نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

طراحی و پیاده سازی یک سامانه ترجمه فارسی به انگلیسی

فائزه میرزائی[۱]، محسن بیگلری[۲] ، احمد یوسفان[۳] و عماد بیات[4]
دانشگاه کاشان؛ دانشکده مهندسی، گروه مهندسی کامپیوتر
fmirzaei@grad.kashanu.ac.ir[۱]
biglari@grad.kashanu.ac.ir[۲]
yoosofan@kashanu.ac.ir [۳]
emadbayat85@yahoo.com [4]

چکیده

ترجمه‌ی خودکار ماشینی از شاخه‌های فعال پردازش زبان طبیعی در سال‌های اخیر بوده است. هر سیستم پردازش زبان طبیعی نیاز به یک بخش پیش‌پردازش برای تبدیل ورودی به شکل مناسب دارد. این پیش‌پردازش می‌تواند شامل جداسازی عبارت ورودی به جملات، جملات به کلمات و اصلاح املای کلمات در صورت نیاز باشد.

در این مقاله یک سیستم ترجمه‌ی فارسی به انگلیسی با استفاده از روش مبتنی بر قوانین ارائه شده است که بخش پیش‌پردازش را نیز شامل می‌شود. در این سیستم بخش مهمی از پردازش‌ها به زمان طراحی سیستم منتقل شده است؛ و به موجب این انتقال، سرعت و دقت سیستم در تجزیه‌ی جملات افزایش چشمگیری پیدا کرده است. برای واژگان فارسی لغت‌نامه دهخدا به کارگرفته شده و برای ترجمه‌ی انگلیسی یک فرهنگ واژگان دوزبانه با نزدیک به ۴۵۰۰۰ لغت به کار گرفته شده است. در زمان اجرای سیستم جاری، تنها نیاز به تشخیص مواردی چون نشانه‌های جمع، زمان افعال و ضمایر متصل مفعولی می‌باشد. سیستم ارائه شده به صورت کامل پیاده‌سازی و ارزیابی شده است و نتایج بدست آمده، نشان داده است که چنین روشی کاملا کارآمد است.

کلمات کلیدی

ترجمه‌ی ماشینی، ترجمه‌ی فارسی به انگلیسی، لغت‌نامه دهخدا، پردازش زبان طبیعی

۱- مقدمه

ترجمه‌ی یک متن از یک زبان طبیعی به زبانی دیگر، ترجمه‌ی ماشینی نامیده می‌شود. در ترجمه‌ی خودکار ماشینی، عملیات ترجمه توسط کامپیوتر صورت می‌گیرد. این شاخه از پردازش زبان طبیعی در ۵۰ سال اخیر، موضوع پژوهش بسیاری از محققان بوده و توجه زیادی را به خود جلب کرده است. روش‌های ارائه شده را می‌توان به صورت کلی به چند بخش تقسیم کرد:

روش‌های مبتنی بر فرهنگ واژگان: در این روش‌ها، منبع اصلی ترجمه یک فرهنگ واژگان دوزبانه است که بسیار تعیین کننده بوده و قدرت ترجمه نیز تا حدود زیادی به آن بستگی دارد.

روش‌های مبتنی بر دانش یا مثال: در این روش‌ها، با استفاده از دانش و نوشتجات دوزبانه (مثال‌ها)، تک‌تک موارد از زبان مبدا به زبان مقصد نگاشت می‌شوند.

روش‌های مبتنی بر قوانین: این روش‌ها از مجموعه‌ای از قوانین برای ترجمه بهره می‌برند که براساس دانش زبان‌شناسی انسان نوشته شده است.

- روش‌های آماری: این روش‌ها، براساس مثال‌هایی از ترجمه‌ی انجام شده توسط انسان سعی بر یادگیری ترجمه دارند.

