



۱۳۵۹
وحد، مشهد



مرکز آموزش عالی علمی- کاربردی
جهاد دانشگاهی مشهد

ترجمه ماشینی آماری

رشته تحصیلی : برنامه نویسی تحت وب

استاد راهنما :

دکتر محمد عبدالهی

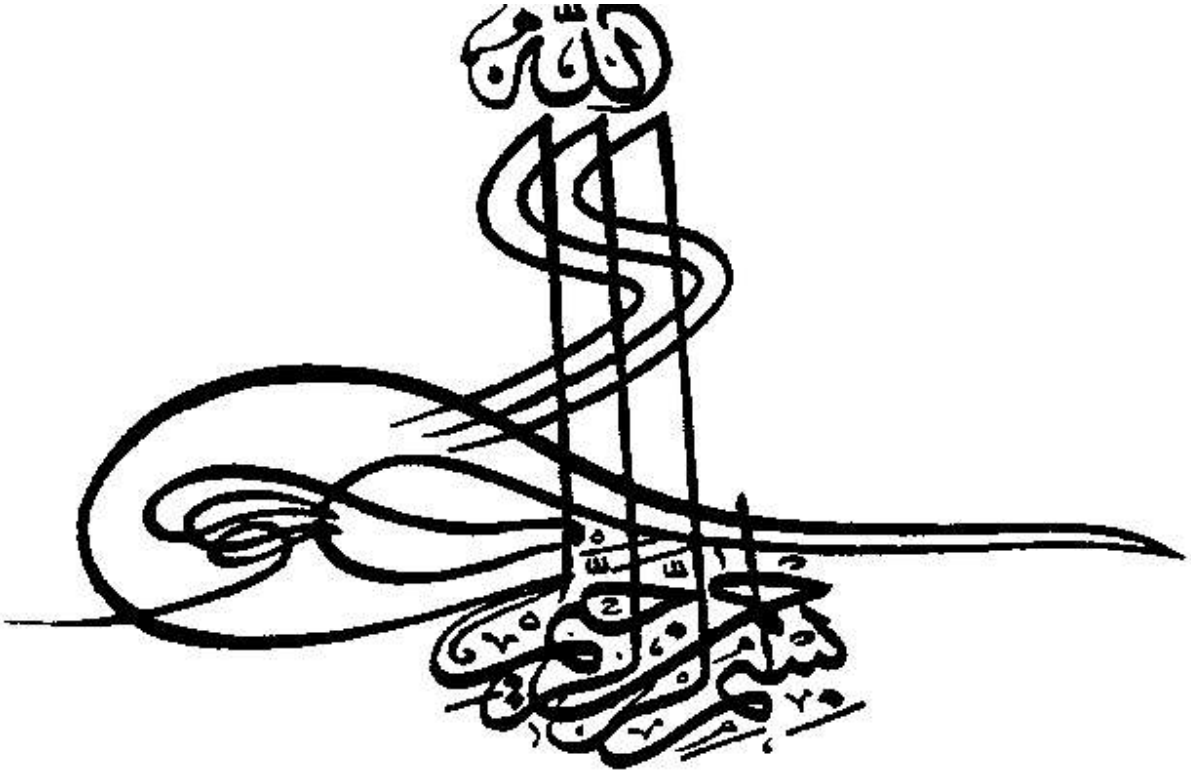
تهیه و تنظیم :

محبوبه جباری , ملیحه مصلی , زهرا جهانزاده

استاد مشاور:

مهندس نیک فرجام

تابستان 1395



تقدیر و تشکر

سپاس گذار استادی، هستیم که اندیشیدن را به ما آموخت، نه اندیشه را... .

از جناب آقای دکتر محمد عبداللہی کہ طی تکمیل این پایان نامہ دلسوزانہ راہنما و مشاور ما بودند و راہنمایی ما و مشاورہ های ایشان، ہموارہ کجاک

و کارکشای ما بودہ است، کمال تشکر را داریم و ہم چنین جناب آقای مهندس نیک فرجام کہ در کرد آوری این مجموعہ مشاور و

راہنمای ما بودند تشکریم.

