

ترجمه ماشینی متون انگلیسی به زبان فارسی

مقاله‌ای که در زیر از نظر خوانندگان می‌گذرد، در اصل نامه‌ای است که آقای هوشنگ مهر جردیان از آلمان خطاب به دکتر مشکوة الدینی، استاد زبانشناسی دانشگاه فردوسی نوشته‌اند. آقای مهر جردیان (فارغ‌التحصیل دوره فوق لیسانس علوم کامپیوتر از دانشگاه صنعتی شریف) در حال حاضر در مرحله پایانی نالیف رساله دکترای خود با عنوان «ترجمه ماشینی متون علمی انگلیسی به زبان فارسی» در دانشکده انفورماتیک دانشگاه بُن می‌باشند. از آنجا که این نامه حاوی دعوتی عام برای همکاری خطاب به مؤسسات زیربسط و افراد علاقه‌مند در جهت اتمام مراحل بعدی طرح فوق می‌باشد، عیناً به چاپ آن اقدام کردیم. از همکار گرامی دکتر مشکوة الدینی که این نامه را در اختیار مجله گذاشتند و نیز توضیحات یادداشت را اضافه کردند، تشکر می‌کنیم. خوانندگان یا مؤسسات علاقه‌مند به همکاری با این طرح می‌توانند برای گرفتن آدرس نویسنده با ما مکاتبه کنند.

موضوع رساله دکترای اینجانب «ترجمه ماشینی متون علمی انگلیسی بزبان فارسی» میباشد. بدیهی است چنین کاری در عمل بسیار حجیم و سنگین است و از عهده یک نفر بر نمی‌آید و در هر مرحله قابلیت تکمیل و به‌سازی دارد.

سیستم طراحی شده اصولاً بر سه پایه استوار است:

- ۱ - تجزیه و تحلیل متن انگلیسی و تبدیل متن به یک نمایش میانی
 - ۲ - تبدیل نمایش میانی متن انگلیسی به نمایش مشابهش در زبان فارسی بکمک واژه‌نامه‌های دو زبانه
 - ۳ - تولید جملات فارسی بر مبنای نمایش میانی فوق‌الذکر و واژه‌نامه‌های یک زبانه فارسی
- مطلب قابل توجه دیگر اینکه خوشبختانه بزرگترین پروژه تحقیقاتی جامعه مشترک اروپا در زمینه ترجمه ماشینی، تحت عنوان Eurotra، در بدو شروع مطالعات تحقیقاتی اینجانب اعلام آمادگی نمود که بخش نخست فوق‌الاشاره را در اختیار اینجانب قرار دهند. با تحقق این امر لازم بود بناچار مراحل بعدی طرح ترجمه ماشینی مورد نظر اینجانب نیز تا حدودی خود را با سیستم Eurotra وفق دهند. با توجه باینکه کلیه زبانهای جامعه مشترک اروپا (انگلیسی، آلمانی، فرانسوی، اسپانیولی، هلندی، دانمارکی، پرتغالی، یونانی) توسط سیستم مزبور مورد پردازش قرار می‌گیرند، تئوریهای طرح شده برای انجام هر یک از مراحل سه گانه فوق نیز باید قابلیت این همه گوناگونی را داشته باشند. بهر حال زبان فارسی نیز توانست

