روش‌های نوین برای ارزیابی سامانه‌های ترجمه ماشینی

پیمان ابوالقاسمی[۱]، محمد آزادنیا[۲]

[۱] پژوهشگر، پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران)، تهران،
abolghasemi@itrc.ac.ir

[۲] عضو هیات علمی رسمی، پژوهشگاه ارتباطات و فناوری اطلاعات (مرکز تحقیقات مخابرات ایران)، تهران،
azadnia@itrc.ac.ir

چکیده

این مقاله با هدف ارائه روش‌های ارزیابی مترجم‌های ماشینی تدوین گردیده است. بدین منظور ابتدا رویکردهای اصلی توسعه سیستم‌های ترجمه ماشینی تشریح می‌گردد. در بخش‌های بعدی روش‌هایی جهت ارزیابی این سیستم‌ها ارائه می‌گردد. این ارزیابی-ها در سه سطح ارزیابی کیفیت ترجمه، میزان سرویس دهی و سطح محبوبیت انجام می‌گیرد و برای هریک از این سه بخش ابزاری مجزاء طراحی و پیاده سازی شده‌است. همچنین برای هریک از این روش‌ها، معیارها و شاخص‌هایی استخراج شده است که با تحلیل نتایج بدست آمده از هر بخش می‌توان ارزیابی کاملی از سیستم‌های ترجمه نمود.

کلمات کلیدی

مترجم ماشینی، رویکرد آماری، مبتنی بر قانون، مبتنی بر عبارت، BLEU، TER ، N-gram و لاگ ترجمه.

۱- مقدمه

پیشینه ترجمه ماشینی به حدود ۶۰ سال پیش برمی‌گردد. تقریبا با ظهور اولین کامپیوتری که برای رمزگشایی کد در جنگ جهانی دوم به کار گرفته شد، ایده اولیه ترجمه ماشینی نیز بطور ضمنی مطرح شد: هر متن به زبان خارجی چیز جدیدی نیست مگر متن اصلی به‌صورت کد شده. از ابتدا تا کنون مسیرهای بسیاری در ترجمه ماشینی پیموده شده است. از روش‌های ساده ترجمه مستقیم که کار نگاشت ورودی به خروجی را انجام می‌داد، یا روش های پیچیده تر مبتنی بر انتقال که تحلیل ریخت شناسی و ساختاری را به خدمت می‌گیرد تا روش زبان میانی که از روش نمایش انتزاعی معنا استفاده می‌نماید. تفاوت رویکردهای متفاوت در طراحی یک مترجم ماشینی از آنجا ناشی می‌شود که محققان مختلف سطوح متفاوتی از دانش زبان شناسی در ارتباط با جمله مبدا و مقصد را برای تولید خروجی نهایی به استخدام گرفتند. از میان این روش ها، ابتدا روش هایی که بیشتر مبتنی بر قوانین زبان‌شناسی بودند در ترجمه ماشینی متون بکار گرفته شدند. پس از آن، حوزه ترجمه ماشینی بر روی روش های آماری متمرکز شدند.

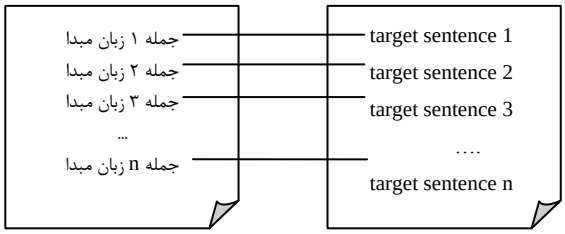
در این مقاله روش‌ها و معیارهایی جهت ارزیابی سیستم‌های ترجمه از جنبه‌های گوناگون ارائه شده‌است.

۲- روش های آماری

یکی از اصلی‌ترین رویکردهای ترجمه ماشینی روش‌های آماری است که سیستم‌های منطبق بر این ایده تا کنون موفقیت‌های چشمگیری نسبت به سیستم‌های قبلی داشته اند. رویکرد آماری ترجمه ماشینی را، مبتنی بر پیکره[۱] یا مشتق از داده[۲] نیز می‌نامند. چرا که سیستم طراحی شده، تمام اطلاعات مورد نیاز خود را، از یک پیکره موازی دو زبانه[۳] که مجموعه بسیار بزرگی از جملات هم ترجمه است، استخراج می‌نماید. شکل ۱ ساختار یک پیکره موازی را نشان می دهد. این پیکره یک مجموعه از جملات زبان مبدا به همراه مجموعه هم‌اندازه‌ای از جملات زبان مقصد است بدین صورت که ترجمه هر جمله در پیکره زبان مبدا، متناظرا در پیکره زبان مقصد موجود می باشد.