۲- مشکلات ترجمه فارسی به انگلیسی

گام اول در ترجمه، تجزیه‌ی جملات زبان مبدا است؛ از دشواری‌های تجزیه‌ی جملات زبان فارسی، می‌توان به موارد زیر اشاره کرد:

- بعضی از اسم‌ها دارای چند معنی متفاوت هستند، مانند شیر.
- بطور معمول، حروف صدادار کوچک در نوشتار فارسی ظاهر نمی‌شوند که موجب ابهام در کلمات متشابه می‌شود. برای مثال، کلمه‌ی «گل» را می‌توان به دو صورت «گِل» و «گُل» تفسیر کرد.

- انعطاف و تنوع جملات فارسی: قید در هر جای جملات (غیر از آخر) می‌تواند قرار گیرد. بسته به شرایط، ترتیب مفعول و متمم جابه‌جا می‌شود. اگر مفعول با «را» بیاید، قبل از متمم قرار می‌گیرد و اگر با «ی» بیاید، هم قبل و هم بعد از متمم می‌تواند قرار گیرد؛ و در صورتی که هیچکدام از این‌دو را نداشته باشد، معمولا بعد از متمم قرار می‌گیرد. در زبان عامیانه گاهی فعل در ابتدای جمله ظاهر می‌شود.
- حرف اضافه برای ارتباط دادن دو کلمه (اسم و صفت)، معمولا در نوشتار فارسی آورده نمی‌شود. مانند «گلِ قرمز»:«گل قرمز».
- وجود استثنائات زیاد در قوانین فارسی: برای مثال، علامت جمع فارسی «ان» اگر به کلمه‌ای که با «ه» پایان می‌یابد، اضافه شود، به «گان» تبدیل می‌شود.

۳- مروری با کارهای انجام شده

به صورت کلی، پروژه‌های انگشت‌شماری در رابطه با ترجمه‌ی فارسی به انگلیسی و برعکس صورت گرفته است؛ که اغلب آن‌ها نیز از روش‌های مبتنی بر قوانین استفاده کرده‌اند. پروژه‌ی شیراز [1] یکی از سیستم‌های نخستین در این رابطه است که با استفاده از یک ساختار از پیش تعریف شده و یک فرهنگ واژگان تقریبا ۵۰۰۰۰ کلمه‌ای پیاده‌سازی شده است. این سیستم از یک تجزیه‌کننده‌ی نحوی و تحلیل‌گر ریخت‌شناسی[1] بهره برده است. مراجع [2-4] نیز با استفاده از گرامر درخت مجاورت[2] یا TAG یک سیستم ترجمه‌ی انگلیسی به فارسی با استفاده از روش مبتنی بر قوانین ارائه کردند؛ سپس نسخه‌ی توسعه‌یافته‌ای از این سیستم را با استفاده از درخت‌های تصمیم آموزش دیده معرفی کردند. مترجم‌های تجاری همچون پارس و آریا نیز موجودند که دارای دقت بالایی نیستند. در [5] یک مترجم دوطرفه‌ی فارسی به انگلیسی و انگلیسی به فارسی به نام PEnTrans ارائه شده است؛ این مترجم از دو بخش اصلی با نام‌های PEnT1 برای ترجمه‌ی انگلیسی به فارسی و PEnT2 برای ترجمه‌ی فارسی به انگلیسی تشکیل شده است؛ در این مرجع، پس از معرفی سیستم مربوطه، مقایسه‌ای بین کارکرد آن و سیستم‌های موجود نیز صورت گرفته است تا برتری خود را نشان دهد.

۴- روش پیشنهادی

سیستم ارائه شده در این مقاله، از ترکیبی از روش مبتنی بر قوانین و فرهنگ واژگان بهره می‌برد. این سیستم از دو بخش اصلی تشکیل شده است:

۱. تجزیه‌ی ورودی فارسی
۲. تبدیل به معادل انگلیسی

بخش اول بسیار حائز اهمیت است زیرا از جهات مختلفی تعیین کننده است؛ هر اشتباهی در این بخش، منجر به ترجمه‌ی نادرست در بخش بعد خواهد شد. از طرفی، سرعت کل سیستم به صورت عمده توسط این بخش تعیین می‌شود. مشکل اغلب روش‌های ارائه شده در تحلیل و تجزیه‌ی جملات فارسی، سرعت پردازش است؛ این روش‌ها در زمان اجرا، به تحلیل کلمات در سطح‌های مختلف نحوی و معنایی می‌پردازند که موجب کاهش عمده‌ی سرعت می‌شود؛ در صورتی که بخش بزرگی از این پردازش‌ها تکراری است و قابل انجام در زمان طراحی: مانند ریشه‌یابی کلمات. از این‌رو، سیستم ارائه شده، سعی بر افزایش چشم‌گیر سرعت با انتقال حداکثر مقدار ممکن از پردازش‌ها به زمان طراحی دارد. این بهینه‌سازی به بخش اول سیستم اعمال شده است و در بخش مربوطه به تفصیل توضیح داده شده است.

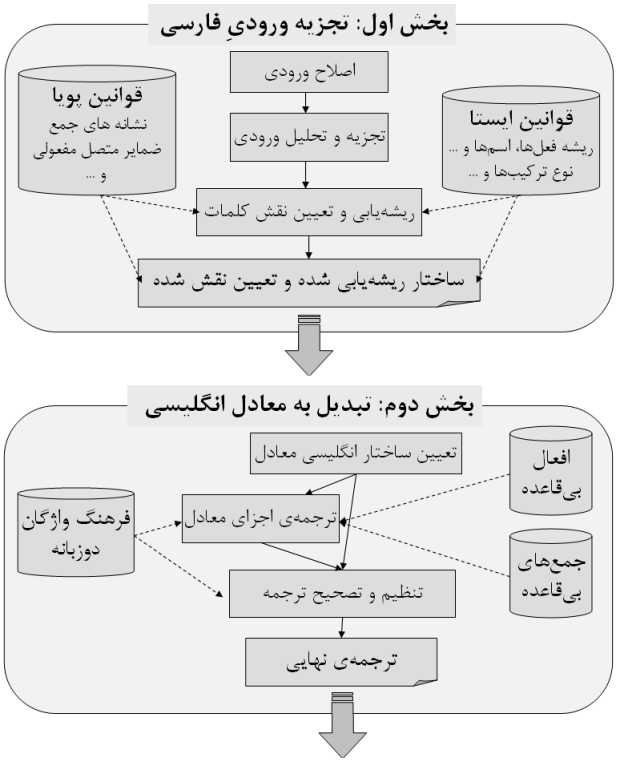
سیستم در گام اول یک فرهنگ لغت غنی تشکیل می‌دهد. سپس قوانین ساخت کلمات را برروی این فرهنگ لغت اعمال کرده و به تجزیه‌ی کلمات در سطح‌های مختلفی می‌پردازد؛ در نهایت، حاصل پردازش‌ها با یک ساختار مشخص در پایگاه داده قرار داده می‌شود. با استفاده از این ساختار، هر کلمه می‌تواند بدون هیچ پردازش اضافی، به اجزای تشکیل دهنده‌ی خود تجزیه شود. در زمان اجرا، سیستم با استفاده از پایگاه داده‌ی ایجاد شده، اقدام به تجزیه‌ی جملات ورودی کرده و نقش بخش‌های مختلف را تعیین می‌کند. حاصل این تجزیه، یک ساختار ریشه‌یابی شده و تعیین نقش شده است که برای بخش دوم سیستم ارسال می‌شود. بخش دوم سیستم که مسئول ترجمه‌ی انگلیسی است، ابتدا براساس ساختار جمله‌ی فارسی، ساختار انگلیسی معادل را محاسبه می‌کند؛ سپس با استفاده از فرهنگ واژگان دوزبانه، کلمات موردنظر را ترجمه کرده و براساس ساختار موردنظر، در قالب مشخصی قرار می‌دهد؛ در این مسیر، افعال بی‌قاعده و اسم‌هایی که دارای جمع‌های بی‌قاعده هستند نیز درنظر گرفته می‌شوند. در نهایت پس از تنظیم و تصحیح احتمالی، خروجی نهایی تولید می‌شود. شکل ۱، سیستم ارائه شده را نمایش می‌دهد.