چکیده

با توسعه‌ی روز افزون تعاملات اجتماعی و میزان اطلاعات و حجم داده‌ها برقراری ارتباط یکی از موضوعات بسیار مهم در زندگی روزمره انسان‌ها می‌باشد. مشکل عمده‌ای که در این زمینه وجود دارد عدم برقراری ارتباط یا استفاده از داده‌ها و اطلاعات به زبان دیگر است.

بنابراین مهمترین مسئله بوجود آمده در زندگی امروزه ارائه راه حل و راهکاری جهت ترجمه یک زبان به زبان دیگر است.

ترجمه ماشینی راهکاری است که برای حل این مشکل ارائه شده است و به خاطر اهمیت آن در دنیای امروز توجه بسیار زیادی به آن شده است.

ترجمه ماشینی، ترجمه‌ای است که توسط کامپیوتر وبدون دخالت انسان انجام می‌شود که زیر شاخه‌ای از زبان شناسی محاسباتی بوده و در آن یک متن از یک زبان طبیعی به زبان دیگر توسط کامپیوتر ترجمه می‌شود.

در این پروژه تحقیقاتی استفاده از تکنیکهای زبان شناسی مد نظر است که امکان ترجمه متون پیچیده تر شامل تشخیص عبارات، ترجمه اصطلاحات، تشخیص عبارات نا متعارف با کیفیت بسیار بالاتر را با استفاده از برنامه های کامپیوتری فراهم می‌کند که البته لازمه این کار رمز گشایی از متن مبدا و کد گذاری مجدد از مفاهیم درک شده به زبان مقصد می‌باشد.

مهمترین چالش ترجمه ماشینی تحلیل متن همانند انسان وتولید متن جدید با توجه به معیارهای زبان مقصد است که امروزه با توسعه جهانی اینترنت و شبکه های مختلف اجتماعی بستر بزرگی برای نقل وانتقال اطلاعات ،افکار و فرهنگ ایجاد شده است.

ترجمه ماشینی آماری به عنوان یکی از بهترین روش های ممکن برای ترجمه ی زبان مبدا به زبان مقصد می باشد.

در واقع ترجمه ماشینی آماری برای زبان‌هایی که از لحاظ ساختاری بهم نزدیک هستند خروجی یا به عبارتی ترجمه مناسبی را ارائه می‌دهد، اما برای برخی از زبانها مانند زبان های انگلیسی و فارسی به علت تفاوت‌های ساختاری دوزبان وعدم وجود پیکره دوزبانه بزرگ موجب شده است که این روش برای ترجمه فارسی به انگلیسی ویا بالعکس ترجمه‌ی مطلوب و قابل قبولی را تولید نکند.

در این پایان نامه سعی می‌شود در مورد این موانع ومشکلات با استفاده از اطلاعات زبان شناسی راه حلی مناسبی ارائه گردد.

کلمات کلیدی :

ترجمه ماشینی، ترجمه ماشینی آماری، مدل‌های زبانی، مدل ترجمه

فصل اول:

مقدمه

پیشگفتار :

وجود زبان‌های مختلف ورشد وگسترش تعاملات بین‌المللی در زمینه‌های متفاوت در جای‌جای دنیا مشکلات فراوانی را برای افراد به منظور برقراری ارتباط و تعامل با یکدیگر ایجاد کرده است.

از طرفی نمی‌توان برای رفع این معضل آموزش زبان‌های گوناگون برای همه افراد یا جوامع اجباری نمود و همچنین دسترسی به افراد مترجم در همه جا و در هر زمان مورد نیاز امکان پذیر نمی‌باشد و در اینجاست که استفاده از کامپیوتر بسیار مورد نیاز می‌باشد. به این نوع ترجمه که توسط کامپیوتر صورت می‌پذیرد ترجمه ماشینی گفته می‌شود.

در واقع اولین اقدامات در زمینه ترجمه ماشینی از سال 1940 شروع و تا به امروز ادامه داشته و در این زمینه پیشرفت‌های بسیار زیادی بدست آمده است.