ایده ترجمه ماشینی با استفاده از روش‌های آماری از نظریه اطلاعات نشأت گرفته است. مبنای عملکرد یک مترجم آماری، نظریه تصمیم آماری است. تئوری تصمیم آماری روش شناخته شده ای برای ساخت یک سیستم تصمیم‌گیری مرکب از چندین منبع اطلاعاتی موجود، با هدف حداقل‌سازی

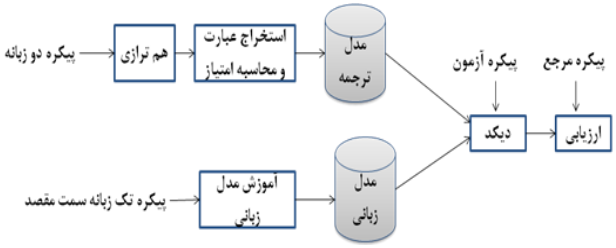
خطای تصمیم‌گیری است. پارامترهای چنین سیستمی با استفاده از مجموعه بزرگی از دادگان آموزشی تخمین زده می‌شوند. مدل های آماری شامل روش ترجمه مبتنی بر لغت، مبتنی بر عبارت، مبتنی بر نحو و سلسله مراتبی می باشد. روش مبتنی بر عبارت یکی از متداول ترین این روشهاست و در این مقاله توضیح داده می شود.



شکل (۱): ساختار یک پیکره دو زبانه موازی

۲-۱- مترجم آماری مبتنی بر عبارت

ترجمه آماری مبتنی بر عبارت نیز بر این مبنا استوار است که ترجمه جملات را با لحاظ کردن بسترشان و توجه به مفهوم متن، انجام می دهد. بدین شکل که برای ترجمه متن ورودی به جای لحاظ کردن کلمات به عنوان اجزای اصلی جملات، عبارات را به عنوان اجزای اصلی در نظر می‌گیرند. مفهوم عبارت در ترجمه مبتنی بر عبارت، با مفهوم عبارت در زبان‌های طبیعی متفاوت است. یک عبارت می‌تواند شامل تعدادی از کلمات همسایه باشد که در زبان طبیعی، معنای مستقلی ندارد و صرفا به خاطر تکرار بیشتر در متن آموزشی و تناظر لغوی، به عنوان یک واحد پرکاربرد، در نظر گرفته شده‌اند. بهره‌گیری از عبارات به جای کلمات، امکان یادگیری آسان‌تر جابجایی‌های محلی عبارات را ایجاد می‌کند. همچنین استفاده از عبارات، ترجمه گروه‌های مختلف کلمات، مانند اصطلاحات چند کلمه‌ای و درج و حذف کلمات را ساده‌تر می‌کند. شکل ۲ شمای کلی این روش را نمایش می دهد.



شکل (۲): شمای کلی ترجمه مبتنی بر عبارت

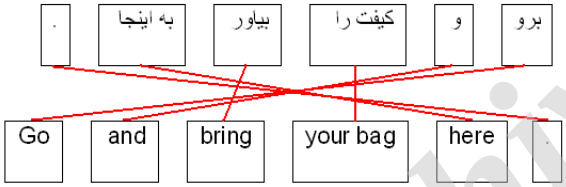
همانطور که در این شکل نمایان است، برای استخراج مجموعه عبارت یک پیکره، ابتدا لازم است جملات پیکره هم‌تراز شوند. این هم‌ترازی بایستی به‌صورت چند به چند باشد. برای این منظور عموما از متقارن‌سازی استفاده می‌شود بدین ترتیب که ابتدا هم‌ترازی جملات در هر دو طرف، از مبدا به مقصد و از مقصد به مبدا، تولید می‌شود و با الگوریتم هایی مبتنی بر اجتماع و یا اشتراک دو جدول را ترکیب می‌نمایند و هم‌ترازی چند به چند ایجاد می شود [۱].

در مرحله بعد با استفاده از جدول هم‌ترازی چند به چند، عبارات استخراج می‌شوند. برای استخراج عبارت دو قانون کلی بایستی رعایت شود و تنها عباراتی مجاز شمرده می‌شوند که از این دو قانون تبعیت نمایند.

$$\forall phrase\ pair\ (e,f) \Leftrightarrow \forall e_i \in e:(e_i,f_j) \in A \rightarrow f_j \in f$$
$$AND \quad \forall f_j \in f:(e_i,f_j) \in A \rightarrow e_i \in e$$

e و f جفت عبارت های ممکن می باشند و e_i و f_j کلمات این عبارات می باشند و A هم ترازی ایجاد شده می‌باشد. پس از استخراج عبارات، امتیاز آنها به روش شمارش محاسبه می شود.

