دانشکده برق و کامپیوتر

پایان‌نامه دکتری در رشته مهندسی کامپیوتر (سیستمهای نرم افزاری)

ترجمه ماشینی مبتنی بر روش‌های داده کاوی و ساختارهای استنتاجی

توسط:

سید مصطفی فخراحمد

اساتید راهنما:

دکتر منصور ذوالقدری جهرمی

دکتر محمد هادی صدرالدینی

اردیبهشت ۱۳۹۱


بسم الله الرحمن الرحيم

به نام خدا

اظهارنامه

اینجانب سید مصطفی فخراحمد (۸۶۳۲۸۴) دانشجوی رشته‌ی مهندسی کامپیوتر گرایش سیستمهای نرم افزاری دانشکده‌ی برق و کامپیوتر اظهار می‌کنم که این پایان‌نامه حاصل پژوهش خودم بوده و در جاهایی که از منابع دیگران استفاده کرده‌ام، نشانی دقیق و مشخصات کامل آن را نوشته‌ام. همچنین اظهار می‌کنم که تحقیق و موضوع پایان‌نامه-ام تکراری نیست و تعهد می نمایم که بدون مجوز دانشگاه دستاوردهای آن را منتشر ننموده و یا در اختیار غیر قرار ندهم. کلیه حقوق این اثر مطابق با آیین‌نامه مالکیت فکری و معنوی متعلق به دانشگاه شیراز است.

نام و نام خانوادگی: سید مصطفی فخراحمد

تاریخ و امضاء: ۹۱/۲/۸

به نام خدا

ترجمه ماشینی مبتنی بر روشهای داده کاوی و ساختارهای استنتاجی

به وسیله‌ی:

سید مصطفی فخراحمد

پایان نامه

ارائه شده به تحصیلات تکمیلی دانشگاه به عنوان بخشی

از فعالیت‌های تحصیلی لازم برای اخذ درجه دکتری

در رشته:

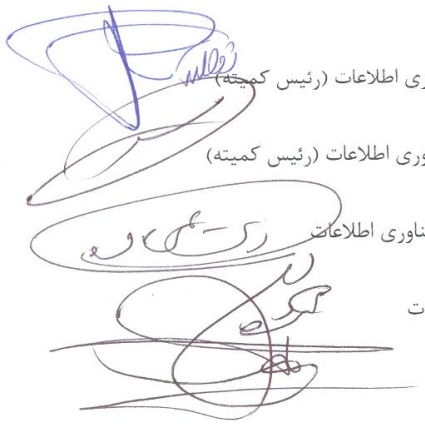
مهندسی کامپیوتر- سیستمهای نرم افزاری

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته پایان نامه با درجه:



دکتر منصور ذوالقدری جهرمی، استاد بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات (رئیس کمیته)

دکتر محمد هادی صدرالدینی، دانشیار بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات (رئیس کمیته)

دکتر غلامحسین دستغیبی فرد، استادیار بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات

دکتر اشکان سامی، استادیار بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات

اردیبهشت ۱۳۹۱

تقدیم

به

روح بلند پدر بزرگوارم

و

به دو بال پروازم

مادر عزیزم و همسر مهربانم

سپاسگزاری

اکنون که به یاری خداوند متعال موفق به اتمام این پایان نامه شده ام، لازم می‌دانم از زحمات ارزنده و کمک‌های بی‌دریغ اساتید گرانقدرم جناب آقای دکتر منصور ذوالقدری جهرمی و جناب آقای دکتر محمد هادی صدرالدینی که در تمامی مراحل پایان نامه از راهنمایی‌های ارزشمند خود اینجانب را بهره‌مند ساختند، قدردانی کرده و برای ایشان آرزوی سلامتی و موفقیت کنم.

همچنین از دوستان و خانواده عزیزم بویژه همسرم و برادرم سید مجتبی فخراحمد که با مساعدت‌های ارزشمند خود بنده را در انجام این کار یاری کردند، کمال سپاسگزاری را دارم.

