

شماره مستند: ۱۹۰/۲۵۳۷/۱/۱



جمهوری اسلامی ایران

دبیرخانه شورای عالی اطلاع رسانی

**جمع‌آوری اطلاعات چالش‌ها و روش‌های ترجمه ماشینی زبان انگلیسی
به فارسی و بالعکس**

نسخه ۱.۰

دانشگاه علم و صنعت ایران

فروردین ۸۸

فهرست مطالب

۴ اهمیت، چالش‌ها و روش‌ها
۴ ۱.۱ مقدمه
۴ ۱.۲ ضرورت وجود ترجمه ماشینی
۵ ۱.۳ طبقه‌بندی کاربردهای مترجم
۵ ۱.۳.۱ ترجمه ماشینی با کمک انسان
۵ ۱.۳.۲ ترجمه انسانی با کمک ماشین
۶ ۱.۳.۳ حافظه ترجمه
۶ ۱.۴ مروری بر روش‌های ترجمه
۶ ۱.۴.۱ روش‌های مبتنی بر قوانین
۸ ۱.۴.۲ روش‌های آماری
۸ ۱.۴.۳ ترجمه ماشینی مبتنی بر مثال
۸ ۱.۴.۴ مقایسه روش‌های آماری در مقابل مبتنی بر قوانین
۸ ۱.۵ کارهای انجام شده در زبان فارسی
۸ ۱.۵.۱ مترجم پارس
۹ ۱.۵.۲ مترجم LanguageWeaver
۹ ۱.۵.۳ پروژه شیراز
۹ ۱.۵.۴ مترجم شرکت سیا سافت
۹ ۱.۵.۵ مترجم شرکت LEC
۱۰ ۱.۵.۶ مترجم گوگل
۱۰ ۱.۵.۷ سایر محصولات فارسی
۱۰ ۱.۶ کارهای انجام شده در زبانهای دیگر
۱۰ ۱.۶.۱ مترجم SYSTRAN
۱۰ ۱.۶.۲ مترجم گوگل
۱۱ ۱.۷ سرویس‌های ترجمه رایگان
۱۲ ۱.۷.۱ مترجم‌های موبایل
۱۲ ۱.۸ مشکلات ترجمه فارسی به انگلیسی
۱۲ ۱.۸.۱ مشکلات ناشی از رسم‌الخط
۱۲ ۱.۸.۲ تشخیص نقش کلمات در جمله و معنی آن‌ها
۱۲ ۱.۹ سرویس‌ها و خدمات قابل ارائه در مترجم نور
۱۴ ابزارها، نیازمندی‌ها و چگونگی
۱۴ ۲.۱ انتخاب روش
۱۴ ۲.۲ دادگان و چالش‌ها
۱۴ ۲.۲.۱ تشخیص کران کلمه

۱۴.....	۲.۲.۲ تطبیق دادن کلمات.....
۱۴.....	۲.۲.۳ تشخیص کران جمله.....
۱۵.....	۲.۲.۴ تطبیق دادن جملات.....
۱۵.....	۲.۲.۵ استخراج هم‌نشینی‌ها.....
۱۵.....	۲.۳ تجزیه‌کننده.....
۱۵.....	۲.۴ ساختار ارائه.....
۱۵.....	۲.۵ مبدل.....
۱۶.....	۲.۶ منابع و مآخذ.....

اهمیت، چالش‌ها و روش‌ها

۱.۱ مقدمه

ترجمه ماشینی عبارت است از تبدیل یک متن از زبان مبدا به زبان مقصد با استفاده از یک نرم‌افزار ترجمه کننده؛ چه این عمل با کمک انسان باشد چه بدون کمک انسان.

سابقه ترجمه ماشینی به حدود نیم قرن پیش بر می‌گردد. اگر چه از آن زمان تا کنون ترجمه ماشینی رشد زیادی داشته‌است، هنوز هم نقص‌های فراوانی دارد. اصولاً چون کامپیوترها نمی‌توانند مانند انسان هوشمند باشند، ترجمه‌ای هم که توسط آنها انجام می‌شود، ترجمه کاملی نخواهد بود. نمی‌توان انتظار داشت که با استفاده از یک نرم‌افزار مترجم، هر متنی به آسانی ترجمه شود. نرم‌افزارهای مترجم، در بهترین حالت، عمل ترجمه را با دقتی در حدود ۷۰ درصد انجام می‌دهند. برای به دست آوردن نتیجه بهتر، لازم است قبل و بعد از ترجمه، مقداری ویرایش روی متن انجام شود.

در طی چند دهه اخیر و هم‌زمان با گسترش و پیشرفت زبان‌شناسی رایانه‌ای، در بسیاری از کشورهای پیشرفته، تلاش‌های همه‌جانبه و پیگیر در جهت ترجمه متون از طریق کامپیوتر انجام گرفته و حاصل کار با توجه به تنگناها و مسائل خاص ترجمه درخور توجه بوده است. این حقیقت را باید پذیرفت که ترجمه متون از طریق کامپیوتر محدودیت‌های خاص خود داشته است. در بعضی از زمینه‌ها حاصل کارکرد واقعا رضایت‌بخش است ولی در بعضی موارد نتایج به دست آمده را علی‌رغم قابل فهم بودن، باید ویراستاری کرد. طبیعتاً نوع متن و میزان پیچیدگی آن اهمیت زیادی در نتیجه کار دارد.