حال طی این دو مرحله مدل ترجمه آموزش دیده است و می توان از این مدل برای ترجمه استفاده نمود. این مرحله دیکد نام دارد. مرحله دیکد در واقع شامل ساخت گراف حالات مختلف فرضیه‌های ممکن برای ترجمه جمله ورودی است. در این مرحله در رویکرد مبتنی بر عبارت ابتدا جمله ورودی به عبارت شکسته می‌شود. برای هر عبارت تمام ترجمه‌های ممکن از زبان مقصد لحاظ می‌شود. سپس در صورت نیاز عبارات زبان مقصد در جمله نهایی جابجا می‌شود. در انتها برای یافتن ترجمه مناسب برای جمله ورودی این گراف حالت جستجو شده و بهترین مسیر یافته شود. برای حل بهینه این مرحله از الگوریتم‌های جستجو مانند A* و جستجوی شعاعی استفاده می‌شود. همچنین گراف جستجوی حاصل هرس می‌شود تا حالت‌های تکراری و یا فرضیات با احتمال پایین از فضای حالت کم شده و پیچیدگی جستجو تا حدودی کاهش یابد. در نهایت پس از انجام ترجمه، می توان کیفیت آن را بر اساس روش های موجود بررسی نمود. این مرحله ارزیابی نام دارد. علاوه بر مدل ترجمه، مدل دیگری در کیفیت ترجمه تاثیر دارد که مدل زبانی نام دارد. آموزش مدل زبانی با استفاده از سمت زبان مقصد پیکره صورت می گیرد و نقش آن کنترل کیفیت و میزان فصاحت جمله تولید شده در زبان مقصد می‌باشد. شکل ۳ مثالی از خروجی این سیستم را نشان می‌دهد.



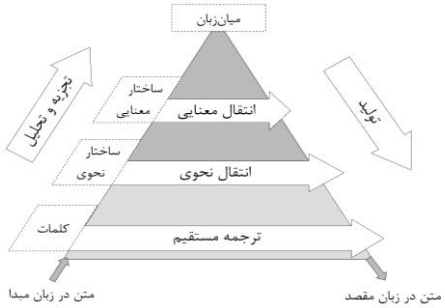
شکل(۳): مثالی از روش ترجمه مبتنی بر عبارت

۳- سیستم‌های مترجم ماشینی مبتنی بر قاعده

سیستم‌های مترجم ماشینی مبتنی بر قاعده در مجموع به سه دسته اصلی زیر تقسیم می‌شوند:

- ترجمه ماشینی مستقیم (Direct)
- ترجمه ماشینی انتقالی (Transfer)
- ترجمه ماشینی میان‌زبانی (Interlingua)

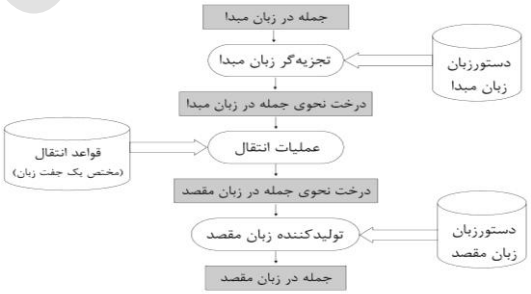
معمولا از مثلث شکل ۴ که هرم واکوئیس[۴] نام دارد، به عنوان بیان تصویری این دسته‌بندی‌ها استفاده می‌شود.قسمت پایینی این مثلث سیستم‌هایی را نشان می‌دهد که در فرآیند ترجمه، از هیچ سطحی از تجزیه زبانی بر روی جمله ورودی استفاده نمی‌کنند. با حرکت به سمت بالا در این مثلث، با سیستم هایی که از مقداری تجزیه صرفی-نحوی بهره می‌برند مواجه می‌شویم. در بالاترین نقطه این هرم، یک تجزیه معنایی بر روی جمله ورودی انجام شده و مفهوم جمله استخراج شده و ترجمه می‌گردد.



شکل(۴): مثلث Vaquise

در شکل ۵ معماری کلی یک سیستم انتقالی نشان داده شده است. جمله ورودی توسط یک تجزیه‌گر که معمولا مبتنی‌بردستور است، از لحاظ زبانی تجزیه می‌شود. ساختار جمله در زبان مبدا با کمک تعدادی قواعد انتقال (عملیات دوزبانه) به ساختاری تبدیل می‌گردد که از آن، جمله‌ای در زبان مقصد تولید خواهد شد.

ترجمه ماشینی انتقالی شامل سه مرحله است: تجزیه، انتقال و تولید. مرحله انتقال به مانند پلی است میان مراحل تجزیه جمله در زبان مبدا و تولید جمله در زبان مقصد. پس از اینکه جملات مبدا تجزیه گردیدند، به مجموعه‌ای از قواعد انتقال معنایی و انتقال لغوی نیاز است تا ساختار جمله در زبان مقصد تولید شود. قواعد انتقال معنایی کمک می‌کنند تا درخت تجزیه از زبان مبدا به زبان مقصد منتقل شود. این بخش در حقیقت نگاشتی بین ساختارهای درختی به یکدیگر است. ترجمه کلمات با استفاده از قواعد انتقال معنایی انجام می‌شود.