برای ایجاد یک مترجم ماشینی اصولاً از دو رویکرد استفاده می‌شود که عبارتند از :

1- رویکرد مبتنی بر قانون

2- رویکرد مبتنی بر پیکره

در رویکرد نخست براساس زبان مبدأ و زبان مقصد یکسری قانون‌ها تدوین شده و براساس آن عمل ترجمه انجام می‌پذیرد که یکی از محدودیت‌ها و موانع اصلی آن همین محدود بودن آن به قوانین می‌باشد.

و اما رویکرد دوم : این رویکرد بر اساس نمونه‌های قبلی و ترجمه‌های انسانی انجام شده به ترجمه متون جدید می‌پردازد. در این رویکرد نیازی به قوانین برای ترجمه نیست بلکه نیازمند به یک پیکره موازی و دوزبانه هستیم.

یکی از روش‌های پر اهمیت در این رویکرد، روش ترجمه آماری می‌باشد که به علت عملکرد بسیار مناسب آن ، در سال‌های اخیر بسیار مورد توجه قرار گرفته است.

هدف این روش این است که از روی پیکره‌های موجود ، مدل‌های لازم و مورد نیاز استخراج گردد. این مدل‌ها تشکیل شده است از مدل زبانی و مدل ترجمه. از این رو می‌توان به سهولت و بدون نیاز به دانش زبانی زیاد از این مزیت استفاده نمود.

البته آنچه که در این روش حائز اهمیت بوده این است که یک پیکره بزرگ داشته باشیم. اگر بخواهیم دانش زبانی که بصورت ضمنی داخل پیکره موازی موجود است را استخراج کنیم و در قالب مدل‌های زبانی و ترجمه‌ای بیان کنیم، نیازمند

یک پیکره بزرگ موازی هستیم. در واقع مشکل و مانع اصلی که در این روش وجود دارد این است که ما دسترسی به این پیکره را نداریم.

برای ایجاد این پیکره در ابتدا نیاز به منابع لازم برای ایجاد آن و همچنین روشی برای ترازبندی جمله به جمله در این پیکره هستیم.

در واقع چالشهای بزرگ برای ایجاد و یا تولید پیکره‌های موازی از این قرار است:

1- وجود منابع لازم برای ساخت و ایجاد این پیکره

2- روشی برای ترازبندی جمله به جمله برای پیکره ایجاد شده

بنابراین باید راهکارهایی بیابیم که با استفاده از پیکره‌های کوچکتر بتوانیم خروجی بهتر و مناسب‌تری تولید نماییم. برای این کار باید برخی اطلاعات زبانی را بطور واضح درون این پیکره وارد کرده تا بخشی از مشکلات موجود بخاطر کوچک بودن پیکره مرتفع گردد.

تحقیقات زیادی در این زمینه در مورد زبان‌های مختلف دنیا انجام گرفته است البته هنوز کار و تحقیق زیادی در مورد زبان فارسی انجام نشده است و بیشتر کارها انجام شده و تحقیقات در زبان فارسی ایجاد سیستم ترجمه آماری صرف می‌باشد. البته باید این نکته را در نظر گرفت که استفاده از اطلاعات زبانی در مترجم‌های ماشینی آماری بسته به زبان‌های مورد استفاده در سیستم مترجم متفاوت خواهد شد.

زبان فارسی جزء زبان‌هایی است که ساختار پیچیده‌ای دارد. در واقع این زبان تصریفی (infelective) بوده و کلمات بسته به شخص، زمان، ضمائر مربوط به آن با فرم‌های مختلف در جمله نمایان می‌شوند. همین امر باعث شده که ترجمه زبان فارسی به زبان‌های دیگر با چالش روبرو گردد.

به طور مثال ترجمه فارسی به انگلیسی و یا بالعکس باعث بروز مشکل برای ایجاد و تولید ترجمه با ترتیب و ترازبندی مناسب می‌شود. همانطور که قبلاً گفته شد در صورتی که پیکره موازی ورودی به قدر کافی بزرگ باشد می‌توان بر این گونه موانع و مشکلات فائق آمد اما ایجاد چنین پیکره‌ای بسیار سخت و دشوار است.