۴-۱- بخش اول: تجزیه‌ی جملات فارسی

تشخیص درست و شناسایی کامل کلمات و عبارات، پیش‌نیاز یک ترجمه‌ی صحیح است. گام اول در شناسایی کلمات، ریشه‌یابی آن‌ها است؛ کلمات در زبان فارسی به صورت

کلی به انواع فعل، اسم، صفت، قید، ضمیر و حرف اضافه تقسیم می‌شوند [۶].

در سیستم ارائه شده از یک پایگاه داده‌ای توسعه یافته استفاده می‌شود؛ در این پایگاه داده برای دسته‌بندی کلمات و قرار دادن آن‌ها در پایگاه داده، کلمات همراه با نوع، بخش‌های تشکیل دهنده و اطلاعات اضافه‌تری قرار گرفته‌اند.



شکل (۱): ساختار سیستم ارائه شده

۴-۱-۱- ایجاد پایگاه داده‌ی ایستا

برای افزایش سرعت و کاهش ضریب خطا در ریشه‌یابی و تجزیه‌ی کلمات، قوانین ساخت کلمات به صورت ایستا پردازش و در پایگاه داده‌ای قرار داده می‌شود.

مرجع اصلی استفاده شده برای کلمات فارسی استفاده شده، فرهنگ دهخدا بوده است که در [7] درباره‌ی چگونگی ساخت آن توضیح داده شده است. کلمات براساس نوعشان در گروه‌های زیر از فرهنگ دهخدا استخراج شده‌اند:

- مصدر فعل
- بن ماضی
- بن مضارع
- اسم
- صفت
- قید
- ضمیر
- حرف اضافه
- میانوند
- پیشوند
- پسوند
- فعل‌های پیشوندی

کلمات استخراج شده، بازبینی و کلمات کم‌کاربرد و خاص هرس شده‌اند. این عمل موجب افزایش سرعت شده و از طرفی در مراحل بعدی کار، از تولید نتایج ناخواسته و دور از انتظار جلوگیری می‌کند. در نهایت و پس از پیش‌پردازش‌های موردنیاز، کلمات براساس گروه مربوطه در پایگاه داده قرار داده می‌شوند. در ساخت پایگاه داده، دو سطح قراردادی درنظر گرفته می‌شود؛ در سطح قراردادی اول، قوانینی که دارای درصد اطمینان بالاتر هستند، برروی کلمات موجود در فرهنگ لغت اعمال می‌شوند و در سطح قراردادی دوم، قوانینی که دارای درصد اطمینان پایین‌تر هستند، اعمال می‌شوند (جدول ۱). این روش موجب حذف بسیاری از ترکیبات نادرست می‌شود. جدول ۲ و ۳ چند نمونه از قوانین سطح قراردادی اول و دوم را نمایش می‌دهند.

جدول (۱): قوانین مربوط به سطح‌های قراردادی اول و دوم

سطح قراردادی اول	بن افعال + وند (پس/پیش)
سطح قراردادی دوم	اسم، صفت، ضمیر، قید + وند (پس/پیش)

جدول (۲): چند نمونه از قوانین سطح قراردادی اول برای اسم‌ها

بخش اول	بخش دوم	بخش سوم	نتیجه
بن مضارع	ش	-	اسم مصدر
بن مضارع	نده	-	اسم فاعلی
بن مضارع	گاه	-	اسم مرکب
بن مضارع	و	بن مضارع	اسم مصدر/مرکب

جدول (۳): چند نمونه از قوانین سطح قراردادی دوم برای اسم‌ها

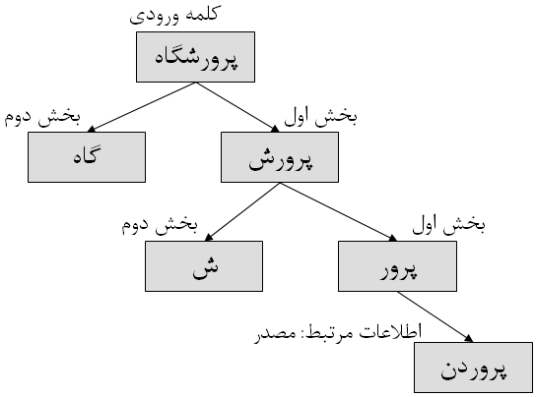
بخش اول	بخش دوم	بخش سوم	نتیجه
اسم	چه	-	اسم
اسم	اسم	-	اسم مرکب
اسم	بن مضارع	-	اسم
صفت	بن مضارع	-	اسم مکان