شکل (۵): ساختار کلی سیستم انتقالی

۴- ارزیابی خودکار کیفیت ترجمه ماشینی

به طور کلی این ارزیابی ها به منظور مقایسه ترجمه های صورت گرفته توسط ماشین با ترجمه های انسانی انجام می گردد، تا با تجزیه و تحلیل نتایج این ارزیابی ها به توسعه و پیش رفت ترجمه های ماشینی خودکار نائل گردیم. این ارزیابی ها در دو دسته قرار دارند:

الف) سنجش انسانی ب) سنجش خودکار و ماشینی

ارزیابی سیستم های ترجمه ماشینی توسط انسان بسیار وقت گیر و زمان بر است در نتیجه سیستم های ارزیابی خودکار ترجمه ماشین در دنیای امروز بسیار پر اهمیت و پر کاربرد می باشند. این سیستم ها از طریق امتیاز دهی به متن خروجی، بر اساس معیار های خاص و تعریف شده ای عمل می کنند. متداول ترین معیار ارزیابی در حال حاضر معیار BLEU۵ می‌باشد. در این مقاله به شرح این معیار می‌پردازیم[۳].

به منظور ارزیابی کیفیت سامانه‌های ترجمه ماشینی از روش‌های خودکار استفاده می‌گردد. به طور کلی روش‌های ارزیابی به دو بخش کلی تقسیم می-گردند:

- معیارهای مبتنی بر سنجش میزان دقت برحسب N-gram
- معیارهای مبتنی بر سنجش میزان خطا

۴-۱- معیارهای سنجش میزان دقت برحسب N-gram:

استراتژی مورد استفاده در این معیارها براساس تقسیم بندی جملات متون مرجع و متن کاندید به بخش های مختلف و بررسی میزان برابری و تطابق این بخش ها می باشد. متداول‌ترین معیار در این حوزه معیار BLEU است. روش BLEU جایگزینی برای روش های ارزیابی انسانی نیست بلکه پیشنهادی است که به ما درارزیابی های مکرر و سریع کمک می‌کند. طرزکار متد BLEU بدین ترتیب است که n-gram های ترجمه‌ای که توسط ماشین انجام شده است (ترجمه کاندید) را با n-gram های ترجمه ی انسانی مقایسه می‌کند و تعداد تطابق‌ها را شمارش می‌کند. این تطابق مستقل از جایگاه n-gramها بوده و هرچه تعداد آن بیشتر باشد نشان‌دهنده کیفیت بهتر متن ترجمه شده است. همان‌گونه که مشاهده می‌شود اساس روش BLEU بر مبنای محاسبه دقت است که از تقسیم تعداد n-gramهای منطبق بر تعداد کلمات در متن کاندید بدست می‌آید.

$$p = \frac{m}{w_t}$$

(در فرمول بالا m تعدادunigramهای نگاشت شده بین عبارت مرجع با عبارت ترجمه ماشین و w_t تعداد unigram ها در متن ترجمه شده است.)

بدین منظور ابتدا ماکزیمم تعداد رخداد هر کلمه در هر متن مرجع شمرده شده و سپس تعداد کل هر کلمه در متن کاندیدا به ماکزیمم تعداد رخداد آن کلمه در متن مرجع کاهش می‌یابد. به عبارت دیگر معیار BLEU تنها یک تطبیق را به ازای هر لغت مرجع امکان پذیر می‌کند.

از همین روش می‌توان برای محاسبه دقت اصلاح شده n-gram در سطح پیکره متنی نیز استفاده کرد. اگرچه ارزیابی سیستم ترجمه ماشینی به ازای کل متن انجام می‌گیرد اما واحد اصلی ارزیابی در سطح جمله است. لذا برای محاسبه امتیاز کلی، تطابق N-gramها جمله به جمله محاسبه می-شود. محاسبه امتیاز دقت در سطح پیکره متنی به شرح زیر است:

$$p_n = \frac{\sum_{ce\{Candidates\}} \sum_{n-gram \epsilon C} Countclip(n-gram)}{\sum_{ce\{Candidates\}} \sum_{n-gram \epsilon C} Count(n-gram)}$$

• ترکیب دقت برای N-gram ها با طول مختلف:

یکی از مسائل مهم در این حوزه چگونگی ترکیب دقت به ازای اندازه‌های مختلف N-gram است. ساده‌ترین روش ممکن میانگین‌گیری ساده از دقت N-gramها به ازای اندازه‌های مختلف است اما تحقیقات نشان داده است که دقت تک‌گرم‌ها به مراتب بیشتر از دقت دوگرم‌ها است که به نوبه خود از سه-گرم‌ها دقت بیشتری دارند. بنابراین روش میانگین‌گیری مناسب باید این روند نمایی را مدنظر قرار دهد. براین اساس معیار BLEU از لگاریتم میانگین با وزن‌دهی یکسان استفاده می‌کند که معادل با استفاده از میانگین هندسی دقت N-gramها است. به صورت تجربی، بهترین همبستگی با ترجمه انسانی در استفاده از ماکزیمم N-gram از درجه ۴ بدست آمده است [۲].

