۱۳۵۹
واحد مشهد

مرکز آموزش عالی علمی- کاربردی
جهاددانشگاهی مشهد

ترجمه ماشینی آماری

رشته تحصیلی : برنامه نویسی تحت وب

استاد راهنما :

دکتر محمد عبدالهی

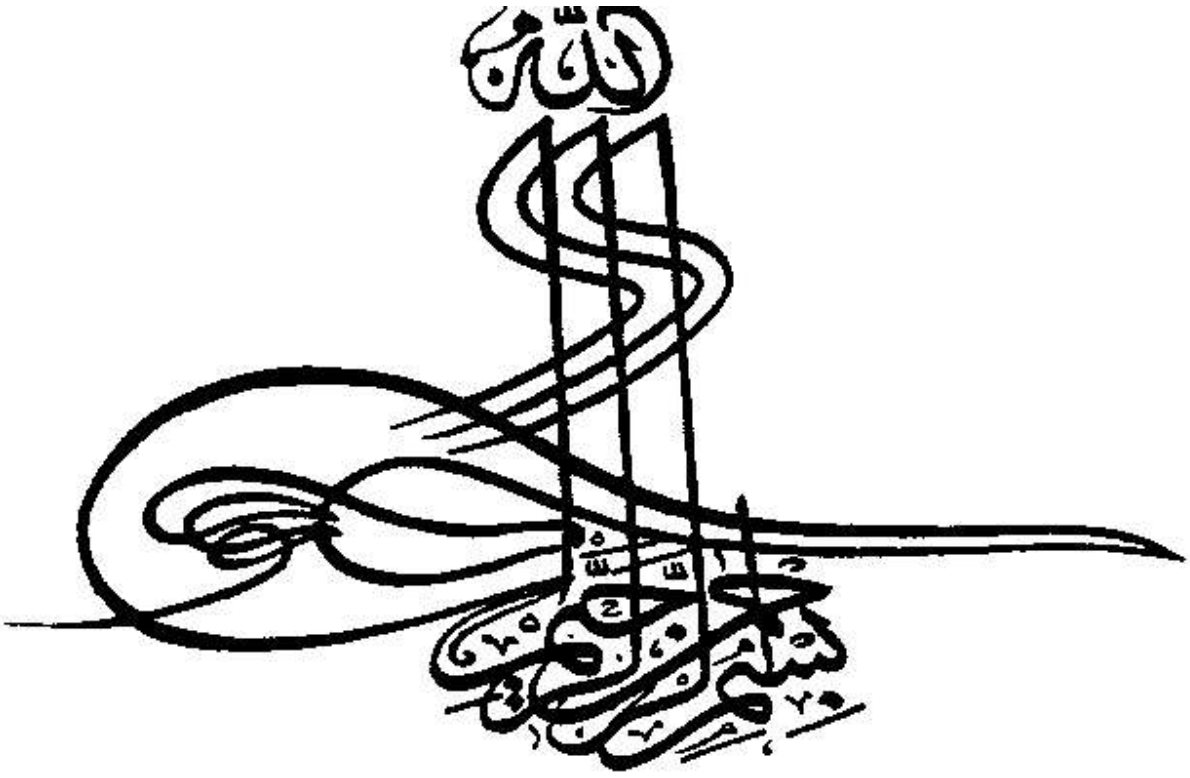
تهیه وتنظیم :

محبوبه جباری, ملیحه مصلی, زهرا جهاندیده

استاد مشاور:

مهندس نیک فرجام

تابستان 1395



تقدیر و تشکر

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

از جناب آقای دکتر محمد عبداللهی که طی تکمیل این پایان نامه دلسوزانه راهنما و مشاور ما بودند و راهنمایی ها و مشاوره های ایشان همواره کمک و کارگشای ما بوده است، کمال تشکر را داریم و هم چنین جناب آقای مهندس نیک فرجام که در گرد آوری این مجموعه مشاور و راهنمای ما بودند متشکریم.

چکیده

با توسعه‌ی روز افزون تعاملات اجتماعی و میزان اطلاعات و حجم داده ها برقراری ارتباط یکی از موضوعات بسیار مهم در زندگی روزمره انسان ها می‌باشد.مشکل عمده‌ای که در این زمینه وجود دارد عدم برقراری ارتباط یا استفاده از داده‌ها واطلاعات 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 محاسباتی‌بوده ودر آن‌یک متن از یک زبان طبیعی به زبان دیگر توسط کامپیوتر ترجمه می‌شود.

دراین 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 بالاتررا بااستفاده از برنامه های کامپیوتری فراهم می‌کند که البته لازمه این کار رمز گشایی از متن مبدا وکد گزاری 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 درمورد این موانع ومشکلات با استفاده ا 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 بین المللی در زمینه های متفاوت در جای جای دنیا مشکلات فراوانی را برای افراد به منظور برقراری ارتباط‌وتعامل 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 جاودر هر زمان مورد نیاز امکان پذیر نمی‌باشد ودر 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 وتا به امروز ادامه داشته ودر این زمینه پیشرفت های بسیار زیادی بدست آمده است.