چکیده

ترجمه ماشینی مبتنی بر روشهای داده کاوی و ساختارهای استنتاجی

به وسیله‌ی:

سید مصطفی فخراحمد

ترجمه ماشینی یکی از جذاب ترین شاخه ها در زمینه پردازش زبان طبیعی(NLP) است. ترجمه ماشینی(MT) یک فرایند تجزیه و تحلیل خودکار متن در زبان مبدا و تولید متن معادل آن در زبان هدف است. روش‌های موجود برای ترجمه ماشینی را می‌توان به دو دسته کلی روش‌های مبتنی بر قانون و روش‌های مبتنی بر مجموعه متون تقسیم کرد. روش‌های مبتنی بر قانون وابسته به زبان بوده و در حل مشکل ابهام معنایی ناتوان هستند، برای حل این مسائل استفاده از روش‌های مبتنی بر مجموعه متون می‌تواند چاره‌ساز باشد. در این تحقیق، سیستم ترجمه جدیدی بنام مترجم برنا را معرفی می کنیم که رویکرد آن ترکیبی از هردو روش مبتنی بر قانون و مبتنی بر مجموعه متون می باشد. بخش مبتنی بر قانون سیستم وابسته به زبان نیست، زیرا قواعد گرامری از زبان به طور خودکار از مجموعه متون موجود استنتاج می شود. ساختار اصلی استفاده شده در سیستم پیشنهادی ساختار آتاماتای متناهی (DFA) است. این اولین بار است که این ساختار در ماشین ترجمه استفاده می شود. مهمترین مزیت این ساختار جداسازی دانش از ماشین ترجمه است. در سیستم برنا، قوانین گرامری استخراجی در قالب ماشین‌های متناهی تو در تویی ارائه خواهند شد که در ماشین ترجمه بصورت بازگشتی یکدیگر را فراخوانی خواهند کرد. در این تحقیق، همچنین تلاش شده است تا برای یکی از پرچالش ترین مسائل ترجمه ماشینی یعنی رفع ابهام سه راهکار جدید ارائه شود. در راهکار اول که یک روش یادگیری نظارتی است از الگوریتم نزدیک ترین همسایه استفاده می‌شود. در این روش به منظور بهبود دقت طبقه بند از یک فرآیند انتخاب خصیصه و یک روش وزن دهی جدید استفاده خواهد شد. در راهکار دوم، یک سیستم طبقه بند مبتنی بر قوانین فازی ارائه می‌شود و به منظور افزایش دقت طبقه بند شیوه‌ی وزن دهی به قوانین معرفی خواهد گردید. اما راهکار سوم، یک سیستم خبره است که اساس کار آن استنتاج دانش بر مبنای روش زنجیره پیشرو است. پایگاه دانش این سیستم مجموعه ای از قوانین همبستگی بین کلمات مختلف زبان است که طی یک فرایند داده کاوی بدست آمده اند. قابلیت استنتاج موجب می شود که روش پیشنهادی قادر باشد با تکیه بر برخی از کلمات موجود معنای سایر کلمات مبهم را تشخیص دهد حتی اگر این کلمات قبلا در مجموعه متون آموزشی هرگز با هم رخ نداده اند.

فهرست مطالب

فهرست جداول

فهرست شکل‌ها

عنوان **صفحه**

فصل اول

۱- مقدمه

۱-۱- ضرورت ترجمه ماشینی

با ماشینی شدن کارها و کاهش نقش مستقیم انسان در به انجام رساندن پروژه‌های مختلف، لزوم وجود نرم‌افزاری هوشمند برای ترجمه روان متون انگلیسی به فارسی و بالعکس بر متخصصین، مهندسان و مترجمان، پوشیده نیست. یک کاربرد مهم ترجمه ماشینی که از ابتدا مطرح بوده، کاربرد نظامی است که شروع آن از وزارت دفاع آمریکا بوده است و هم‌اکنون نیز از آن بطور جدی در بخش نظامی استفاده می شود. جالب است که اخیرا مترجم ماشینی فارسی هم به جمع پروژه‌های مورد استفاده بخش نظامی آمریکا اضافه شده است. یک کنفرانس در زمینه پردازش زبان طبیعی و ترجمه ماشینی برای زبانهای مشابه فارسی و عربی نیز سالانه در آمریکا انجام می شود.