جدول (۱) : نمونه‌ای از متون مرجع و کاندید

متن مرجع ۱	"I always do"
متن مرجع۲	" I invariably do"
متن مرجع۳	"I perpetually do"
متن ترجمه شده۱	"I always invariably perpetually do"
متن ترجمه شده۲	"I always do"

• **طول جملات کاندید:**

طول ترجمه کاندید از جمله مسائل مهم در محاسبه امتیاز BLEU است. در صورتی که این طول کوتاه باشد ولی طول ترجمه جمله مرجع بیشتر باشد در این صورت BLEU امتیاز بیشتری کسب می کند زیرا تعداد N-gram های مشابه بیشتر می‌شود. این امر نشان‌دهنده ضعف BLEU در ارزیابی جملات کوتاه است.

جدول (۲) : میزان تفاوت و شباهت متون

متن مرجع۱	It is the guiding principle which quarantees the military forces always being under the command of the party
متن مرجع۲	It is the practical guide for the army always to heed the directions of the party
متن ترجمه شده	of the

همان‌طور که در جدول ۲ مشاهده می‌نمائید "of the" در همه متون مرجع وجود دارد. در این حالت با وجود اینکه ترجمه کاندید صحیح نیست، امتیاز کامل را دریافت می‌کند.

$$\text{Modified Unigram precision} = \frac{2}{2};\ \text{Modified Bigram precision} = \frac{1}{1}$$

به طور معمول برای حل مشکل دقت در رویارویی با جملات کوتاه از مفهوم یادآوری استفاده می‌شود. اما روش BLEU از چندین متن مرجع استفاده می‌کند که هر کدام ممکن است از کلمات متفاوتی برای ترجمه کلمه اولیه در متن اصلی استفاده کرده باشند. این درحالی است که یک کاندید ترجمه خوب تنها باید از یکی از این انتخاب‌های ممکن استفاده کند زیرا یادآوری همه این لغات به یک ترجمه نامناسب و بی‌کیفیت منجر می‌شود.

برای حل این مشکل در زمانی که تعداد کلمات ترجمه کاندید کمتر از ترجمه مرجع است، جریمه‌ی اختصار[۶] در نظر گرفته می‌شود. با درنظرگرفتن جریمه‌ی اختصار، ترجمه کاندید تنها زمانی امتیاز بالایی دریافت می‌کند که در اندازه، انتخاب کلمات و ترتیب لغات با متن مرجع انطباق داشته باشد. زمانی که طول متن کاندید با یکی از متون مرجع یکسان باشد، مقدار این جریمه معادل با ۱,۰ خواهد بود. بنابراین همواره در محاسبه جریمه اختصار، متن مرجع با بیشترین شباهت طولی[۷] با متن کاندید مدنظر قرار می‌گیرد. نکته مهم دیگر که باید مدنظر قرار گیرد، این است که اگر جریمه اختصار در سطح هر جمله از متن محاسبه و در نهایت میانگین‌گیری شود، انحرافات طولی جملات کوتاه به جریمه سنگینی منجر می‌شود. بنابراین بهتر است محاسبه جریمه در سطح پیکره متنی انجام گیرد تا در انتخاب اندازه جملات، آزادی بیشتری وجود داشته باشد. بر این اساس محاسبه جریمه اختصار به شرح زیر است:

$$BP = \begin{cases} 1 & \text{if } c > \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

در رابطه بالا فاکتور r "طول موثر مرجع"[۸] است که از طریق تجمیع جملاتی که بیشترین شباهت طولی[۹] با هر جمله در متن کاندید را دارند، محاسبه می‌شود و فاکتور c برابر با اندازه پیکره متنی کاندید است.

برای محاسبه BLEU با درنظرگرفتن جریمه اختصار (BP)، ابتدا میانگین هندسی دقت اصلاح‌شده n-gramها (p_n) براساس n-gramهایی با حداکثر طول N و با استفاده از اوزان مثبت w_n با مجموع یک محاسبه می‌شود و سپس نتیجه در فاکتور جریمه اختصار ضرب می‌شود:

$$\text{BLEU} = \text{BP}.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

در معیار اصلی BLEU، مقدارNبرابر با ۴ است و از وزن‌دهی یکنواخت استفاده می‌شود ($w_n = 1/N$). امتیازی که BLEU به ماشین های ترجمه می‌دهد در مقیاس ۰-۱ است و می‌تواند به صورت درصد نیز بیان شود. یکی از مسائل مورد توجه تعداد متون مرجعی است که می‌تواند مورد استفاده قرار گیرد. در حالت ایده‌آل هرچقدر تعداد این متون بیشتر باشد، میزان انطباق بیشتر و کیفیت ارزیابی به طور کلی بهتر است. در روش‌های مبتنی بر متون مرجع عموما از ۴ متن استفاده می‌شود.