به رغم مشکلات و خصوصیات بارزش، خود را با این تئوریه‌ها همگام سازد و طراحی مراحل ۲ و ۳ اشاره شده در فوق طول مدت چهار سال گذشته تحقق یافتند. اینک سیستم مزبور قادر است، کلیه جملات انگلیسی را که از نظر گرامری صحیح بوده و بسیار پیچیده نباشند بزبان فارسی ترجمه نماید. طبیعتاً تحقق این امر مستلزم آنستکه واژه‌های بکار رفته در متن مورد نظر پیشاپیش در اختیار سیستم مترجم قرار گرفته باشد. نظر باینکه مرحله سوم ترجمه، یعنی تولید جملات فارسی بر مبنای نمایش میانی، کاملاً مستقل از زبان مرجع صورت میگیرد، عملاً ترجمه از هر یک از زبانهای اشاره شده در فوق بزبان فارسی نیز با توجه به در دسترس بودن مرحله تجزیه و تحلیل متن، در مقایسه با کل مطلب کاری نسبتاً ساده خواهد بود. بدین منظور لازم است تنها یک واژه‌نامه دو زبانه از هر یک از زبانهای فوق‌الذکر به فارسی تدوین گردیده و با بهره‌گیری احتمالی از چند قاعده بازنویسی، نمایش میانی متن از زبان مرجع را به نمایش میانی مربوطه‌اش در زبان فارسی تبدیل نموده و تولید جمله فارسی متناظر را بعهده بخش سوم سیستم طراحی شده فعلی محلول نمود. در حال حاضر حدود ۲۰۰ واژه گوناگون زبان فارسی، مشابه روش اشاره شده توسط جنابعالی در کتاب «دستور زبان فارسی بر پایه نظریه گشتاری» (۱) در اختیار سیستم قرار گرفته‌اند و بدین‌طریق نیازهای آزمایشی سیستم را برآورده ساخته‌اند. بمنظور استفاده عملی از این سیستم مترجم، استادان مربوطه که به پیوست اسامی و آدرس آنان را تقدیم میدارم، اظهار علاقه مینمایند که با همکاری حداقل یکی از دانشگاههای ایران و بنا به پیشنهاد جنابعالی مبنی بر بهره‌گیری از توانائیهای دانشجویان دوره فوق لیسانس زبانشناسی در ایران در جهت تحقق این امر تلاشهایی را آغاز نمایند. اینجانب فارغ‌التحصیل دوره فوق لیسانس علوم کامپیوتر دانشگاه صنعتی شریف میباشم، لذا متأسفانه با استادان و محققین رشته زبانشناسی تماس و آشنایی چندانی نداشته و ندارم. استدعا دارم چنانچه بالاخص جنابعالی در این زمینه آماده همکاری هستید مستقیماً یا از طریق اینجانب با یکی از مراجع ذیل تماس حاصل نموده و امکانات موجود و بطور کلی نحوه این همکاری را تشریح کنید. اینجانب شخصاً پیشنهاد می‌کنم بخشی از کار تکمیل این پروژه بصورت پایان‌نامه فوق لیسانس در اختیار دانشجویان دوره فوق لیسانس رشته‌های زبانشناسی، زبان و ادبیات فارسی، و زبان و ادبیات زبانهای اشاره شده در ابتدای این نامه قرار گیرد. این امر اساساً متشکل خواهد بود از طراحی واژه‌نامه‌های یک زبانه برای زبان فارسی و واژه‌نامه‌های دو زبانه برای هر یک از زبانهای نه‌گانه فوق‌الذکر به زبان فارسی، که خود شامل واژه‌نامه عمومی و واژه‌نامه‌های تخصصی در هر یک از رشته‌های علوم کامپیوتر، ریاضیات، فیزیک، شیمی، علوم سیاسی... و غیره می‌باشد.

لازم است در اینجا به این مطلب اشاره کنم که سیستم تشریح شده در فوق تحت سیستم عامل UNIX و روی ماشینهای SUN کار میکند. بمنظور ارائه نتیجه عمل ترجمه بزبان فارسی، ضمن طراحی یک جدول نمایش میانی خط و حروف فارسی با استفاده از حروف انگلیسی، در واژه‌پرداز نسبتاً معروف LATEX که بالاخص بمنظور چاپ مقالات علمی کاربرد همه‌جانبه‌ای دارد، تغییرات لازم انجام یافته و این واژه‌پرداز اینک قادر به چاپ خروجی سیستم مترجم فوق‌الذکر بخط فارسی می‌باشد. جدول پیوست

که بکمک این واژه پرداز چاپ گردیده حاوی نمایش میانی حروف فارسی میباشد.

در خاتمه استدعا دارد چنانچه جنابعالی در این زمینه فرصت یا علاقه کافی ندارید با ارسال رونوشتی از نامه اینجناب برای همکاران گرامیتان در دانشگاههای دیگر و یا مراجعی که آنجناب صلاح میدانید آنان را در جریان امر قرار دهید تا در صورت وجود افراد و مؤسسات علاقه مند به تحقق این امر، یعنی دستیابی به یک سیستم مترجم چند زبانه زبان فارسی، بتوان از معلومات و تلاشهای آنان و احتمالاً کمکهای مالی و مشورتی مؤسسات مربوطه بهره مند گردید. بدیهی است چنانچه اشخاص یا موسساتی علاقه مند به مطالعه پایان نامه تحصیلی اینجناب باشند با کمال میل برایشان نسخه ای ارسال خواهد شد. اشاره مینماید که متأسفانه رساله مزبور در حال حاضر تنها بزبان آلمانی در دسترس است و در فرصتهای آتی ترجمه آن بزبانهای فارسی و انگلیسی انجام خواهد یافت.