یک انسان برای ترجمه متون از زبان مبدا به مقصد، نیازمند یادگیری هر دو زبان و بکارگیری گرامرها و لغات، جهت ترجمه است. فردی که با زبان انگلیسی و فارسی آشنایی دارد، با روبرو شدن با یک جمله انگلیسی، ابتدا کلمات را در فرهنگ لغات ذهن خود جستجو کرده و با تشخیص نقش دستوری هر کدام بوسیله همان فرهنگ لغات، جایگاه درست را برای لغات پیدا می کند؛ او با دیدن لغات دیگر جمله و نقش‌های آن‌ها، متوجه ساختار گرامری جمله شده و ترجمه صحیح را از ذهن گذرانده، بر زبان می آورد و یا بر روی کاغذ می‌نویسد. از این رویکرد در روشهای قدیمی‌تر ترجمه ماشینی که موسوم به روشهای مبتنی بر قانون [4, 5] هستند، الگوبرداری شده و فرایند ترجمه بر مبنای آن انجام می‌گیرد. روشهای مبتنی بر قانون اخیرا به دلایلی از جمله قدرت محدود این روشها، محدودیتهای تولید قوانین و نیاز به متخصصان زبانشناسی دیگر بر روی آن کاری انجام نمی‌شود.

اما تشخیص صحیح گرامر جمله زبان مبدا و الگوی ترجمه (با کمک روشهای مبتنی بر قانون) برای رسیدن به یک ترجمه قابل قبول کافی نیست. مشکلات عمده‌ای که در فرایند ترجمه وجود دارند عمدتا وابسته به پویایی ذهن انسان و قابلیتهای فردی و انسانی مترجم هستند. بدین معنی که میزان تسلط فرد بر دو زبان و تجربه او در امر ترجمه بر کیفیت ترجمه بسیار تاثیرگذار خواهد بود. از مهمترین مشکلات ترجمه بعد از تشخیص گرامر جمله می توان موارد زیر را نام برد: تشخیص ریشه کلمات تغییر شکل یافته، تشخیص نقش یک کلمه در جمله در موارد مبهم، تشخیص زمان فعل در جملاتی که فاقد قید زمان هستند، انتخاب معنای مناسب برای کلمات دارای چند معنا، تشخیص افعال چند قسمتی بخصوص افعال جدایی پذیر، انتخاب ترجمه مناسب برای عبارات و ترکیب‌های معین و ... . برخی از این مشکلات، گلوگاههای اصلی در ساخت یک ماشین مترجم محسوب می‌شوند. بخشهایی از عملیات ترجمه که نیاز به تجربه و تسلط مترجم دارد، موضوع بحث روشهای ترجمه مبتنی بر مثال [۶-۹] است. اما بزرگترین مشکل این روشها نیاز به تهیه حجم زیادی از متون ترجمه شده در زبانهای مختلف است. در این روشها که بر مبنای مقایسه با ترجمه های صحیح می باشند، سیستم حاوی یک پایگاه دانش بسیار بزرگ و متنوع می‌باشد. در این پایگاه دانش، نمونه هایی از جملات و عبارات زبان مبدا به همراه ترجمه آنها در زبان مقصد ذخیره می شود. با ورود یک جمله جدید برای ترجمه، سیستم از همین عبارات موجود که با یکدیگر منطبق شده اند، کمک می گیرد تا ترجمه عبارت جدید را بدست آورد.