۱.۲ ضرورت وجود ترجمه ماشینی

امروزه با توسعه اینترنت، افراد زیادی با مشکل زبان در استفاده از آن مواجه هستند. ۸۰ درصد مطالب منتشر شده در اینترنت هنوز هم به زبان انگلیسی هستند شاید به این دلیل که بیشتر استفاده کنندگان اینترنت با این زبان آشنائی دارند. اما در بسیاری از کشورهای جهان مانند ایران، افراد زیادی هستند که به زبان انگلیسی آشنائی ندارند. برای این دسته از کاربران ابزارهای ترجمه ماشینی می‌تواند گزینه مناسبی برای گردش در اینترنت باشد. اما آیا چنین ابزارهایی به اندازه کافی دقت دارند؟ برای چه زبانهایی ابزارهای ترجمه مجانی وجود دارد؟

در ترجمه ماشینی ویژگی‌هایی وجود دارد که نه تنها از نظر جذبه و کشش علمی بلکه از دیدگاه اقتصادی و دیگر ضرورت‌ها و اقتضاهای عصر، انجام آن را کاملاً توجیه می‌کند. به عنوان مثال در مقر سازمان ناتو در بروکسلو جامعه اروپا علی‌رغم آنکه حدود ۱۲۰۰ مترجم ورزیده به کار اشتغال دارند، در حال حاضر از ترجمه ماشینی نیز استفاده می‌شود. دلیل این امر سرعت و هزینه است. میزان کاری که مترجمی ورزیده در خلال چندین روز انجام می‌دهد، توسط کامپیوتر در عرض چند دقیقه انجام می‌شود. حتی اگر کیفیت و دقت ترجمه ماشینی کمتر از حاصل کار مترجم باشد، باز هم از جهات مختلف اهمیت و ارزش خاص آن چشمگیر است.

ترجمه ماشینی یکی از ابزارهایی است که از همان ابتدا کاربرد نظامی داشته است. شروع این پروژه از وزارت دفاع آمریکا بوده است و آخرین اخبار هم نشان دهنده استفاده جدی در بخش نظامی است. جدیداً مترجم ماشینی فارسی هم به جمع پروژه‌های مورد استفاده بخش نظامی آمریکا اضافه شده است.

۱.۳ طبقه‌بندی کاربردهای مترجم

سیستم‌های ترجمه ماشینی را از نظر نقش ماشین و انسان در ترجمه، می‌توان بصورت زیر دسته بندی کرد:

Fully Automatic High Quality Translation - FAHQT	ترجمه کامل و با کیفیت
Human-Aided Machine Translation – HAMT	گروه ترجمه ماشینی با کمک انسان
Machine Aided Human Translation - MAHT	ترجمه انسانی با کمک ماشین
Traditional Human Translation	ترجمه انسانی

پیشرفته‌ترین ترجمه به وسیله ماشین، ترجمه کامل و با کیفیت بدون نیاز به انسان می‌باشد که در بالای این لیست قرار دارد. از بالا به پایین به ترتیب نقش کامپیوتر در ترجمه کم شده و به ترجمه انسانی می‌رسیم. در حال حاضر بیشتر سیستم‌های ترجمه ماشینی، در دو گروه ترجمه ماشینی با کمک انسان و ترجمه انسانی با کمک ماشین قرار می‌گیرند.

۱.۳.۱ ترجمه ماشینی با کمک انسان

در اینگونه سیستم‌های ترجمه ماشینی، یک برنامه عمل ترجمه را انجام می‌دهد و انسان در دو حالت می‌تواند به برنامه در ترجمه کمک کند. این کمک می‌تواند در حین ترجمه انجام شود که به آن روش تعاملی می‌گویند. در این روش برنامه در حین ترجمه در مواردی که به حالت‌های مبهم می‌رسد، به کاربر این امکان را می‌دهد که از بین حالت‌های ممکن بهترین را انتخاب کند. این ابهام می‌تواند در معنی یک واژه باشد یا اینکه در شناسایی گرامری کلمات در جمله باشد. اینگونه سیستم‌ها می‌توانند مورد استفاده افرادی قرار گیرند که با ترجمه آشنائی دارند. سرعت استفاده از اینگونه سیستم‌ها به دلیل نیاز به کمک انسان، بالا نیست.

در روش دیگر کمک انسان، برنامه قبل یا بعد از ترجمه نیاز به این کمک دارد. به این صورت که ابتدا کاربر بایستی بتواند متن داده شده را ساده کند. جملات بسیار طولانی را به جملات کوتاه‌تر بشکند. در نهایت متن را به برنامه مترجم بدهد و ترجمه انجام شده توسط ماشین در این مرحله بدون نیاز به راهنمایی انسان، انجام می‌گیرد. پس از اینکه ترجمه به پایان رسید، متن ترجمه شده بایستی توسط کاربر ویرایش شود. البته پس ویراستاری و پیش ویراستاری یک متن می‌تواند به کیفیت ترجمه ماشینی در هر دو روش کمک زیادی بکنند. در این زمینه در یک بخش جداگانه توصیه‌هایی برای بهبود کیفیت ترجمه ماشینی، ارائه خواهد شد.

۱.۳.۲ ترجمه انسانی با کمک ماشین

اینگونه از سیستم‌ها بیشتر مبتنی بر تجربیات انسانی بوده و ماشین در نقش یک دستیار در ترجمه کاربرد دارد. برنامه‌های ویراستار متن، واژه‌نامه‌ها و در سیستم‌های پیچیده‌تر برنامه‌های حافظه ترجمه^۱، از این گروه هستند. برنامه‌های ویراستار متن و واژه‌نامه‌ها در جامعه جایگاه خود را یافته و به عنوان ابزار ترجمه مدتهاست که مورد استفاده مترجمین حرفه‌ای و نوآموز قرار می‌گیرند. اما برنامه‌های حافظه ترجمه، در بازار ایران ناشناخته مانده است ولی در بازار ترجمه در سطح بین‌المللی کاربرد وسیعی پیدا کرده است. این برنامه‌ها با توجه به محدودیت‌هایی که برای سیستم‌های ترجمه ماشینی وجود دارد، توانسته‌اند به مترجمین حرفه‌ای کمک شایانی بکنند. روش کار این برنامه بر اساس ترجمه‌های انجام شده می‌باشد. هر متنی که برای ترجمه به برنامه‌ها داده می‌شود، ابتدا به جملات قابل ترجمه شکسته می‌شود. برنامه سعی می‌کند که جملاتی را که قبلاً ترجمه شده‌اند را پیدا کرده و در متن

ترجمه شده قرار دهد. در این میان جملاتی که برنامه برای اولین بار با آنها برخورد می‌کند را در بخش دیگری و با روش دیگری پردازش می‌کند. جملاتی که برای برنامه ناشناخته باشند نیز آنالیز شده و نزدیکترین جملات و ترجمه آن‌ها به مترجم نشان داده می‌شوند، در اینجا مترجم با استفاده از ترجمه‌های مشابه ترجمه دقیق جملات جدید را انجام داده و به برنامه یاد می‌دهد. دفعات بعد که برنامه به این جملات برخورد کند، ترجمه دقیق آنها را ارائه می‌دهد. اینگونه از برنامه بر اساس یکی از مهمترین ویژگی‌های کار ترجمه، یعنی تکراری بودن ترجمه، طراحی می‌شوند. هرچند برای زبان فارسی به صورت خاص برنامه‌ای طراحی نشده است، اما برنامه‌های مشابه برای زبان عربی وجود دارند که تا اندازه‌ای می‌توانند به مترجمین برای ترجمه متون به زبان فارسی یاری برسانند.

