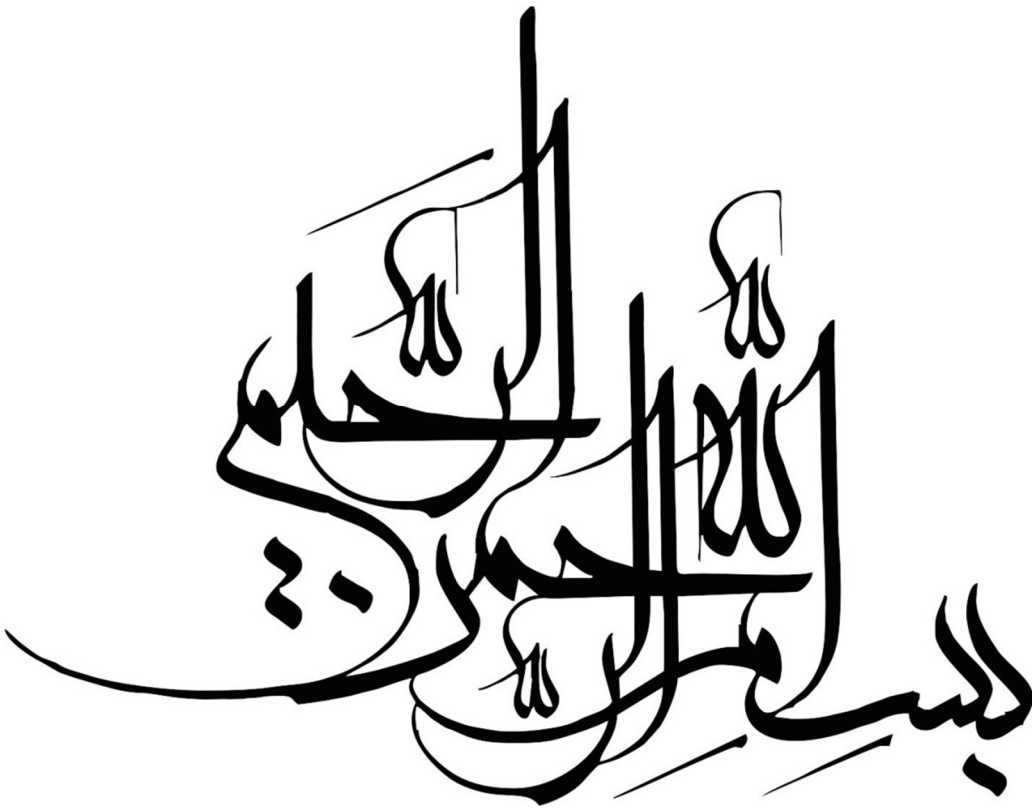
بسم الله الرحمن الرحيم

واحد بین الملل

پایانامه‌ی کارشناسی ارشد در رشته

مهندسی کامپیوتر – نرم‌افزار

تخمین موثری از شباهت متون در ترجمه ماشینی مبتنی بر مثال

به کوشش

رضا اکبری

استاد راهنما

دکتر محمد هادی صدرالدینی

بهمن ۱۳۹۲

به نام خدا

اظهارنامه

اینجانب رضا اکبری (۹۰۸۸۸۵) دانشجوی رشته مهندسی کامپیوتر گرایش نرم‌افزار دانشکده‌ی مهندسی اظهار می‌کنم که این پایانامه 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 پایانامه‌ام تکراری نیست و تعهد می‌نمایم 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: رضا اکبری

تاریخ و امضاء: ۹۲/۱۲/۱۸

به نام خدا

تخمین موثری از شباهت متون در ترجمه ماشینی مبتنی بر مثال

به کوشش

رضا اکبری

پایان‌نامه

ارائه شده به تحصیلات تکمیلی دانشگاه به عنوان بخشی از فعالیت‌های تحصیلی لازم

برای اخذ درجه‌ی کارشناسی ارشد

در رشته‌ی

مهندسی کامپیوتر-نرم افزار

از دانشگاه شیراز-واحد بین الملل

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته پایان‌نامه با درجه: عالی

دکتر محمد هادی صدرالدینی، دانشیار بخش مهندسی و علوم کامپیوتر (استاد راهنما) .....