در سال های ۱۹۵۰و۱۹۶۰، ترجمه ماشینی حوزه تحقیقی مهمی در زبان شناسی رایانه‌ای به حساب می آمد. در شروع، هدف مورد انتظار ترجمه خودکار انواع متون و اسناد در کیفیتی معادل یا حتی بهتر از مترجمان انسانی بود. اما خیلی زود معلوم شد که این هدف در آینده‌ای قابل پیش‌بینی غیر ممکن است. اگر نتایج ترجمه بخواهد به شکلی قابل چاپ درآید، اصلاح و تجدید نظر برونداد ترجمه ماشینی به وسیله انسان ضروری است. در عین حال این نکته نیز قابل درک است که برونداد ترجمه ماشینی به شکل اصلاح نشده و خام نیز می تواند برای بسیاری از مقاصد از جمله برای خواندن متن به منظور پی بردن به ایده و محتوای کلی آن زبان نا آشنا مفید واقع گردد. اما این کاربرد ترجمه ماشینی برای سال ها مورد غفلت واقع گشت.

بیشترین حجم ترجمه در دنیا از متونی نیست که از موقعیت ادبی وفرهنگی بالایی برخوردار باشند. اکثریت قریب به اتفاق مترجمان حرفه ای دست اندرکار ترجمه اسناد و متون علمی وفنی، معاملات تجاری وبازرگانی، اساسنامه‌های مدیریتی، اسناد حقوقی، دستورالعمل ها،

کتاب‌های کشاورزی و پزشکی، پروانه‌های صنعتی، نشریه‌های تبلیغاتی، گزارشات روزنامه‌ای و غیره می‌باشند. مقداری از این کار پردرد سر و مشکل است. اما بیشتر آن کسل کننده و تکراری است ولی در عین حال نیاز به صحت وثبات دارد. تقاضا برای چنین ترجمه هایی با سرعتی بیش از قابلیت حرفه مترجمی رو به افزایش است. مساعدت یک کامپیوتر جذابیت‌های واضح وضروری دارد. سودمندی عملی یک سیستم ترجمه ماشینی نهایتاً به وسیله کیفیت برونداد آن تعیین می شود. اما آنچه که به عنوان یک ترجمه خوب صرف نظر از اینکه به وسیله کامپیوتر باشد یا انسان تلقی شود مفهومی است که تعریف دقیق آن مشکل به نظر می رسد. در عین حال بیشتر آن وابسته به موقعیت‌های ویژه‌ای که در آن‌ها ترجمه صورت می‌گیرد و نیز دریافت‌کننده‌های ویژه‌ای که ترجمه برای آن انجام می‌شود است. صداقت، صحت، وضوح، سبک مناسب و سیاق همه از جمله ملاک‌هایی هستند که می‌توان از آن‌ها استفاده کرد، اما باز هم این ها قضاوت‌های ذهنی هستند. تا آنجایی که به ترجمه ماشینی مربوط می شود آنچه که در عمل مهم است این است که چه مقدار تغییرات بایستی صورت گیرد تا برونداد ترجمه به حد استاندارد قابل قبولی از نظر مترجم یا خواننده انسانی رسانده شود. با چنین مفهوم دشوار و بی-ثباتی چون ترجمه، محققان و سازندگان ترجمه ماشینی نهایتاً آرزو دارند که حداقل بتوانند ترجمه‌هایی تولید کنند که در موقعیت‌های ویژه‌ای که مجبور به توضیح دادن اهداف تحقیق هستند مفید واقع شوند و یا کاربردهای مناسب دیگری برای ترجمه‌هایشان بیابند.