۱.۳.۳ حافظه ترجمه

حافظه ترجمه پایگاه داده‌ای است که قطعات یا عبارتهای از پیش ترجمه شده را در خود نگه می‌دارد تا در زمان‌های بعدی که کاربر به همان عبارتهای تکراری رسید، بتواند با استفاده از عبارتهای ذخیره شده کار کمتری انجام دهد. همزمان با عمل ترجمه توسط کاربر، سیستم حافظه ترجمه، کلمات، عبارتهای، جملات و پاراگراف‌ها را به صورت قابل بازیافتن سریع در خود نگه می‌دارد؛ به محض اینکه کاربر به قطعه‌ای مشابهی رسید، متون قبلی را به کاربر عرضه می‌کند. در واقع سیستم حافظه ترجمه یک ابزار کمکی برای کاربر انسانی است و موفقیت آن در موارد بسیاری به اثبات رسیده است. نرم‌افزارهایی که این سیستم را در خود تعبیه دارند با عنوان ترجمه‌ی مدیریت شده با حافظه^۲ شناخته می‌شوند. حافظه‌های ترجمه عموماً در کنار نرم‌افزارهایی مانند: «نرم‌افزارهای کمکی ترجمه»، «نرم‌افزارهای پردازش کلمه»، «واژه‌نامه‌های چند زبانه» و حتی «مترجم‌های ماشینی» ارایه می‌شوند. آمارها نشان داده است که بسیاری از شرکت‌هایی که با متون چندزبانه سر و کار دارند از سیستم‌های حافظه ترجمه استفاده می‌کنند.

هرچند این روزها سیستم‌های حافظه ترجمه مورد توجه مترجمین حرفه‌ای قرار گرفته اند، اما باید اشاره داشت که ایده این نوع سیستم‌ها زیاد جدید نیست. تعدادی از شرکت‌های ترجمه قبل از این بانک‌های اطلاعاتی داشتند که برای صرفه جوئی در هزینه و زمان ترجمه از آنها استفاده می‌کردند. اما معرفی این سیستم‌ها بصورت تجاری و در بازار نرم‌افزار باعث استفاده گسترده آن در بین شرکت‌ها و افراد مختلف در زمینه ترجمه شده است.

هر کار ترجمه‌ای که سه ویژگی مهم زیر را داشته باشد را می‌توان با یک سیستم حافظه ترجمه با هزینه کمتر انجام داد: مطالب تکراری، استفاده از فایل‌های الکترونیکی، حجم قابل توجه متن برای ترجمه. بیشتر افرادی که شرایط مذکور را داشته‌اند و از چنین سیستم‌هایی استفاده کرده‌اند، ۲۵ الی ۳۰ درصد در هزینه‌های ترجمه صرفه جوئی کرده اند. هر چه حجم بانک اطلاعاتی سیستم افزایش پیدا می‌کند یا به عبارت دیگر هر چه با این سیستم‌ها ترجمه بیشتری انجام می‌شود، مقدار صرفه جوئی در هزینه ترجمه افزایش پیدا می‌کند.

۱.۴ مروری بر روش‌های ترجمه

۱.۴.۱ روش‌های مبتنی بر قوانین

در ترجمه ماشینی می‌توان از روشی مبتنی بر قوانین^۳ زبان استفاده کرد که در آن، کلمات در از طریق قوانین زبانی ترجمه می‌شوند؛ یعنی مناسب‌ترین کلمات در زبان مقصد، جایگزین کلمات متناظرشان در زبان مبدا می‌شوند. اغلب بحث می‌شود که ترجمه ماشینی نیازمند این است که ابتدا مساله فهم زبان طبیعی حل شود.

^۲ translation memory manager

^۳ Role-Based

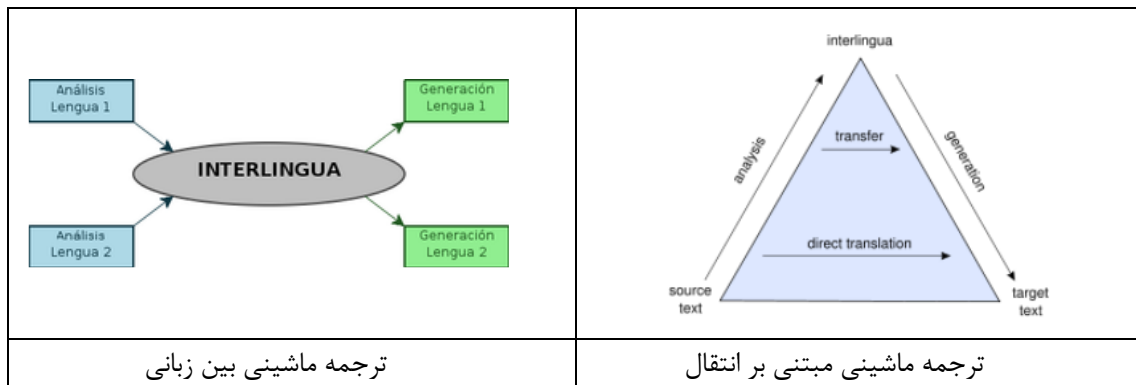
به طور کلی، روش‌های مبتنی بر قوانین، ابتدا متن را تجزیه می‌کنند، و سپس معمولاً یک نمایش نمادین واسطه‌ای (زبان میانی) برای آن بر می‌گزینند. متن زبان مقصد از روی این نمایش نمادین تولید می‌شود. با توجه به ماهیت این نمایش میانی، عناوین متفاوتی برای روش‌های متناظر برگزیده می‌شود:

- ترجمه ماشینی بین زبانی^۴
- ترجمه ماشینی مبتنی بر انتقال^۵

این روش‌ها نیازمند واژه‌نامه‌های بسیار وسیع به همراه اطلاعات ریخت‌شناسی، نحوی و معنایی می‌باشند. همچنین، مجموعه‌ی بزرگی از قوانین تعریف‌شده نیاز دارند. وجود شبکه معنایی کلمات^۶ کمک شایانی به کیفیت ترجمه ارائه شده خواهد کرد.