دکتر سید مصطفی فخراحمد، استادیار بخش مهندسی و علوم کامپیوتر (استاد مشاور) .....

دکتر فرشاد خون‌جوش، استادیار بخش مهندسی و علوم کامپیوتر (داور متخصص داخلی) .....

بهمن ۱۳۹۲

تقدیم به منجی عالم بشریت حضرت مهدی (عج)

سپاسگزاری

به نام آنکه جان را فکرت آموخت

چراغ دل به نور جان بر افروخت

سپاس خدایی را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت‌های او ندانند و کوشندگان، حق او را گزاردن نتوانند. اکنون که به مدد ایزد توانا، توانسته‌ام رساله کارشناسی ارشد خود را فراهم آورم، بر خود می‌دانم کمال تقدیر و تشکر را از استاد با کمالات و شایسته، جناب آقای دکتر محمد هادی صدرالدینی که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این رساله را بر عهده گرفتند، و همچنین از استاد صبور و با تقوا، جناب آقای دکتر سید مصطفی فخراحمد که زحمت مشاوره این رساله را در حالی متقبل شدند که بدون مساعدت ایشان، این پروژه به نتیجه مطلوب نمی‌رسید، نهایت تشکر را داشته باشم. "من ستایشگر معلمی هستم که اندیشیدن را به من آموخت، نه اندیشه‌ها را".

و نیز از استاد فرزانه و دلسوز، جناب آقای دکتر خون‌جوش که زحمت داوری این رساله را متقبل شدند و نیز از نماینده محترم تحصیلات تکمیلی، کمال تشکر و قدردانی را می‌نمایم.

همینطور از دوستان و خانواده عزیزم که با مساعدت‌های ارزشمند خود بنده را در پیشبرد پایان‌نامه یاری نمودند، کمال سپاسگزاری را دارم. بر خود نیز لازم می‌دانم از همه اساتید و دانشجویانی که از طروق متفاوت بنده را راهنمایی کردند، و به ابهامات من پاسخ دادند، سپاسگزاری کنم.

باشد که این نوشته کوتاه، بخشی از زحمات آنان را سپاس گوید.

چکیده

تخمین موثری از شباهت متون در ترجمه ماشینی مبتنی بر مثال

به کوشش
رضا اکبری

ترجمه ماشینی یکی از مهمترین شاخه‌های تحقیقاتی در زمینه پردازش زبان طبیعی می‌باشد. ترجمه ماشینی عبارت است از برگردان متنی از یک زبان به زبان دیگر توسط ماشین به طوری که مفهوم متن در زبان مبدأ بدون تغییر به زبان مقصد منتقل شود. یکی از سیستم‌های ترجمه ماشینی، سیستم مبتنی بر مثال می‌باشد. در این رویکرد برای ترجمه یک ترکیب از کلمات، به مجموعه متون ترجمه شده قبلی مراجعه می‌شود تا بجای ترجمه لفظ به لفظ و نامتعارف، یک ترجمه مشابه ترجمه انسانی بدست آید.در این رساله، ما یک مدل را در جهت اندازه‌گیری میزان شباهت دو جمله در ترجمه ماشینی مبتنی بر مثال ارائه کرده‌ایم. در مدل ارائه شده از الگوریتم ژنتیک و یک تابع برازندگی جدید که مبتنی بر بار معنایی منطبق شده بین دو جمله می‌باشد، استفاده گردیده است. ما فعل‌ها را به عنوان قلب یک جمله در نظر گرفته‌ایم چون بخش اساسی یک جمله بشمار می‌آیند و مقادیر زیادی از بار معنایی جمله را حمل می‌کنند. بنابراین ما در تابع برازندگی ارائه شده توجه بیشتر خود را بر روی افعال جمله قرار داده‌ایم. نتایج حاصل از معیارهای اندازه‌گیری Precision و Recall نوید دهنده آن است که متد ارائه شده، کیفیت جملات منطبق شده بازیافتی را بهبود بخشیده است.