طرح پیشنهادی

در این پایان نامه استفاده از بعضی اطلاعات زبان‌شناسی برای رفع برخی از مشکلات به عنوان یکی از اهداف اصلی مطرح شده است. به صورتی که بتوان با استفاده از پیکره‌های کوچکتر موجود عمل ترجمه را بهبود بخشید. برای رسیدن به این هدف دو راهکار پیشنهاد شده است.

راهکار نخست: کاهش تفاوت ساختاری میان جملات فارسی و انگلیسی می‌باشد. منظور این است که بتوان ساختار جملات زبان فارسی و انگلیسی را از نظر نحوی مشابه هم نماییم. این کار باعث می‌شود که عباراتی که ترجمه یکدیگر هستند در یک توالی یکسان در دو جمله هم‌تراز فارسی و انگلیسی قرار بگیرند. البته لازمه این کار این است که ساختار هر دو زبان مورد بررسی قرار گرفته و یکسری قوانین برای تبدیل ساختاری استخراج گردد.

این تبدیلات در سمت زبان انگلیسی (زبان مبدأ) صورت می‌گیرد. با اجرا کردن چنین تغییراتی انتظار می‌رود که مدل ترجمه ایجاد شده نسبت به قبل بهبود یابد. این بهینه‌سازی در نهایت منجر به بهبود نهایی مترجم ماشینی می‌شود.

و اما راهکار دوم: این راهکار از اطلاعات زبانی بیشتر درون پیکره استفاده می‌کند. در حالت معمول درون پیکره فقط جملات انگلیسی (زبان مبدأ) و جملات فارسی (زبان مقصد) وجود دارند. همانطور که گفته شد بسیاری از لغات و عبارات در زبان فارسی با ساخت و صرف‌های متفاوت از یک عبارت، متفاوت از یکدیگر در نظر گرفته می‌شوند و ارتباطی بین آنها برقرار نمی‌شود. به همین دلیل می‌توان اطلاعات زبانی دیگر نظیر ریشه (stem) کلمات و برچسب‌های بخش‌های سخن (part of speech) را هم داخل پیکره وارد نمود. این اطلاعات قادر است تا حدی مشکلاتی از این دست را مرتفع کند و ما را در مراحل ایجاد مدل‌های مورد نیاز برای ترجمه یاری کند. نتایج بدست آمده از آزمایشات نمایانگر بهبود روش‌های پیشنهادی در نتایج نهایی ترجمه می‌باشد. ارزیابی بر روی نتایج در قالب دو معیار BLEU و همچنین NIST صورت گرفته که نتایج بدست آمده، افزایش هر دو معیار را نشان می‌دهد.

ساختار پایان نامه

مطالب این پایان نامه شامل چهار بخش دیگر می‌باشد که به شرح زیر است:

فصل دوم (مرور ادبیات): در این بخش ابتدا مروری بر مشکلات ترجمه ماشینی خواهیم داشت و در ادامه به روش‌های گوناگون ترجمه ماشینی اشاره شده است. همچنین برخی تعاریف در زمینه پردازش زبان طبیعی نیز در این فصل ارائه می‌گردد و در انتهای این فصل مروری بر کارهای مربوطه داریم.

فصل سوم (روش‌های پیشنهادی): در این فصل دو روش برای ترکیب اطلاعات ترجمه آماری و زبان‌شناسی ارائه گردیده است. این روش‌ها بصورت اعمال پیش پردازشی بر روی پیکره صورت می‌پذیرد تا در مرحله ایجاد مدل‌های مورد نظر گردند.

فصل چهارم (پیاده‌سازی و ارزیابی): در این قسمت ابتدا مراحل و نحوه‌ی ساخت پیکره دوزبانه مورد بررسی قرار می‌گیرد سپس نحوه‌ی پیاده‌سازی روش‌های ارائه شده را بیان می‌کند و در انتها نتایج حاصل از ترجمه در قالب معیارهای BLEU و NIST بیان می‌شود.

فصل پنجم (نتیجه‌گیری): در فصل پایانی این پایان نامه به نتیجه‌گیری روش‌های ارائه شده می‌پردازد و پیشنهادهایی برای کارهای آتی ارائه می‌دهد.