برای ایجاد یک مترجم ماشینی اصولا از دو رویکرد استفاده می شود که عبارتند از :

1- رویکرد مبتنی بر قانون
2- رویکرد مبتنی بر پیکره

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

و اما رویکرد دوم : این رویکرد بر اساس نمونه‌های قبلی وترجمه‌های 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 می‌باشدکه 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 ومورد نیاز استخراج گردد.این مدل‌ها تشکیل شده است از مدل زبانی و مدل ترجمه.از این رو می توان به سهولت وبدون 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 ودر قالب مدل‌های زبانی و ترجمه‌ای بیان کنیم,نیازمند

یک پیکره بزرگ موازی هستیم.در واقع مشکل ومانع 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 وهمچنین روشی برای ترازبندی جمله به جمله در این پیکره هستیم.

درواقع چالشهای بزرگ برای ایجاد ویا تولید پیکره‌های موازی از این قرار است:

1- وجود منابع لازم برای ساخت وایجاد این پیکره
2- روشی برای ترازبندی جمله به جمله ,برای پیکره ایجاد شده

بنابراین بایدراهکارهایی بیابیم که با استفاده از پیکره های کوچکتر بتوانیم خروجی بهتر ومناسب‌تری تولید نماییم.برای این کار باید برخی اطلاعات زبانی را بطور واصح درون این پیکره وارد کرده تا بخشی از مشکلات موجود بخاطر کوچک بودن پیکره مرتفع گردد.

تحقیقات زیادی در این زمینه درمورد زبان های مختلف دنیا انجام گرفته است البته هنوز کار وتحقیق زیادی در مورد زبان فارسی انجام نشده است وبیشتر کارها انجام شده وتحقیقات 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 پیچیده ای دارد . در واقع این زبان تصریفی(infelective) بوده وکلمات بسته به شخص,زمان,ضمایر 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 ویا بالعکس باعث بروز مشکل برای ایجاد وتولید ترجمه با ترتیب وتررازبندی مناسب می‌شود.همانطور که قبلا گفته شد در صورتی که پیکره موازی ورودی به قدر کافی بزرگ باشد می‌توان براین گونه موانع ومشکلات فائق آمد اما ایجاد چنین پیکره ای بسیار سخت ودشوار 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 بااستفاده 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 وانگلیسی می‌باشد.منظور این است که بتوان ساختار جملات زبان فارسی وانگلیسی را از نظر نحوی مشابه هم نماییم.این کار باعث می‌شود کهعباراتی که ترجمه یکدیگر هستند در یک توالی یکسان در دو جمله همتراز فارسی وانگلیسی قرار بگیرند.البته لازمه این کار این است که ساختار هر دوزبان مورد بررسی قرار گرفته ویکسری 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.

واما راهکار دوم:این راهکار از اطلاعات زبانی بیشتر درون پیکره استفاده می‌کند.در حالت معمول درون پیکره فقط جملات انگلیسی(زبان مبدأ) وجملات فارسی(زبان مقصد)وجود دارند. همانطور که گفته شدبسیاری از لغات وعبارات در زبان فارسی با ساخت وصرف‌های متفاوت از یک عبارت,متفاوت از یکدیگر در نظر گرفته می‌شوندوارتباطی بین آنها برقرارنمی‌شود به همین دلیل می‌توان اطلاعات زبانی دیگر نظیر ریشه (stem)کلمات وبرچسب‌های بخش‌های سخن(part of speed) را هم داخل پیکره وارد نمود.این اطلاعات قادر است تا حدی مشکلاتی از این دست را مرتفع کند ومارا در مراحل ایجاد مدل‌های مورد نیاز برای ترجمه یاری کند.نتایج بدست آمده از آزمایشات نمیانگر بهبود روش‌های پیشنهادی در نتایج نهایی ترجمه می‌باشد.ارزیابی برروی نتایج در قالب دو معیار BLEUوهمچنینNIST صورت گرفته که نتایج بدست آمده ,افزایش هردو معیار را نشان می‌دهد.

ساختار پایان نامه

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

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

فصل سوم(روش‌های پیشنهادی): دراین فصل دو روش برای ترکیب اطلاعات ترجمه آماری وزبان‌شناسیارائه گردیده است.این روش‌ها بصورت اعمال پیش پردازشی برروی پیکره صورت می‌پذیردتا در مرحله ایجاد مدل‌های مورد نظر گردند.

فصل چهارم(پیاده‌سازی وارزیابی): در این قسمت ابتدا مراحل ونحوه‌ی ساخت پیکره دوزبانه مورد بررسی قرار می گیرد سپس نحوه‌ی پیاده‌سازی روش‌های ارائه شده رابیان می‌کند ودر انتها نتایج حاصل از ترجمه در قالب معیار های 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.