کلمات کلیدی: ترجمه ماشینی، ترجمه ماشینی مبتنی بر مثال، فاز تطابق، تشابه جملات، بار معنایی، الگوریتم ژنتیک.

فهرست مطالب

عنوان **صفحه**

عنوان **صفحه**

فصل سوم: مروری بر کارهای مرتبط

فصل چهارم: ابزارها و روابط معنایی

عنوان **صفحه**

فصل پنجم: راهکار پیشنهادی

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فصل اول

مقدمه

۱-۱- شرح مسئله

ترجمه‌ی ماشینی[1] زیر شاخه‌ای از زبان‌شناسی محاسباتی می‌باشد که عبارت است از ترجمه‌ی متنی از یک زبان طبیعی به زبانی دیگر، توسط کامپیوتر. در سطح مقدماتی، ترجمه ماشینی یک جایگزینی ساده برای کلمات از زبان طبیعی به زبان دیگری است. با استفاده از تکنیک‌های زبان‌شناسی پیکره‌ای[2]، ترجمه‌های پیچیده بیشتری قابل دسترسی هستند. همچنین این تکنیک‌ها کنترل بهتر تفاوت‌های گونه شناسی در زبان، تشخیص عبارت و ترجمه‌ی اصطلاحات را به خوبی و درستی جدا کردن عبارات نامتعارف در متن، مقدور می‌سازند.

بطورکلی ترجمه ماشینی، به معنای ترجمه خودکار جملات از یک زبان طبیعی به زبان دیگر است. به گونه‌ای که فرد بومی زبان مقصد، همان مفهومی را دریافت کند که گوینده مدنظر داشته است [۱]. اگر چه قدمت ترجمه ماشینی به سال ۱۹۴۰ بر می‌گردد [۲]، اما همچنان به عنوان یک شاخه زنده، رو به رشد و پویا به شمار می‌رود. گروه‌های بسیاری، در دانشگاه‌ها و مؤسسات تحقیقاتی در سراسر دنیا مشغول انجام تحقیقات در این زمینه می‌باشند.

دلیل اصلی پویایی این شاخه، اهمیت و کاربرد ترجمه برای افراد و سازمان‌ها است که ضرورت دستیابی به روش‌های بهتر و سریع‌تر برای ترجمه را ایجاب می‌کند. با گسترش روابط و حرکت به سوی ارتباطات گسترده‌تر جوامع، لزوم دسترسی آسان به ترجمه متون و گفتار، از زبانی به زبان دیگر، ضروری می‌نماید. از طرفی دیگر، با توجه به هزینه بالا و محدودیت در

[1] Machine Translation
[2] Corpus Linguistics

امکان استفاده از مترجم انسانی، اهمیت مطالعه و تحقیق بر روش‌هایی برای ایجاد امکان ترجمه خودکار توسط ماشین، مشخص می‌شود.

گذشته از گروه‌های تحقیقاتی در سطح دانشگاه‌ها، سازمان‌های زیادی به مبحث ترجمه ماشینی ابراز علاقه نموده‌اند. یکی از این سازمان‌ها، وزارت دفاع آمریکا است که محور فعالیت خود را، بر روی زبان کشورهایی قرار داده که از نظر نظامی، سیاسی یا اقتصادی، تهدیدی برای ایالات متحده به حساب می‌آیند[۳].

به علاوه، کاربرد گسترده ترجمه در بحث ارتباطات، اهمیت ترجمه ماشینی و لزوم بهبود آن را دوچندان می‌کند. صفحات اینترنتی حاوی اطلاعات بسیار زیادی هستند که می‌توانند برای هر فرد، در هر موضوعی، راهگشا باشند. یک مترجم ماشینی ساده، می‌تواند ایده کلی یک وب سایت را برای افراد مختلف در زبان‌های متفاوت، بیان کند.