۴-۲- معیارهای مبتنی بر سنجش میزان خطا:

در این معیارها تعداد حداقل جایگزینی، حذف، جابجایی و اضافه نمودن کلمات جملات ترجمه برای رسیدن به جمله مرجع مورد سنجش قرار می‌گیرد. از این معیارها به طور گسترده در بازشناسی گفتار نیز استفاده می‌شود.

در این دسته، معیار TER به عنوان اصلی ترین معیار ارزیابی نرخ خطای ترجمه در نظر گرفته می شود و در ارزیابی‌ها از این معیار استفاده می‌شود.

معیار TER برابر حداقل تعداد ویرایش جمله مورد نظر برای رسیدن به یک جمله ای دقیقا مشابه با یکی از جملات مرجع است. با توجه به انکه در این روش حداقل ویرایش مورد توجه قرار می گیرد، ما تنها تعداد ویرایش های نزدیک ترین جمله مرجع را در نظر می گیریم. به طور مشخص این معیار به صورت زیر محاسبه می شود:

$$\text{TER} = \frac{\text{تعداد کل ویرایش ها}}{\text{میانگین تعداد کل کلمات جملات مرجع}}$$

به طور کلی ویرایش های ممکن در این معیار عبارتند از:

۱-حذف. ۲-جابجایی. ۳-جایگزینی. ۴- الحاق

در ویرایش حذف کلماتی که در جمله کاندید موجود بوده ولی در جملات مرجع موجود نمی باشد حذف می گیردد. در جابجایی نیز به منظور هماهنگی جمله مرجع با جملات مراجع، دنباله ای از کلمات از یک نقطه از جمله به نقطه ای دیگر منتقل می گردند. نکته قابل توجه در این معیار این است که علائم نگارشی نیز به عنوان کلمه معمولی در نظر گرفته می شوند.

۴-۳- ارزیابی کیفیت ترجمه سیستم‌های ترجمه ماشینی انگلیسی-فارسی

با استفاده از معیار BLEU که در سامانه ارزیابی مورد نظر توسعه داده شده-است، سیستم‌های ترجمه مورد ارزیابی و نمره‌دهی قرار گرفتند. در جدول (۳) نتایج این ارزیابی‌ها را مشاهده می‌نمائید.

جدول (۳) : نتایج ارزیابی سامانه‌های ترجمه انگلیسی-فارسی

نام سرویس	میزان کیفیت بر حسب معیار BLEU
مترجم ترگمان	۳۱
مترجم فرآزین	۲۷
مترجم پارس	۱۵
مترجم گوگل	۲۷
مترجم شورای عالی اطلاع رسانی	۸

۵- ارزیابی سطح و میزان سرویس دهی سیستم-های ترجمه ماشینی

ارزیابی سطح سرویس‌دهی در اصل سنجش کمی میزان پاسخ‌گوئی سیستم-های ترجمه ماشینی می‌باشد. بدین منظور ابزاری طراحی و پیاده‌سازی گردیده است که توانایی تحلیل و بررسی لاگ‌های ترجمه را در این سامانه‌ها دارا می-باشد. این لاگ‌ها در زمان ارسال درخواست ترجمه از سوی کاربران بر روی سرورهای ترجمه ذخیره می‌گردند و دارای اطلاعات مفیدی بوده که می‌توان از آنها در ارزیابی‌ها استفاده نمود. این ابزار با هدف تحلیل و استخراج اطلاعات مورد نیاز از لاگ های ترجمه طراحی و پیاده سازی شده است. این ابزار دارای بخش های مختلفی می باشد که این بخش ها عبارتند از:

- سنجش تعداد لغات ترجمه گردیده
- سنجش مدت زمان ترجمه لغات
- سنجش سرعت پاسخگوئی
- سنجش تعداد کاربران یکتا درخواست کننده
- سنجش تعداد درخواستهای ترجمه ارسالی
- سنجش میزان پایداری و خطا دسترسی

۵-۱- معیار و شاخص‌های بکار رفته در تحلیل لاگ-های ترجمه

به منظور امتیازدهی به سیستم‌های ترجمه برحسب اطلاعات بدست آمده از لاگ‌های آنها، معیارهایی استخراج گردیده است و مطابق با آنها به سیستم-های ترجمه نمره دهی می‌گردد.

کلیه اطلاعات موجود در لاگ‌ها را می‌توان در بازه‌های زمانی متفاوت و برحسب مدت زمان‌های انتخابی (واحدهای تقسیم بندی نمودارها) از طریق ابزار تحلیل‌گر لاگ به نمایش در آورد.

در زیر شاخص‌های اصلی بکار رفته در ابزار تحلیل‌گر لاگ‌های ترجمه ارائه گردیده است.

۱) مجموع تعداد درخواست های ارسالی:

این مؤلفه تعداد کل درخواست‌های ترجمه ارسالی از کاربران به وب سایت در بازه درخواست شده را به نمایش می گذارد. توجه داشته باشیم که هر کاربر می‌تواند چندین درخواست ترجمه را ارسال نماید.