برای هر کلمه در پایگاه‌داده ساختار زیر درنظر گرفته شده است:

- یک کلید یکتا
- نوع کلمه
- سطح قراردادی
- اشاره‌گر به بخش اول کلمه (در صورت وجود)
- اشاره‌گر به بخش دوم کلمه (در صورت وجود)
- اشاره‌گر به بخش سوم کلمه (در صورت وجود)
- اطلاعات مرتبط با کلمه: این بخش برای هر نوع کلمه، می‌تواند کاربرد منحصری داشته باشد. برای مثال، برای افعال، می‌تواند مصدر/بن ماضی/بن مضارع آن را مشخص

کند. برای اسم‌ها، می‌تواند اسامی هم‌معنی آن را مشخص کند. برای صفت‌ها، می‌تواند شامل صفات هم‌معنی/متضاد آن باشد.

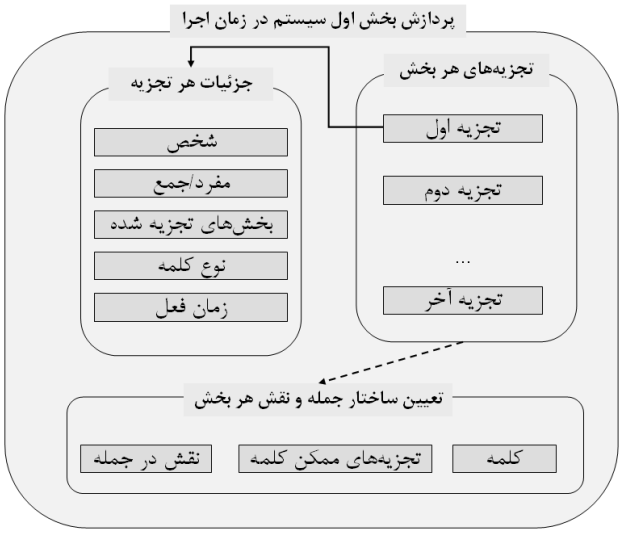
پایگاه داده‌ی جاری به شکلی طراحی شده است که با مراجعه به هر کلمه، می‌توان به بالاترین سطح از تجزیه‌ی آن دسترسی پیدا کرد. هر کلمه دارای یک کلید یکتا است و از یک یا چند بخش تشکیل شده است؛ هر بخش خود به کلمه‌ای دیگر اشاره می‌کند که می‌توان برای تجزیه‌ی بیشتر به آن مراجعه کرد. برای مثال کلمه‌ی «پرورشگاه» را درنظر بگیرید؛ که از دو بخش «پرورش» و «گاه» تشکیل شده است؛ حال اگر به جزئیات بیشتری نیاز باشد، می‌توان به کلمه‌ی «پرورش» مراجعه کرد که خود از دو بخش «پرور» و «ش» تشکیل شده است؛ با مراجعه به کلمه‌ی «پرور» می‌توان پی‌برد که این کلمه بن مضارع مصدر «پروردن» است.



شکل (۲): روند تجزیه‌ی یک کلمه تا آخرین سطح با استفاده از پایگاه داده‌ی طراحی شده

۴-۱-۲- پردازش پویا

منظور از پردازش پویا، اعمال سیستم در زمان اجراست؛ که شامل تجزیه‌ی جملات به بخش‌های کوچکتر و یافتن نقش هر بخش می‌باشد. هر کلمه در جمله ممکن است به چند طریق قابل تجزیه باشد، بنابراین بعد از جداسازی کلمه‌های موجود در جمله، تجزیه‌های ممکن هر بخش استخراج می‌شود؛ سپس برای هر تجزیه، اطلاعات مرتبط با آن محاسبه می‌شود؛ در نهایت مجموع این اطلاعات برای زیربخش بعدی که وظیفه‌ی تعیین نقش هر بخش از جمله را دارد، ارسال می‌شود. شکل ۳ روند عملیات را نمایش می‌دهد.