هر دو روش فوق ایده یکسانی را دنبال می‌کنند: لازم است تا یک «نمایش میانی» برای بدست آوردن معنی زبان مبدا داشته باشیم و از روی آن ترجمه درست را تولید کنیم. تفاوت دو روش فوق از این زبان میانی نشات می‌گیرد. در روش بین زبانی، این «نمایش میانی» از هر دو زبان مبدا و مقصد مستقل است؛ در حالی که در روش مبتنی بر انتقال، این نمایش میانی، تا حدی وابستگی به زبان‌های مبدا و مقصد آن دارد.

در یک سیستم ترجمه مبتنی بر انتقال، ابتدا متن ورودی از نظر ریخت‌شناسی و نحوی تجزیه و تحلیل می‌شود تا یک نمایش معنایی برای آن بدست آید. این نمایش معنایی سپس می‌تواند به یک سطح بالاتر تجزیه، پالایش شود. در این پالایش بر بخش‌های مربوط به ترجمه تاکید می‌شود و سایر اطلاعات زاید نادیده گرفته می‌شود. در فرایند انتقال، این نمایش سطح بالا به نمایشی در همان سطح اما در زبان مبدا تبدیل می‌شود. این دو نمایش با عنوان نمایش‌های میانی شناخته می‌شوند. برای تولید زبان خروجی فرایند نمایش به صورت معکوس انجام می‌شود. در روش ترجمه بین زبانی، زبان مبدا به یک زبان مصنوعی عمومی^۷ تبدیل می‌شود. در واقع این زبان یک نمایش مجرد سطح بالا است که از روی آن می‌توان، متن را به هر زبان دیگر ترجمه کرد.



روش‌های ترجمه ضعیف‌تری نیز وجود دارند که در دسته روش‌های مبتنی بر قوانین قرار می‌گیرند:

- ترجمه بر مبنای دیکشنری^۸: در این روش از مدخل‌های دیکشنری برای ترجمه استفاده می‌شود؛ یعنی کلمات همانگونه که هستند، معنی آنها از دیکشنری استخراج می‌شود و به جای آنها قرار می‌گیرد.

^۴ Inter-Lingual

^۵ Transfer-Based

^۶ Semantic Dictionary مانند WordNet

^۷ Interlingua

^۸ Dictionary-Based

- ترجمه انتقالی کم‌عمق^۹: برای زبان‌هایی که ساختار نزدیکی به همدیگر دارند از این روش می‌توان استفاده کرد. از جمله می‌توان از نرم‌افزار منبع باز Apertium نام برد [4].

۱.۴.۲ روش‌های آماری

در روش‌های آماری^{۱۰} سعی می‌شود تا ترجمه مورد نظر با استفاده از روش‌های آماری بر اساس متون دو زبانه موجود به دست آید. وقتی مقدار زیادی متون دوزبانه در دسترس باشند، به نتایج بسیار شگفت‌انگیزی در ترجمه می‌توان دست یافت، ولی متاسفانه حجم این متون هنوز بسیار اندک است. در این روش‌ها با توجه به هم‌نشینی کلمات در متون مبدا و مقصد، کلمات مناسب را در ترجمه انتخاب می‌کند.

۱.۴.۳ ترجمه ماشینی مبتنی بر مثال

ویژگی اصلی ترجمه ماشینی مبتنی بر مثال^{۱۱}، استفاده از متون دوزبانه شامل متون موازی، به عنوان پایگاه دانش آن در زمان ترجمه است. در شالوده ترجمه ماشینی، ایده ترجمه بر مبنای قیاس قرار دارد. این همان کاری است که توسط مترجم انسانی انجام می‌شود. یعنی بر اساس استدلال موردی از تجربیات قبلی شکل می‌گیرد. در این سیستم به جای تجزیه و تحلیل عمیق جملات توسط مترجم انسانی، جملات ابتدا به عبارت‌های کوچکتر شکسته می‌شوند، سپس این عبارت‌ها بر اساس آموزش‌ها از مثال‌های قبلی ترجمه می‌شوند و نهایتاً این عبارت‌ها در غالب یک جمله‌ی طولانی به هم می‌پیوندند. ترجمه عبارت‌ها در مقایسه با ترجمه‌های قبلی بدست می‌آید. قوانین ترجمه ماشینی از روی مثال‌هایی بدست می‌آید که برای آموزش ماشین تهیه می‌شوند.

۱.۴.۴ مقایسه روش‌های آماری در مقابل مبتنی بر قوانین

اشکال عمده روش‌های ترجمه آماری در این است که فراهم کردن متون دوزبانه برای آموزش آن بسیار دور از دسترس است و فقط شرکت‌ها و موسسات خاصی به متونی از این دست دسترسی دارند. اما دقت این روش‌ها نسبتاً خوب است و ترجمه قابل قبولی ارائه می‌دهد.

در روش‌های مبتنی بر قوانین اشکال عمده در این است که نیازمند افراد بسیار متخصص و خبره در زبان‌های مبدا و مقصد هستیم که ساختارهای زبان را به خوبی بشناسند تا بتوان این ساختارها را به نرم‌افزار آموزش داد.

۱.۵ کارهای انجام شده در زبان فارسی

۱.۵.۱ مترجم پارس

نرم افزار مترجم پارس یک سیستم ترجمه ماشینی است که متن انگلیسی را به جمله های فارسی ترجمه می نماید اولین نگارش تجاری مترجم در شهریور ماه ۱۳۷۶ به عموم عرضه گردید و آخرین نگارش ارتقاء یافته مترجم پارس در خرداد ماه ۱۳۸۱ ارائه شده است در سیستم ورودی مترجم پارس، تایپ متن انگلیسی یا انتخاب محتویات از یک فایل استفاده می گردد در حال حاضر، موتور مترجم پارس قادر به شناسایی و تجزیه بیش از یک میلیون و پانصد هزار واژه عمومی و اصطلاحات تخصصی متداول در ۳۳ رشته علمی است.