۲) مجموع تعداد کاربران یکتا :

این مولفه تعداد کاربران یکتا استفاده کننده از سیستم ترجمه را در بازه‌های زمانی متفاوت نمایش می‌دهد. محاسبه تعداد کاربران از طریق اختصاص یک شماره ID منحصر به فرد به هریک کاربران انجام می‌گیرد. این شماره ID به صورت خودکار بروی سیستم‌های کاربران (در قالب یک کوکی از مرورگر آنها) ذخیره می‌شود. این امر به منظور تشخیص کاربران حقیقی از کاربران غیرحقیقی (ربات‌ها) صورت پذیرفته‌است.

۳) میانگین طول کلمات ترجمه شده:

میانگین تعداد کلماتی که در هر درخواست در بازه تعیین شده برای سامانه ارسال شده است توسط این مؤلفه قابل بازیابی است.

۴) مجموع تعداد لغات ترجمه گردیده:

تعداد کل کلمات ترجمه شده توسط سامانه ترگمان در بازه مشخص در این قسمت از جدول نشان داده می‌شود. این اطلاعات قابل سنجش در بازه‌های زمانی متفاوت می‌باشد.

۵) بیشترین طول کلمات:

بیشترین تعداد کلمات موجود در یک متن ارسالی جهت ترجمه در بازه مشخص شده توسط تحلیل‌گر را نشان می‌دهد.

۶) کمترین طول کلمات:

کمترین تعداد کلمات موجود در یک متن ارسالی جهت ترجمه در بازه مشخص شده توسط تحلیلگر را نشان می‌دهد.

۷) میانگین مدت زمان ترجمه:

میانگین زمانی را که سامانه در آن به درخواست‌های ارسالی ترجمه در بازه دلخواه پاسخ داده است را می‌توان در این بخش مشاهده نمود.

۸) بیشترین مدت زمان ترجمه:

حداکثر زمانی را که سامانه در آن به درخواست‌های ارسالی ترجمه در بازه دلخواه پاسخ داده است را می‌توان در این بخش مشاهده نمود.

۹) کمترین مدت زمان ترجمه:

حداقل زمانی را که سامانه در آن به درخواست‌های ارسالی ترجمه در بازه دلخواه پاسخ داده است را می‌توان در این بخش مشاهده نمود.

۱۰) تعداد درخواست‌های موفق و ناموفق:

در این بخش می‌توان کلیه درخواست‌های ترجمه که با موفقیت پاسخ داده شده‌اند و همچنین تعداد درخواست‌هایی که با خطا مواجه شده‌اند را مشاهده نمود. علاوه بر موارد فوق نوع خطاهای مواجه شده (شبکه‌ای، سیستمی و...) را نیز تشخیص می‌دهد و به نمایش کاربران در می‌آورد.

اطلاعات خروجی ابزار تحلیل لاگ نتایج مهمی را در اختیار تیم‌های توسعه قرار می‌دهد و با توجه به اطلاعات دریافتی از این ابزار، می‌توانند کلیه جوانب و قسمت های سرویس ترجمه را با در نظر گرفتن کلیه ویژگی ها و خصوصیات این سیستم‌ها، مورد ارزیابی قرار دهند.

۶- ارزیابی سطح محبوبیت و تحلیل رفتار کاربران بازدید کننده

با نصب و راه اندازی ابزار تحلیل‌گر وب سایت بروی سایت مترجم‌های ماشینی می‌توان اطلاعات کاملی از رفتار کاربران بدست آورد. این ارزیابی‌ها امکان بررسی و تحلیل مراجعات انجام شده به وب سایت را در اختیار ما قرار می‌دهد. این سامانه اطلاعات گوناگونی از وضعیت وب سایت ترجمه، در قالب نمودارها و جداول در اختیار تیم توسعه قرار می دهد. نمودارهای اصلی این سامانه عبارتند از:

- تحلیل و دسته بندی بازدیدها از جنبه های گوناگون مانند:
  - ✓ بازدیدهای بلادرنگ

- ✓ دسته بندی از جنبه زمان و تاریخ
- ✓ بازدیدها براساس مکان جغرافیایی
- نمایش میزان زیربار بودن سرورها در ساعات شبانه روز
- نمودار بازدید بر حسب تعداد روزهای هفته
- ثبت لاگ برای هر یک از کاربران بازدید کننده
- نمودار دوره تکامل وب سایت ترجمه
- نمایش تعداد کاربران بازگشتی به همراه گراف فعالیت
- نمایش نحوه ارجاع کاربران با سایت ترجمه

۶-۱- اطلاعات بکار رفته در تحلیل رفتار کاربران

۱) میزان تکامل و رشد سایت ترجمه:

در این بخش می‌توان سیر تکاملی وب سایت ترجمه را برحسب معیار هایی از قبیل تعداد بازدیدها مشاهده نمود.