ترجمه ماشینی بخشی از حوزه وسیع تر "تحقیق محض" در پردازش زبان طبیعی مبتنی بر کامپیوتر در زبان شناسی رایانه‌ای و هوش مصنوعی می‌باشد که مکانیزم‌های اساسی زبان و ذهن را به وسیله مدلسازی و شبیه سازی در برنامه‌های کامپیوتری مورد بررسی قرار می‌دهد. تحقیق در زمینه ترجمه ماشینی با پذیرش و بکارگیری جنبه‌های نظری وروش‌های عملی در فرآیندهای ترجمه شدیداً با این امور سروکار دارد و به نوبه خود می تواند بینش‌ها و راه‌حل-هایی را برای مشکلات و مسائل خاص خود بیابد. علاوه بر این ترجمه ماشینی می تواند بستر آزمایشی مناسبی را در مقیاس بزرگتر برای نظریه‌ها و روش‌هایی که به وسیله آزمایشات و تحقیقات در مقیاس کوچک در زبانشناسی رایانه ای و هوش مصنوعی انجام می شود فراهم نمایند.

مانع اصلی بر سر راه ترجمه به وسیله کامپیوتر مانع کامپیوتری یا محاسبات نیست بلکه مانع زبانی با زبانشناختی می باشد. مشکلات زبانی عبارتند از: ابهام واژگانی، پیچیدگی نحوی، تفاوت‌های واژگانی بین زبان‌ها، ساختارهای غیر دستوری و مستتر و به طور خلاصه استخراج

معنای جملات و متون از ترجمه نشانه‌های نوشتاری و تولید جملات و متون به مجموعه دیگری از علامات زبانی با معنای معادل. در نتیجه ترجمه ماشینی به شدت متکی به پیشرفت-های تحقیقاتی زبانی و به ویژه شاخه‌هایی از آن با درجات بالایی از فرمالیته بودن است و در حقیقت همیشه همین طور بوده و خواهد بود. اما ترجمه ماشینی نمی‌تواند نظریه‌های زبانی را مستقیماً به کار گیرد. زبان شناسان با شرح مکانیزم زیر ساختی تولید و درک زبان سروکار دارند. آن ها بر روی مشخصه‌های قطعی متمرکزند و نه کوشش برای توصیف یا توضیح در مورد هر چیزی. در مقابل، ترجمه ماشینی با متون قطعی واقعی سروکار دارد. سیستم‌های ترجمه ماشینی با دامنه وسیعی از پدیده های زبانی، پیچیدگی اصلاحات، غلطهای املایی، واژه‌های جدید و جنبه‌های اجرایی که همیشه هم در رابطه با زبان شناسی نظری مجرد نیستند سرو کار دارد.

به طور خلاصه، ترجمه ماشینی به خودی خود یک شاخه مستقلی از تحقیق محض نیست بلکه هرگونه ایده، روش و فنی را که ممکن است در توسعه سیستم‌های پیشرفته سهیم باشد خواه از زبان شناسی یا علم کامپیوتر، هوش مصنوعی و یا نظریه ترجمه باشد، اقتباس می کند. در واقع ترجمه ماشینی تحقیق کاربردی است ، اما شاخه‌ای که از حجم مهمی از فنون و مفاهیم که به نوبه خود می‌توانند در حوزه‌های دیگر پردازش زبان مبتنی بر کامپیوتر به کار روند تشکیل شده است.

محققین و توسعه‌دهندگان امر ترجمه ماشینی برای رسیدن به ترجمه مطلوب نیاز به ابزار-های متنوعی دارند و با مشکلات متعددی روبرو هستند. این مشکلات و نیازمندی‌ها باعث به وجود آمدن روش‌های متعددی[۱۰ و۱۱] برای ترجمه ماشینی شده‌است. اما هر یک از این روش‌ها دارای مشکلات مخصوص به خود بوده و نتوانسته‌اند در همه‌ی موارد ترجمه‌ای صددرصد درست ارائه دهند.