نمونه‌ای از ترجمه نرم‌افزار «مترجم پارس»:

Shallow Transfer^۹

Statistical^{۱۰}

Example-Based^{۱۱}

موضوع: سیاست
India has complained in the past that attacks on its soil have been carried out by groups based in Pakistan, although relations between the two countries have improved in recent years and Pakistani leaders were swift to condemn the latest attacks. ¹²
هند در گذشته شکایت کرده است که حمله می کند خاکش بوده است خارج بوسیله مبنی بر گروه ها در پاکستان، ادامه بدهد اگرچه رابطه مابین دو کشورها در سالی اخیر اصلاح کنند و رهبرهای پاکستانی تندرو بودند که آخرین حملات مورد اعتراض قرار بدهند.

۱.۵.۲ مترجم LanguageWeaver

یک شرکت آمریکائی با همکاری دانشگاه کالیفرنیا جنوبی جدیداً زبان فارسی را به مجموعه زبانهای مترجم ماشینی خودش اضافه کرده است. این شرکت که بیشتر اعضای مدیریت آن از اساتید دانشگاه کالیفرنیا جنوبی است توانسته با استفاده از کارهای تحقیقاتی انجام شده در دانشگاه، سیستم ترجمه ماشینی مناسبی را ارائه نماید. نام این شرکت LanguageWeaver است.

البته یکی از دلایل موفقیت چنین پروژه هائی بیشتر به حجم زیاد کاریست که در دانشگاه بر آنها انجام شده است. این شرکت به دلیل پیشرفتی که در سه سال گذشته داشته موفق به دریافت جوایز زیادی شده است. بد نیست نگاهی به بودجه های چند ملیون دلاری این شرکت بیندازیم که بیشتر آنها توسط DARPA^{۱۳} و موسسه ملی استاندارد آمریکا به این شرکت داده شده است.

۱.۵.۳ پروژه شیراز

تنها کار جدی که قبل از پروژه LanguageWeaver در زمینه ترجمه ماشینی انجام شده بود، توسط دانشگاه نیومکزیکو پشتیبانی میشد که در سالهای ۱۹۹۱ دلیل مسائل مالی و عدم تخصیص وامهای مناسب متوقف ماند. نام این پروژه شیراز^{۱۴} بود. این پروژه از محدود پروژه هائی است که مستندات مناسبی در این زمینه ارائه داده است.

۱.۵.۴ مترجم شرکت سیا سافت

یکی دیگر از شرکت‌های آمریکائی که در زمینه ترجمه ماشینی برای زبان فارسی فعالیت می کنند شرکت سیا سافت می باشد. این شرکت در سالهای گذشته بر روی زبان‌های فارسی، عربی، فارسی دری و پشتو فعالیت‌های زیادی انجام داده است و در حال حاضر با حضور در بازار کشور افغانستان توانسته به عنوان یکی از سیستم‌های ترجمه فعال در زمینه زبان‌های این کشور فعالیت نماید.

۱.۵.۵ مترجم شرکت LEC

شرکت Language Engineering Co (LLC) یا LEC تعدادی سرویس برای ترجمه ارائه می کند که زبان فارسی نیز جزو آنها است. نام محصولات این شرکت را در زیر می بینید:

Product: LEC Translate DotNet
Source: <http://www.lec.com/w2/translate-dotnet.asp>

Product: LEC Passport Premium
Source: <http://www.lec.com/w2/passport-premium.asp>

Product: LEC Passport

^{۱۲} این متن از سایت خبری BBC برداشته شده است.

^{۱۳} Defense Advanced Research Projects Agency

^{۱۴} Shiraz Machine Translation

Access: <http://www.lec.com/w2/passport.asp>

Product: LEC Translate

Source: <http://www.lec.com/w2/listProductFamily.asp>

۱.۵.۶ مترجم گوگل

شرکت گوگل در سال‌های گذشته برای مدتی امکان ترجمه انگلیسی-فارسی را در سایت خود ارائه کرد، اما پس از مدتی آن را برداشت.

۱.۵.۷ سایر محصولات فارسی

برخی از شرکت‌های دیگر نیز پشتیبانی از زبان فارسی را در محصولات خود ادعا می‌کنند. لیست این شرکت‌ها را در زیر می‌آوریم.

Product: 1-800-Translate

Company: 1-800-Translate

Site: <http://www.1-800-translate.com>

Product: Postchi

Site: <http://www.postchi.com/main/LowK/index.php>

Product: TranSphere

Company: AppTek Inc.

Site: <http://www.apptek.com/products/transphere.html>

Product: WebTrans

Company: AppTek Inc

Source: [http://www.apptek.com/products/pdf/WebTrans\(sellsheet\).pdf](http://www.apptek.com/products/pdf/WebTrans(sellsheet).pdf)

۱.۶ کارهای انجام شده در زبانهای دیگر

۱.۶.۱ مترجم SYSTRAN

یکی از قدیمی‌ترین پروژه‌های ترجمه ماشینی است که بیش از ۳۰ سال سابقه دارد. این برنامه توسط یکی از اولین طراحان سیستم‌های ترجمه ماشینی برای نیروی هوایی آمریکا طراحی شد و در سال ۱۹۶۹ تست شد. این برنامه در ابتدا برای زبانهای روسی و انگلیسی طراحی و عرضه شد. پس از تست موفقیت‌آمیز اولیه، وزارت خارجه و همچنین سازمان فضائی آمریکا نیز متقاضی استفاده از این برنامه در ترجمه مستندات خود شدند. در سال ۱۹۷۵ نگارشی از این برنامه برای زوج زبان انگلیسی و فرانسه نیز به سفارش اتحادیه اروپا تهیه شد. در حال حاضر این برنامه بیش از ۲۰ زبان را پوشش می‌دهد. این برنامه در سایت رسمی خود سرویس ترجمه بر خط را به صورت رایگان ارائه می‌کند. تنوع محصولات ارائه شده توسط شرکت نرم افزاری SysTran نیز یکی از ویژگی‌های این گروه می‌باشد.