۲) مشخصات کاربران از جنبه منطقه جغرافیایی :

این بخش به نمایش آمار کاربران به تفکیک قاره، کشور، شهر و منطقه می-پردازد. همچنین می‌توان روند تکامل وب سایت ترجمه به همراه درصد بازدیدکنندگان از هر بخش را مشاهده نمود.

۳) بازدیدها بر اساس مدت زمان سپری‌شده از آخرین بازدید:

در این قسمت می‌توان علاقه افراد به استفاده مستمر از وب‌سایت را مورد بررسی قرار داد، در صورتی که فاصله کمتری میان بازدیدهای کاربران با بازدید آخر آنها وجود داشته باشد نشان از علاقه کاربران به وب سایت می-باشد.

۴) سنجش میزان زیربار بودن سرورها:

این بخش اطلاعاتی در زمینه میزان استفاده از سرورها به تفکیک ساعات شبانه روز را نمایش می‌دهد. اطلاعات با اهمیتی که از این نمودار می توان استخراج نمود این است که ساعات پرکار سرویس ترجمه را استخراج نموده و برای آن زیرساخت بیشتری در نظر گرفت.

۵) میانگین زمان سپری شدن بازدید کنندگان در سایت:

در این قسمت می‌توان مدت زمان سپری کردن هر کاربر در سایت را ثبت نمود. هر چه این مدت زمان بیشتر باشد، نشان از استقبال کاربران دارد.

۶) بیشترین عملیات انجام گردیده از جانب کاربران:

در این قسمت می‌توان تعداد عملیات کاربران بروی سایت و همچنین میانگین فعالیت‌های انجام داده از کاربران بروی سایت را مورد سنجش قرار داد.

۷) گراف بازدید کنندگان بازگشتی:

در این گراف‌ها می‌توان آماری از میزان بازگشت کاربران سایت و تعداد عملیاتی که در هر بازدید بروی سایت انجام می‌دهند را بدست آورد.

۸) دسته بندی کاربران براساس تعداد بازدیدهای آنها:

برای این بخش جدولی تعبیه شده‌است که کاربران را برحسب تعداد بازدیدشان از وب‌سایت تقسیم بندی می‌نماید. تعداد بازدیدهای از یک بازدید به بالا می‌باشد.

۹) بازدیدها بر اساس روزهای هفته:

این بخش شامل تعداد مراجعه کنندگان به تفکیک روزهای هفته می‌باشد.

۷- نتیجه

در این مقاله پس از تشریح روش‌های پیاده‌سازی سیستم‌های ترجمه ماشینی، روش‌ها و معیارهایی جهت ارزیابی این سیستم‌ها ارائه شده است. در پایان هریک از سامانه‌های ترجمه انگلیسی–فارسی موجود از جنبه کیفیت ترجمه ارزیابی شده و همچنین شاخص‌ها و معیارهایی نیز جهت ارزیابی این سامانه‌ها از جنبه‌های گوناگون ارائه شده است.

مراجع:

[1] F. J. Och, and H. Ney, (2003), “A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, volume 29, number 1, pages 19-51, Cambridge, MA, USA.

[2] Papineni, k., Roukos, S., Ward, T. and Zhu, W.J (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[3] Lavie,A. (2010) Evaluating the Output of Machine Translation Systems, AMTA 2010 Tutorial. Denver, Colorado, USA, October 31, 2010

[4] Al-Onaizan, Yaser, Jan Curin, Michael Jahr, Kevin Knight, John D. Lafferty, I. Dan Melamed, David Purdy, Franz J. Och, Noah A. Smith, and David Yarowsky. 1999. Statistical machine translation. Final Report, JHU Workshop.

[5] 2009, Callison-Burch, C., P. Koehn, C. Monz and J. Schroeder, “Findings of the 2009 Workshop on Statistical Machine Translation”, In Proceedings of the Fourth Workshop on Statistical Machine Translation at EACL-2009, Athens, Greece, March 2009. Pages 1- 28.

[6] 2009, Snover, M., N. Madnani, B. Dorr and R. Schwartz, “Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric”, In Proceedings of the Fourth Workshop on Statistical Machine Translation at EACL-2009, Athens, Greece, March 2009. Pages 259-268.

[7] 2007, Lavie, A. and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments" . In Proceedings of the Second Workshop on Statistical Machine Translation at the 45th Meeting of the Association for Computational Linguistics (ACL-2007), Prague, Czech Republic, June 2007. Pages 228- 231.

[8] 2008, Agarwal, A. and A. Lavie. "METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output" . In Proceedings of the Third Workshop on Statistical Machine Translation at the 46th Meeting of the Association for Computational Linguistics (ACL-2008), Columbus, OH, June 2008. Pages 115-118.

زیرنویس‌ها

1 Corpus-based
2 Data driven
3 Bilingual parallel corpus
4 Vauquois
5 Bilingual Evaluation Understudy
6Brevity Penalty Factor
7Best Match Length
8 Effective Reference Length