در پایان باین مطلب مهم اشاره می‌کنم که پروژه EUROTRA تاکنون بیش از ۷۰ میلیون دلار هزینه در برداشته است و دست‌اندرکاران این پروژه در مورد پرداخت هزینه‌های بیشتر برای بالا بردن قابلیت سیستم مترجم خود ابائی ندارند و تنها در صورتیکه اطمینان لازم در مورد تحقق‌پذیری پیشنهادات فوق به آنان داده شود و یا نحوه انجام کار بطرز معقولی برایشان توجیه گردد، امکان بهره‌گیری از تسهیلات ابزارگانی، مشورتی، علمی و احتمالاً مالی آنان وجود داشته و خواهد داشت.

با نهایت سپاسگزاری و آرزوی همکاری بیشتر:

هوشنگ مهرجردیان - دانشکده انفورماتیک دانشگاه بُن

یادداشت:

۱ - منظور از روش ارائه شده در کتاب «دستور زبان فارسی بر پایه نظریه گشتاری» نوشته مهدی مشکوة‌الدینی (۱۹۷۰)، تعیین مشخصه‌های نحوی - معنایی واژه‌ها به صورت دو ارزشی (binary) و بر پایه ویژگیهای خاصی است که امکان می‌دهد هر واژه (یا واحد واژگانی) (lexical unit) با واژه‌های مناسب به کار رود، و از این راه جمله‌های درست تولید شود. در دستور زبان گشتاری، بر پایه مجموعه محدودی از قاعده‌های ساختاری یا بازنویسی (phrase-structure rules, rewriting rules) به همراه تعداد محدودی قاعده بازگشتی (recursive rules) و نیز با توجه به مشخصه‌های نحوی - معنایی یاد شده در بالا، تمامی جمله‌های بالقوه بی‌شمار زبان تولید می‌شود. قاعده‌های بازنویسی از نوع نمونه‌های زیر است:

جمله \Leftarrow گروه اسمی + گروه فعلی + عنصر صرفی

$$\left\{ \begin{array}{l} \text{(وابسته) + اسم} \\ \text{ضمیر} \end{array} \right\} \Leftarrow \text{گروه اسمی}$$

$$\left. \begin{array}{l} \text{فعل} \\ \text{گروه اسمی (را) + فعل} \\ \\ \text{گروه اسمی + فعل رابط} \\ \text{صفت} \\ \text{صفت + به نظر رسیدن / به نظر آمدن} \\ \text{گروه حرف اضافه‌ای + فعل رابط} \end{array} \right\} \leftarrow \text{گروه فعلی} \Leftarrow$$

عنصر صرفی \Leftarrow زمان + شناسه

قاعده‌های بازگشتی از نوع زیر است:

گروه اسمی \Leftarrow گروه اسمی + جمله

گروه فعلی \Leftarrow فعل + جمله

مشخصه‌های نحوی-معنایی واژه‌ها از راه تعداد محدودی قاعده‌های دستوری خاص که به آنها قاعده‌های زیر مقوله‌ای (subcategorization rules) گفته می‌شود به دست می‌آید. این قاعده‌ها از نوع نمونه‌های زیر است:

(الف) اسم \Leftarrow [+ اسم، \pm ذات]

(ب) [+ ذات] \Leftarrow [\pm شمردنی]

(ج) [+ شمردنی] \Leftarrow [\pm عام]

(د) [+ عام] \Leftarrow [\pm جاندار]

(ه) [- عام] \Leftarrow [\pm جاندار]

(و) [+ جاندار] \Leftarrow [\pm انسان]

(ز) [+ انسان] \Leftarrow [\pm مذکر]

(ح) [- انسان] \Leftarrow [\pm مذکر]

با به کار بستن قاعده‌های زیر مقوله‌ای از نوع نمونه‌های بالا برای هر واژه یک نماد مرکب (complex

symbol) می‌توان به دست داد، مانند نمونه‌های زیر:

طاووس [+ اسم، + ذات، + شمردنی، + عام، + جاندار، - انسان، {ویژگیهای تمایز دهنده}]

آب [+ اسم، + ذات، - شمردنی، {ویژگیهای تمایز دهنده}]

به ترتیب یاد شده در بالا، برای همهٔ واژه‌های اسم، فعل، صفت و قید می‌توان نماد مرکب به دست داد. بر پایهٔ نمادهای مرکب حاصل از قاعده‌های زیر مقوله‌ای، برای هر واژه محدودیت‌های بافتی (contextual restrictions) خاصی مشخص می‌شود و در نتیجه، تنها جمله‌های درست تولید می‌گردد. همچنین بر پایهٔ نمادهای مرکب، چگونگی انحراف (deviation) یا غیر دستوری بودن زنجیره‌های نادرست (ungrammatical strings) نیز توضیح داده می‌شود.