در ترجمه ماشینی دو رویکرد اصلی وجود دارد:

- روش‌های مبتنی بر انتقال
- روش‌های مبتنی بر داده

در روش‌های مبتنی بر انتقال، علاوه بر مفهوم، ساختار جمله نیز از زبان مبدأ به زبان مقصد منتقل می‌شود. یکی از دشواری‌های این روش، لزوم تسلط کامل بر قواعد زبان مبدأ و زبان مقصد است. دو ایده اصلی در این زیر شاخه مطرح است: استفاده از زبان میانی[۱] و روش‌های قانون محور[۲].

دو رویکرد اصلی برای روش مبتنی بر داده، روش آماری و روش مبتنی بر مثال[۴] می‌باشد. ایده اولیه در روش مبتنی بر مثال، استفاده از ترجمه‌های انسانی موجود برای ترجمه متن‌های جدید است. لذا کافی است متون جدید به قطعه‌های کوچک شکسته شود و ترجمه معادل این قطعات، در پایگاه داده‌ای از قطعات ترجمه شده توسط انسان جستجو شده و ترجمه مورد نظر تولید گردد.

روشن است که الگوریتم تطبیق در این سیستم ترجمه دارای اهمیت فراوانی می‌باشد. از آنجا که تطبیق رشته‌ها در این سیستم‌ها به صورت تطبیق دقیق است. لذا ممکن است جمله‌ای

[1] Interlingua
[2] Rule-based

در پیکره متنی موجود باشد که فاصله بسیار کمی با جمله یا عبارت ورودی داشته باشد اما بدلیل عدم تطبیق دقیق، سیستم قادر به ارائه پاسخ مناسب نباشد. این فاصله می‌تواند به دو صورت معنایی و ساختاری باشد.

این روش دارای محدودیت دادگان می‌باشد. جمع‌آوری مجموعه مثال‌های بسیار بزرگ نیز کل زبان را پوشش نمی‌دهد. از طرفی نیز با بزرگ بودن پیکره متنی دو زبانه و با توجه به اینکه سیستم‌های ترجمه مبتنی بر مثال بر مبنای مقایسه عمل می‌کنند، لذا این سیستم‌ها کند بوده و پردازش‌ها بسیار زمان‌بر هستند[۵]. بنابراین معمولاً این روش برای زیر مجموعه‌های محدودی از یک زبان استفاده می‌شود. به دلیل استفاده مستقیم از ترجمه انسان، بهره‌مندی ترجمه نهایی از ظرافت‌های ترجمه که تنها یک انسان می‌تواند تولید نماید، از مزایای این روش بشمار می‌رود، بطوریکه ارتقاء کیفیت نتیجه نهایی چشمگیر خواهد بود. اما هر چه قدر هم پیکره متنی بزرگ باشد باز هم ممکن است ترجمه جمله ورودی به خوبی قابل استخراج نباشد. اگر بتوانیم از تطابق تقریبی جملات استفاده نماییم، دقت ترجمه به مراتب بهتر خواهد شد. مزیت‌های گفته شده در بالا، علاقه‌مندان به این سیستم را بر آن داشت که فرم ساخت یافته-تری از آن را مطرح نمایند و این ایده اولیه روش‌های مبتنی بر عبارت در ترجمه آماری می-باشد.