شکل (۳): روند عملیات سیستم برای پردازش یک جمله‌ی ورودی

در سیستم پیاده‌سازی، جملات خبری ساده مورد بررسی قرار گرفتند. هر جمله‌ی خبری به صورت کلی به دو بخش گروه نهادی و گروه گزاره‌ای قابل تقسیم است. گروه نهادی از اولین کلمه‌ی جمله آغاز و با حرف‌های ربطی همچون «و» و «یا» گسترش می‌یابد. باقی جمله نیز گروه گزاره‌ای خواهد بود. در صورتی که جمله دارای فعل اسنادی باشد، نقش بخش اول «نهاد» و در غیراینصورت «فاعل» خواهد بود. برای تعیین نقش هر کلمه در بخش دوم (گروه گزاره‌ای)، از اطلاعات موجود در جمله و اطراف هر کلمه استفاده می‌شود؛ برای مثال، مفعول بعد از نشانه‌ی مفعولی ظاهر می‌شود. نتیجه‌ی اجرای جمله‌ی ورودی «بچه و پدرش به مدرسه می‌روند» در شکل ۴ نمایش داده شده است. ساختارهای پشتیبانی شده در جدول ۴ فهرست شده‌اند.

جدول (٤): ساختارهای قابل شناسایی در تجزیه جملات

ساختار	مثال
فاعل + فعل	باران می‌بارد
فاعل + مفعول + فعل	کارگر آجرها را حمل می‌کند
فاعل + مفعول + متمم + فعل	مادر کودکش را به مدرسه می‌برد
نهاد + مسند + فعل اسنادی	هوا سرد است
فاعل + متمم + فعل	دانشجو به دانشگاه می‌رود
فاعل + متمم + مفعول + فعل	راننده با بوق مسیرش را باز کرد
نهاد + متمم + مسند + فعل اسنادی	هوا مانند زمستان سرد است

گروه نهادی (فاعل)

کلمه	سایر اطلاعات
بچه	اسم، مفرد
و	حرف اضافه، مفرد
پدرش	اسم، [پدر، ش]، سوم شخص مفرد

گروه گزاره‌ای

کلمه	نقش	سایر اطلاعات
به	نشانه متمم	حرف اضافه، مفرد
مدرسه	متمم	اسم، مفرد
می‌روند	فعل	فعل، [می، رو، ند]، سوم شخص جمع، مضارع اخباری، از مصدر رفتن

شکل (۴): نتیجه‌ی تجزیه یک نمونه جمله‌ی ورودی

۵- بخش دوم: ترجمه‌ی انگلیسی

فرهنگ واژگان دوزبانه‌ی کامل یکی از ضرورت‌های مهم در ترجمه‌ی انگلیسی است. در سیستم جاری از یک فرهنگ واژگان ۴۵۰۰۰ کلمه‌ای استفاده شده است. روش کار بخش دوم سیستم به این شکل است که برای هرکدام از هفت ساختار استخراج شده توسط بخش اول، ساختار معادل انگلیسی را درنظر می‌گیرد. در صورتی که ساختار ورودی یکی از این هفت حالت نباشد، از ترجمه‌ی بخش به بخش استفاده شده است؛ در این حالت بهینه‌سازی‌هایی برروی ورودی و خروجی نیز اعمال می‌شود؛ هرچند درنهایت خروجی به دقت حالت‌های قبل نیست. اسامی خاص با استفاده از حروف معادل انگلیسی جایگزین می‌شوند؛ هرچند به دلیل عدم وجود حرکت در اسامی خاص فارسی، ممکن است معادل‌سازی به صورت دقیق انجام نشود.