ظهور اینترنت این امکان را فراهم ساخته که به تازه‌ترین دستاوردهای علم و تکنولوژی بشر دسترسی پیدا کنیم، با محققین مختلف در سطح ملی یا بین‌المللی ارتباط برقرار کنیم، از مهمترین تحولات منطقه‌ای و فرامنطقه‌ای اطلاع کسب کنیم، به راحتی با فرهنگ‌ها و آداب ملل مختلف آشنا شویم و بسیاری از موارد دیگر که به علت گستردگی، صحبت در مورد آنها در این‌کار نمی‌گنجد. اما برای دستیابی به این دستاوردها و مطالب ارزنده مشکلات متعددی وجود دارد که شاید بتوان زبان را به عنوان مهمترین سد و مانع در انتقال این اطلاعات نام برد. بنابراین ظهور اینترنت، ضرورت وجود ترجمه ماشینی را در کنار خود مضاعف می‌کند.

ترجمه ماشینی جنبشی است که می‌تواند در زمینه علمی، انقلابی اساسی ایجاد کند و متون علمی و غیرعلمی زبان‌های مختلف دنیا را در اختیار محققان و پژوهشگران رشته‌های مختلف قرار دهد. ظهور اینترنت و ضرورت وجود ترجمه ماشینی در کنار آن، باعث شده است که تلاش‌های تحقیقاتی زیادی در این زمینه انجام، و حجم سرمایه‌گذاری‌ها در این عرصه روزبه‌روز مضاعف گردد.

علاوه بر این حقایق آشکار، ترجمه ماشینی یک زیرشاخه‌ی علمی است که جذابیت و کشش علمی آن برای محققان و توسعه‌دهندگان امر ترجمه ماشینی انکارناپذیر است. و رسیدن به ترجمه ماشینی با خروجی مطلوب همواره یکی از رویاهای دانشمندان این حوزه علمی بوده-است.

نکته‌ی دیگری که سرمایه‌گذاری در این عرصه را توجیه می‌کند وجود مزیت‌های اقتصادی است که در ورای این قضیه نهفته است. در مقر سازمان ناتو در بروکسل و جامعه اروپا علیرغم آنکه تعداد زیادی مترجم ورزیده به کار اشتغال دارند، در حال حاضر از ترجمه ماشینی نیز استفاده می‌شود. می‌توان دلایل متعددی برای این قضیه ذکر کرد که از جمله این موارد سرعت بسیار زیاد و غیرقابل مقایسه یک مترجم ماشینی نسبت به انسان است. میزان کاری که مترجمی ورزیده در خلال چندین روز انجام می‌دهد، توسط کامپیوتر در عرض چند دقیقه انجام می‌شود. از طرف دیگر انسان قادر است مقدار محدودی اطلاعات را به ذهن خویش بسپارد و بر اساس آن‌ها تصمیم اتخاذ می‌کند ، در صورتی که ماشین قادر است بر اساس مقدار اطلاعاتی که در حافظه‌اش ضبط گردیده تصمیم بگیرد و در تصمیم‌گیری تمام جنبه‌های اطلاعاتی خویش را در نظر داشته باشد.

از طرفی دیگر ماشین ترجمه می‌تواند بدون احساس خستگی ۲۴ ساعت در شبانه روز یک دسته عملیات را بی‌وقفه تکرار کند در حالی که تکرار یک رشته عملیات یکنواخت برای مدت طولانی جسم و روح انسان را فرسوده می‌کند.

به این مسائل هزینه بسیار کمتر مترجم ماشینی نسبت به یک مترجم انسانی را نیز باید افزود. حال با توجه به این مزیت‌ها حتی اگر کیفیت و دقت ترجمه ماشینی کمتر از حاصل کار مترجم انسانی باشد، بازهم سرمایه‌گذاری و تلاش شبانه‌روزی در این رشته یک امر کاملاً توجیه‌پذیر است.

متاسفانه این موضوع در کشور ما علی‌رغم این حقایق آشکار، مورد غفلت قرارگرفته است و ابعاد واقعی قضیه و گستره‌ی کاربردی آن نیز نامکشوف مانده است.