[4] در تراز بندی جفت جملات انگلیسی و فارسی جزء مهم‌ترین بخش‌های سیستم پیشنهادی می‌باشد. علاوه بر این‌ها استفاده از ابزار زبان شناسی و انجام برخی اصلاحات در آن‌ها به منظور کاربرد بهتر در ساختار مدل ترازبندی، از دیگر امتیازات این روش می‌باشد.

در یک نگاه کلی، روشی که در این پایان نامه به آن اشاره می‌شود به این صورت است که ابتدا یک پیکره تراز شده با حجم و دقت قابل قبول به یک سیستم یادگیرنده آموزش داده می‌شود. پس از آن، سیستم توانایی انجام عملیات ترازبندی، بر روی عبارات جفت متن موازی را خواهد داشت.

مهم‌ترین ویژگی مدل پیشنهاد شده در این پایان نامه، عدم وابستگی آن به جفت زبان‌های مربوط به پیکره – در حالت کلی- می‌باشد. البته در صورت در دسترس بودن هر یک از خصوصیات زبانی موثر در ترازبندی، می‌توان با ساخت مؤلفه به شیوه‌ای که بعدا خواهد آمد، از این خصوصیات برای بهبود دقت عملیات ترازبندی استفاده نمود.

به طور کلی نوآوری‌های موجود در این روش را می‌توان در دو بخش طبقه بندی کرد.

در بخش اول و در قسمت استخراج ویژگی، استفاده از مولفه های زبانی برچسب دستوری در ساخت مؤلفه های بردار ویژگی به عنوان یک رویکرد جدید مطرح می‌باشد. همچنین در همین قسمت استفاده از اصلاحگر های برچسب زن دستوری نیز از دیگر مواردی است که تا کنون به آن پرداخته نشده است. همچنین بکارگیری ابزار مترجم ماشینی گوگل به عنوان یک ابزار کمکی سبب شده است که عملیات ساخت ترجمه تحت‌اللفظی عبارات انگلیسی، ارزش دقیق‌تری را به میزان نزدیکی دو عبارت فارسی و انگلیسی کاندید برای ترازبندی تولید نماید.

---

[3] Decision tree

[4] Support Vector Machine