۱-۲- انگیزش تحقیق

یکی از مهمترین فازهای ترجمه ماشینی مبتنی بر مثال، فاز تطابق می‌باشد که به عنوان فاز اول و حیاتی ترجمه زبان مبدأ به زبان مقصد با استفاده از این سیستم قلمداد می‌گردد. زیرا در این فاز شبیه‌ترین جملات موجود در قسمت زبان مبدأ پیکره دو زبانه نسبت به جمله ورودی جستجو شده و به عنوان جملات کاندید انتخاب می‌شوند. سپس از ترجمه کاندیدهای انتخاب شده موجود در قسمت زبان مقصد پیکره دو زبانه، برای ترجمه جمله ورودی استفاده می‌گردد. حال اگر از متد نامناسبی در این فاز برای تعیین شباهت استفاده نماییم باعث می‌شود که با

وجود فازهای قوی دیگر در این سیستم، ترجمه مناسبی برای جمله ورودی حاصل نشود. تعیین شباهت دو جمله از نظر معنا با توجه به قدرت زبان طبیعی و متغیر بودن بیان عبارات، کاری دشوار است. انگیزه اصلی این پژوهش بررسی و ارائه راهکاری برای مسئله محاسبه شباهت معنایی و ساختاری، و استفاده از آن در عملیات تطابق تقریبی سیستم‌های ترجمه ماشینی مبتنی بر مثال می‌باشد.

۱-۳- ساختار پایان‌نامه

در فصل دوم ابتدا به تعریف ترجمه ماشینی و تاریخچه آن پرداخته شده است. سپس فاز تطابق که مهمترین فاز ترجمه ماشینی مبتنی بر مثال می‌باشد، تشریح شده است. همچنین گذری به معیارهای ارزیابی از قبیل ارزیابی فنی و کارایی پرداخته شده است. در فصل سوم به بررسی کارهای مرتبط با تشریح کامل روش‌های تطابق جملات و شباهت معنایی جملات پرداخته شده است. ابزارهایی مانند وُردنِت و ریشه یاب و همچنین مجموعه داده‌ها و روابط معنایی مانند فرکانس، هایپرنیم و ترادف و ... در فصل چهارم این رساله توضیح داده شده است. در فصل پنجم رویکرد ارائه شده از جمله مرحله پیش پردازش جملات و همچنین پیاده سازی الگوریتم با استفاده از الگوریتم ژنتیک با در نظر گرفتن عملگرهای جهش و ترکیب، شرح داده شده است. در این فصل همچنین به ارزیابی متد خود در مقایسه با روش‌های استاندارد بازیابی اطلاعات و متدهای جدید ارائه شده دیگر توسط سایر همکاران، پرداخته شده است.

فصل دوم

تعاریف و تشریح مفاهیم

۲-۱- مقدمه

در این فصل ترجمه ماشینی و تاریخچه آن را مورد بررسی قرار خواهیم داد. و انواع ترجمه ماشینی را تحلیل خواهیم کرد. در گام بعد تمرکز بیشتر خود را بر روی ترجمه ماشینی مبتنی بر مثال و فاز تطابق جملات در این سیستم و چالش‌های مربوط به آن خواهیم گذاشت. سپس به بررسی مسئله دشوار و پیچیده محاسبه شباهت جملات در این فاز، با توجه به عبارات و اصطلاحات مختلف ولی هم معنی موجود در جملات، خواهیم پرداخت.

۲-۲- ترجمه ماشینی

با توجه به پیشرفت و گسترش کاربردهای رایانه در علوم مختلف، نیاز به استفاده از توانایی‌های آن، در حوزه‌ی زبان شناسی نیز به شدت احساس می‌شود. حوزه‌های پردازش زبان طبیعی[۱] و زبان شناسی رایانه‌ای[۲] به تلاش برای ماشینی کردن فرآیند زبان شناسی سنتی می‌پردازند. منشأ پیدایش زبان شناسی رایانه‌ای را می‌توان هم‌زمان با شکل گیری تلاش‌هایی برای تولید ماشین ترجمه‌ی خودکار در دهه‌ی ۵۰ میلادی در ایالات متحده آمریکا دانست [۶]. این ماشین ترجمه‌ی خودکار، قرار بود مجلات علمیِ روسی را به انگلیسی ترجمه کند اما با شکست این پروژه، مشخص شد که پردازش خودکار زبان طبیعی بسیار پیچیده‌تر از آن است که

[۱] Natural Language Processing
[۲] Computational Linguistics