۱.۶.۲ مترجم گوگل

سیستم ترجمه ماشینی گوگل از تمام سیستم‌های ترجمه ماشینی موجود بهتر عمل می‌کند. بر اساس آزمایش‌های انجام شده و مقایسه بین نرم افزارهای موجود از شرکت‌های آی.بی.ام.^{۱۵} و دانشگاه‌های آمریکا در زمینه ترجمه عربی به انگلیسی و چینی به انگلیسی، سیستم ترجمه ماشینی گوگل توانسته است بالاترین رتبه را به خود

اختصاص دهد. این آزمایش بر روی ۱۰۰ مقاله انجام شده است. علت بالاتر بودن کیفیت ترجمه سیستم گوگل از نظر کارشناسان ناشی از دسترسی این غول اینترنتی به حجم بالای متون در اینترنت است. در روش‌های جدید ترجمه ماشینی که بیشتر بر اساس روش‌های آماری است، دسترسی به حجم بالای متون می‌تواند کمک بسیاری به بالا بردن دقت و کیفیت ترجمه بنماید.

موسسه NIST هر ساله سیستم‌های ترجمه ماشینی را بر اساس روش‌های استاندارد، از نظر کیفیت نتایج مورد ارزیابی قرار می‌دهد. طبق این ارزیابی مترجم گوگل در زمینه ترجمه زبان عربی بالاترین رتبه را داشته است. برای مشاهده همه نتایج به آدرس زیر^{۱۶} مراجعه نمائید. شرکت گوگل توانسته است مترجم ماشینی برای زبان انگلیسی چینی و عربی را توسط تیمی تولید کند که هیچکدام از آنها با زبان‌های عربی و چینی آشنائی ندارند.

در سیستم‌های ترجمه ماشینی که قبلاً طراحی شده‌اند همیشه تعدادی کارشناس زبانشناسی در کنار تیم‌ها طراحی نرم‌افزار وجود دارند که تسلط کامل به زبان مقصد یا مبدا دارند. در این سیستم‌ها معمولاً اطلاعاتی زبان‌شناختی نقش اصلی در اطلاعات ورودی به سیستم را بازی می‌کند. اما یکی از روش‌هایی که امروزه مورد استفاده شرکت‌های مختلف قرار گرفته استفاده از متون از قبل ترجمه شده و مدل‌های آماری برای ترجمه ماشینی است. اولین کار در این زمینه در شرکت آی.بی.ام انجام شده است. اما در حال حاضر گوگل به دلیل دسترسی به حجم زیادی از متون زبان طبیعی توانسته مجموعه‌ای از بهترین سیستم‌های ترجمه ماشینی را با استفاده از روش آماری ارائه دهد.

البته این شرکت هنوز برای زبان فارسی به طور رسمی مترجم ارائه نداده است. اما می‌توان انتظار داشت که با سرمایه‌گذاری وزارت دفاع آمریکا این سیستم ترجمه انگلیسی-فارسی هم بزودی ارائه شود. واقعیت این است که شرکت‌هایی مانند گوگل به دلیل دسترسی به حجم بسیار بالایی از اطلاعات، توانائی انجام بسیاری از پروژه‌هایی را دارند که دیگران فقط به دلیل عدم دسترسی به این اطلاعات (بدلیل هزینه زیاد گردآوری آن) نمی‌توانند انجام دهند. اما مهم این است که امکان دسترسی به حجم اطلاعات طبقه‌بندی شده اینترنت در اختیار پروژه‌های انجام نشده‌ای چون ترجمه ماشینی قرار گرفته و در نهایت به نفع کاربران اینترنت خواهد شد. در ایران هم شرکت‌های کامپیوتری با استفاده از روش‌های آماری در زمینه ترجمه ماشینی کارهایی انجام داده‌اند اما به دلیل هزینه زیاد تهیه متون ترجمه شده، هزینه پروژه بالا رفته و روز به روز انجام کار برای آنها سخت‌تر می‌شود.

۱.۷ سرویس‌های ترجمه رایگان

در حال حاضر تعداد شرکت‌هایی که سرویس‌های ترجمه مجانی را ارائه می‌کنند زیاد شده است و حتی شرکتی مانند گوگل هم چنین سرویسی را ارائه کرده است. اما برای زبان فارسی در حال حاضر سرویس‌های زیادی وجود ندارد. یکی از این سرویس‌ها متعلق به مترجم پارس است و دیگری هم به صورت محدود با نرم‌افزار بابیلون^{۱۷} ارائه شده است که البته متعلق به شرکت LEC است. استفاده از سرویس اول رایگان می‌باشد اما سرویس دوم از طریق نرم‌افزار بابیلون قابل استفاده است که برای سرویس ترجمه مجانی بایستی حق عضویت پرداخته شود.

در مورد کیفیت ترجمه مجانی که با کمک ماشین انجام می‌شود بایستی دقت شود. هنوز خطای موجود در ترجمه مجانی ماشینی زیاد است. استفاده از ترجمه مجانی ماشینی محدود به درک مطلب می‌باشد هرچند در این مورد نیز خطا دارد اما با توجه به رایگان بودن آن، مناسب بنظر می‌رسد. شرکت‌های ایرانی و خارجی جدیدی نیز بر روی ترجمه ماشینی کار می‌کنند که در حال حاضر سرویس‌های خود را بصورت رایگان ارائه نمی‌کنند.

^{۱۶} http://www.google.com/language_tools?hl=en

^{۱۷} Babylon

گروه مترجم پارس با همکاری یک شرکت ارائه دهنده سرویس پیام کوتاه، سرویس رایگان ترجمه متن و کلمه فارسی را برای کاربران ایرانی موبایل ارائه داد. این سرویس برای موبایل‌هایی که زبان فارسی را پشتیبانی می‌کنند قابل استفاده می‌باشد.

۱.۷.۱ مترجم‌های موبایل

مدتی است که سرویس‌های ترجمه ماشینی به گوشی‌های موبایل هم رسیده است. البته با توجه به مصرف بالای حافظه و توان پردازشی در سیستم‌های ترجمه ماشینی، هنوز سیستم‌های ترجمه ماشینی قادر به اجرا بر روی موبایل نیستند. اما استفاده از خدمات ترجمه ماشینی و ارائه آن توسط سرویس پیام کوتاه امکان‌پذیر است. در کشور ما هم یک شرکت در حال حاضر سرویس ترجمه متن انگلیسی به فارسی و همچنین ارائه سرویس دیکشنری را آغاز کرده است.