مراحل کلی بخش دوم از سیستم به صورت زیر است:

۱. تعیین ساختار انگلیسی معادل با ساختار فارسی ورودی
۲. ترجمه‌ی اجزای جمله‌ی ورودی به انگلیسی
۳. ترکیب اجزای ترجمه شده براساس ساختار مربوطه

جدول ۵ مثالی از چند جمله‌ی ورودی و معادل ترجمه‌ی شده‌ی آن‌ها را توسط سیستم ارائه شده نشان می‌دهد.

جدول (۵): چند نمونه از ترجمه‌های انجام شده توسط سیستم جاری

جمله‌ی ورودی	ترجمه
آریا آمد	Aria came
او با من آمد	He/She came with me
ناهید پولش را به من بخشید	Nahid spared her/his deposit to me

۶- نتیجه‌گیری

برای بررسی کارایی و سرعت سیستم ارائه شده، آن را توسط زبان برنامه‌نویسی پایتون[۳] و پایگاه‌داده‌ای SQLite که هردو متن‌باز و رایگان هستند، پیاده‌سازی و ارزیابی کردیم [8, 9]. نتایج در محدوده‌ی ساختارهای درنظر گرفته شده بسیار مناسب و دقیق بود. برای محاسبه‌ی یک مقیاس تقریبی از دقت سیستم، ۱۰۰ جمله در قالب هفت ساختار اشاره شده به عنوان ورودی به سیستم داده شد و بعد از بررسی دستی خروجی‌ها، نتایج زیر حاصل شد:

- ۸۳ جمله به صورت درست ترجمه شده بود.
- ۱۰ جمله به صورت تقریبی درست ترجمه شده بود.
- در ۷ جمله‌ی باقیمانده، از واژگان مناسب استفاده نشده بود

از مزایای این سیستم نسبت به سایر سیستم‌های موجود، می‌توان موارد زیر را نام برد:

- سرعت بسیار مناسب در تجزیه‌ی جملات در زمان اجرا
- دقت بالای تجزیه‌ی کلمات موجود در فرهنگ واژگان

و از مشکلات آن نیز می‌توان به موارد زیر اشاره کرد:

- درنظر نگرفتن همه‌ی ساختارهای موجود در جملات فارسی؛ مخصوصات جملات غیررسمی.
- عدم اعمال جنسیت (مرد/زن) در ترجمه‌ها.

مراجع

[1] Amtrup, J.W., et al., Persian-English machine translation: An overview of the Shiraz project. Memoranda in Computer and Cognitive Science MCCS-00-319, NMSU, CRL, 2000.

[2] Feili, H. and G. Ghassem-Sani. An application of lexicalized grammars in English–Persian translation. 2004. Citeseer.

[3] Feili, H. and G. Ghassem-Sani. Using Tree Adjoining Grammar in English-Persian Translation. 2004.

[4] Pilevar, M.T. and H. Faili. Persiansmt: A first attempt to english-persian statistical machine translation. 2010. JADT.

[5] Saedi, C., Y. Motazadi, and M. Shamsfard. Automatic translation between English and Persian texts. 2009.

[۶] گیوی، ا. . دستور زبان فارسی ۱. ویرایش سوم. ۱۳۸۷, تهران: موسسه فرهنگی فاطمی.

[۷] یوسفان, ا., جعفری, خ. و بیگلری, م., "تبدیل خودکار کلمه‌های لغت نامه دهخدا به قالب آوایی IPA "، دومین کنفرانس ملی مهندسی برق ایران. دومین کنفرانس ملی مهندسی برق ایران. دانشگاه آزاد اسلامی واحد نجف آباد ، ایران: اسفند ۱۳۸۸

[۸] بیات، ع. . مترجم فارسی به انگلیسی جملات. ۱۳۸۹. دانشگاه کاشان: دانشکده مهندسی- گروه کامپیوتر.

[۹] میرزائی، ف. ، بیگلری، م. . سیستم ترجمه جملات فارسی به انگلیسی. ۱۳۸۹. دانشگاه کاشان: دانشکده مهندسی- گروه کامپیوتر.

زیرنویس‌ها

[1] Morphological
[2] Tree Adjoining Grammar
[3] Python