در حال حاضر برای استفاده از این سیستم بایستی کارت‌های آن شرکت را خریده و فعال نمائید. به نظر می‌رسد استفاده از سرویس‌های پیام کوتاه به خصوص در زمینه ترجمه کمک زیادی به دانشجویان و مسافری در کشورهای خارجی می‌کند. دیگر نیاز نیست که کتاب مکالمات روزمره انگلیسی را با خودمان به مسافرت ببریم. به همراه داشتن موبایل (که در بیشتر کشورهای اروپائی جواب می‌دهد) برای امر ترجمه کافی است.

۱.۸ مشکلات ترجمه فارسی به انگلیسی

۱.۸.۱ مشکلات ناشی از رسم‌الخط

یکی از مشکلاتی که در زبان فارسی و ابزارهای پردازش آن وجود دارد، خط زبان فارسی است. می‌توان مشکلات ناشی از رسم‌الخط را در موارد زیر بیان کرد:

- مشکلات ناشی از ریخت‌شناسی^{۱۸} بسیار پیچیده‌ی فارسی
- فراوانی و پیچیدگی وندها^{۱۹} و ترکیب‌های مختلف آن‌ها (بیش از ۲۵۰ ترکیب مختلف)
- پیچیدگی و وجود موارد استثنای زیار در قواعد صرف فعلی^{۲۰}
- وجود اعراب
- وجود حروف هم‌آوای^{۲۱} فراوان (مانند: «ظ» و «ز»)
- وجود حروف هم‌شکل^{۲۲} فراوان (مانند: «چ»، «ج»، «خ» و «ح»)
- وجود انواع مختلف از یک حرف (مثلا وجود ۴ کاراکتر با کدهای مختلف برای «ی» با شکل مشابه)
- وجود نگارش‌های مختلف از یک کلمه به خاطر اختلاف سلیقه در گذاشتن یا نگذاشتن فاصله یا نیم‌فاصله
- مشخص نبودن مرز یک کلمه در اثر فاصله‌گذاری

۱.۸.۲ تشخیص نقش کلمات در جمله و معنی آن‌ها

مهمترین بخش، مشخص کردن نقش^{۲۳} و معنای کلمات می‌باشد نقش کلمات با توجه به محل قرار گرفتن آن در جمله مشخص می‌گردد، اما مسئله مهم معنی کلمات می‌باشد؛ چرا که تعدادی از کلمات دارای چند معنی می‌باشد

^{۱۸} Morphology

^{۱۹} Affix

^{۲۰} Stemming

^{۲۱} Homophone

^{۲۲} Close-shaped - Homomorph

که این ابهام باید به روش مناسبی رفع گردد. در ترجمه ماشینی ابهام هم در معنای کلمات زبان مبدا وجود دارد (از نظر درک معنی آن) و هم در زبان مقصد و همین باعث می‌گردد روند ترجمه دچار مشکل شود.

یکی از اولین مشکلاتی که هر سیستم پردازش زبان طبیعی با آن درگیر است، مسئله ابهام معنایی و ساختاری کلمات است. بخش عمده‌ای از این ابهام به کمک روالی به نام نشانه‌گذار بخش‌کننده یا برچسب‌گذار^{۲۴} جملات که برای تعیین نقش کلمات در جمله به کار می‌رود، مرتفع می‌گردد. روال مذکور با تعیین نقش کلمه، ابهام ساختاری آن را مرتفع و در پیدا کردن معنای صحیح کلمات کمک شایانی می‌نماید. البته ابهام معنایی کلمات خیلی پیچیده‌تر از ابهام ساختاری آن است و برای رفع آن به روش‌های پیچیده‌تر نیاز است. ولی روش نشانه‌گذار بخش‌کننده جملات در پیدا کردن معنای دقیق کمک می‌کند و مراحل اولیه آن را انجام می‌دهد. در واقع، رفع ابهام از معنای یک کلمه بر می‌گردد به اینکه آن کلمه در چه جمله‌ای به کار رفته و با چه کلماتی هم‌نشین گردیده است؛ چرا که یک کلمه در یک جمله یک معنی می‌دهد، در حالی که همان کلمه در جمله دیگر معنای دیگری به خود می‌گیرد.

رفع ابهام از معنای کلمات جملات یک زبان، از سه راه کلی می‌تواند انجام پذیرد که هر راه خود روش‌های متعددی دارد: اول روش ابهام با نظارت^{۲۵} است که مبنای آن مجموعه آموزشی برچسب‌دار است. روش دوم رفع ابهام بر مبنای منابع لغوی مانند فرهنگ لغت و یا فرهنگ فرهنگ جامع^{۲۶} است. و بالاخره روش سوم رفع ابهام بدون نظارت است که در این حالت تنها مجموعه لغات و متون بدون برچسب در دسترس است [۲].

برای رفع ابهام از نقش و معنای کلمات نیازمند یک واژه‌نامه با ساختار قابل استفاده برای کامپیوتر می‌باشیم که در آن اطلاعات لازم برای هر کلمه و ارتباطات کلمات با همدیگر به صورت مناسب وجود داشته باشند. متأسفانه هنوز این مجموعه به صورت یک فرهنگ فارسی قابل خواندن برای ماشین تهیه و تدوین نگردیده است. یعنی این اطلاعات به صورت دستی وجود دارند، اما نسخه رایانه‌ای از آن‌ها با یک استاندارد معین برای تعریف هر کلمه و معنی و نقش‌های مختلف آن وجود ندارد.

Part of speech^{۲۳}Tagger^{۲۴}Supervised^{۲۵}Thesaurus^{۲۶}

ابزارها، نیازمندی‌ها و چگونگی

۲.۱ انتخاب روش

جهت انتخاب روش ترجمه‌ی مورد استفاده، با توجه به مقایسه‌ی انجام شده میان روش‌های آماری و مبتنی بر قوانین، روش‌های ترجمه آماری دقت بسیار خوبا ارائه می‌کنند اما نیازمند فراهم کردن متون دوزبانه (موازی) زیادی هستند که کار بسیار سخت و زمان‌بری است.

در روش‌های مبتنی بر قوانین اشکال عمده در این است که نیازمند افراد بسیار متخصص و خبره در زبان‌های مبدا و مقصد هستیم که ساختارهای زبان را به خوبی بشناسند تا بتوان این ساختارها را به نرم‌افزار آموزش داد. و نیز ساختارهای زبانی رایج که در مجموعه‌ی قوانین نمی‌گنجد (موارد استثنا مه در فارسی بسیار زیاد است مانند کنایه و ضرب المثل) با این روش قابل پیاده‌سازی نیستند.

بنا به دلایل ذکر شده، یک روش ترکیبی از روش‌های آماری و مبتنی بر قوانین می‌تواند نتیجه‌ای مطلوب‌تر ارائه دهد.

۲.۲ دادگان و چالش‌ها

همانگونه که ذکر شد، فراهم کردن حجم عظیمی از دادگان موازی مناسب با پوشش کافی، برای زبان‌های مبدا و مقصد امری دشوار، زمان‌بر و پرهزینه است. لذا توصیه می‌شود برای شروع کار، در حوزه‌ای خاص (مثلا جهانگردی و توریسم) به جمع‌آوری دادگان موازی اقدام گردد تا بر اساس این دادگان فرآیند ترجمه ماشینی طراحی و تولید شود. سپس با گذشت زمان می‌توان این سیستم را با دادگان دیگر غنی‌تر ساخت و بر کیفیت ترجمه‌ی آن افزود. چالش‌های موجود در این بخش شامل:

۲.۲.۱ تشخیص کران کلمه

تشخیص کران کلمه شامل تشخیص مرز یک کلمه در زبان مبدا و مقصد است. به عنوان مثال کلمه‌ی «آب‌سردکن» می‌باید به عنوان یک کلمه در نظر گرفته شود نه ۳ کلمه.

۲.۲.۲ تطبیق دادن کلمات

تطبیق دادن^{۳۷} کلمات از زبان مبدا به کلمه‌ی متناظر در زبان مقصد. توجه به این نکته ضروری است که یک کلمه از زبان مبدا می‌تواند به بیش از یک کلمه در زبان مقصد ترجمه شده باشد و بر عکس.

۲.۲.۳ تشخیص کران جمله

تشخیص کران جمله شامل تشخیص حدود جمله در زبان مبدا و مقصد است. توجه به این نکته ضروری است که ممکن است علائم نگارشی (نقطه، علامت تعجب و ...) در زبان مبدا یا مقصد برای برخی جملات وجود نداشته باشد و لذا تکیه بر این علائم جهت تشخیص کران جمله کافی نیست.

۲.۲.۴ تطبیق دادن جملات

تطبیق دادن جملات از زبان مبدا به کلمه‌ی متناظر در زبان مقصد. توجه به این نکته ضروری است که یک جمله از زبان مبدا می‌تواند به بیش از یک جمله در زبان مقصد ترجمه شده باشد و بر عکس.

۲.۲.۵ استخراج هم‌نشینی‌ها

استخراج هم‌نشینی‌های (n-gram)^{۲۸} مناسب از کلمات در جملات متناظر. در این بخش انتخاب وزن مناسب هم‌نشینی اهمیت بسزایی دارد زیرا وزن کم، ممکن است مفهوم مورد نظر را در بر نگیرد و وزن زیاد هم بار زیادی بر سیستم تحمیل کند و هم تفکیک مفاهیم را شامل نشود.

۲.۳ تجزیه‌کننده

یکی از ابزارهای ضروری و زیرساختی در نرم‌افزارهای پردازش زبان طبیعی، خصوصاً مترجم ماشینی، تجزیه‌کننده^{۲۹} است. تجزیه‌کننده می‌تواند موارد زیر را از جملات نتیجه دهد:

- وابستگی کلمات
- نقش کلمات
- برجسب‌گذاری

ایجاد یک تجزیه‌کننده‌ی کامل با دقت و کیفیت مناسب امری دشوار و نیازمند زمان است، توصیه می‌شود که در ابتدا با تجزیه‌کننده‌ای ساده‌تر و با قابلیت‌های کمتر فرایند ترجمه آغاز شود و سپس با تکمیل و بهبود تجزیه‌کننده در فازهای بعد و با موازی با پیشبرد فرآیند تولید مترجم، بر کیفیت ترجمه افزود.

۲.۴ ساختار ارائه

این بخش وظیفه‌ی نگهداری و مدیریت دادگان، هم‌نشینی‌ها، موارد حاصل از تجزیه و ... را دارا است. با توجه به حجم بسیار زیاد اطلاعات، جستجو و بازیابی اطلاعات امری زمان‌بر خواهد شد، بنابر این این بخش می‌باید دارای ساختاری مناسب جهت بازیابی کارا و سریع را ارائه نماید. مثلاً می‌توان از یک درخت نخ‌کشی شده دارای نمایه‌ی معکوس به عنوان ساختار نگهداری و ارائه استفاده نمود.

۲.۵ مبدل

این بخش که مهم‌ترین بخش از فرایند ترجمه است، تبدیل از زبان مبدا به مقصد را به عهده خواهد داشت. یک هم‌نشینی مستخرج از یک جمله که با ساختاری مناسب ارائه شده، با استفاده از تابع تبدیل که می‌تواند یافتن هم‌نشینی متناظر از زبان مقصد با اعمال قوانین زبان مقصد و مبدا باشد، این ساختار شامل هم‌نشینی را به

^{۲۸} N-gram یک هم‌نشینی از n کلمه‌ی مجاور است که n وزن این هم‌نشینی است.
^{۲۹} Parser

ساختاری از زبان مقصد شامل هم‌نشینی معادل تبدیل می‌سازد و فرآیندها این‌بار به صورت معکوس طی می‌شوند تا این ساختار هم‌نشینی و باقی هم‌نشینی‌های مرتبط، یک جمله را شکل دهند.

۲.۶ منابع و مآخذ

- [۱] وبلاگ <http://www.parstranlator.net/mt-news> ، علیرضا عبادت
- [۲] ترجمه ماشینی، امیر شهاب شهابی، دکتر عبدالحسین صراف زاده، مجله زبانشناسی
- [۳] درآمدی بر ترجمه ماشینی، پرستو کریمی ویراستار، طیبه موسوی میانگه تدوین
- [۴] www.apertium.org