



پایایی نمرات ترجمه‌های انسانی با بهره‌گیری از ابزارهای خودکار جانشین ارزیابی کیفیت ترجمه

سمیه کرمی*

دانشجوی دکتری رشته ترجمه، دانشگاه اصفهان، اصفهان، ایران

Email: s.karami@fgn.ui.ac.ir

داریوش نژادانصاری**

(نویسنده مسئول)

استادیار گروه زبان و ادبیات انگلیسی، رشته آموزش زبان انگلیسی، دانشگاه اصفهان، اصفهان، ایران

Email: d.nejadansari@fgn.ui.ac.ir

اکبر حسابی***

استادیار گروه زبان و ادبیات انگلیسی، رشته زبان‌شناسی همگانی، دانشگاه اصفهان، اصفهان، ایران

Email: a.hesabi@fgn.ui.ac.ir



چکیده

با توجه به ماهیت فرایند ارزیابی ترجمه که از لحاظ زمان، انرژی و هزینه قابل تامل می‌باشد، بهره‌گیری از فن‌آوری‌های نوین در حوزه ترجمه ماشینی منطقی به نظر می‌رسد. ابزارهای خودکار جانشین ارزیابی کیفیت ترجمه یکی از این فن‌آوری‌ها است که در حوزه ترجمه ماشینی کاربرد دارد. این پژوهش در صدد یافتن پاسخ این سؤال است که پایایی نمرات این ابزارها در سطح واژگان به ترجمه‌های انسانی (۵۱ دانشجوی سال آخر رشته ترجمه در ایران) با استفاده از ۱، ۲، ... تا ۵ ترجمه معیار به صورت مرحله به مرحله و جداگانه چه تغییری می‌کند. لذا پژوهشی تجربی و کاربردی با رویکردی کمی برای محاسبه میزان پایایی نمرات این ابزارها در مقایسه با میانگین نمرات ۵ ارزیاب متخصص انجام شد. میزان رابطه همبستگی میان این دو مجموعه نمره (در حالت‌های مختلف استفاده از ۱، ۲، ... تا ۵ ترجمه معیار) به منزله پایایی نمرات ابزار خودکار تفسیر شده است. نتایج تحلیل آزمون همبستگی پیرسون نشان داد که استفاده از ۵ ترجمه معیار در ۳۷/۸۰ درصد موارد منجر به بالاترین میزان رابطه همبستگی شده است که بیشتر از هر حالت دیگر در این پژوهش است (۴ ترجمه معیار (۳/۶۵ درصد)، ۳ ترجمه معیار (۱۰،۹۷ درصد)، ۲ ترجمه معیار (۳۱/۷۰ درصد) و ۱ ترجمه معیار (۱۵/۸۵ درصد)). بنابراین، فرضیه پژوهش تایید می‌شود که استفاده از ترجمه‌های معیار بیشتر منجر به رابطه همبستگی بالاتر و پایایی بیشتر نمرات می‌شود. در عین حال، استفاده از ۲ ترجمه معیار جایگاه دوم را از نظر دستیابی به بالاترین میزان رابطه همبستگی دارد و فرضیه پژوهش را نقض می‌کند.

اطلاعات مقاله

تاریخ ارسال: ۱۳۹۹/۰۶/۱۵
تاریخ پذیرش: ۱۳۹۹/۰۷/۱۵
تاریخ انتشار: پاییز ۱۳۹۹
نوع مقاله: علمی پژوهشی

کلید واژگان:

ارزیابی کیفیت ترجمه، ابزارهای خودکار جانشین ارزیابی کیفیت ترجمه، نمره‌دهی خودکار، پایایی، ترجمه معیار

شناسه دیجیتال DOI: 10.22059/jflr.2020.309025.751 کلیه حقوق محفوظ است ۱۳۹۹

کرمی، سومیه، نژادانصاری، داریوش، حسابی، اکبر (۱۳۹۹). پایایی نمرات ترجمه‌های انسانی با بهره‌گیری از ابزارهای خودکار جانشین ارزیابی کیفیت ترجمه. پژوهش‌های زبان‌شناختی در زبان‌های خارجی، (۱۰)، ۳، ۶۱۸-۶۲۹.
Karami, Somayyeh, Nejadansari, Dariush, Hesabi, Akbar (2020). Reliability of Human Translations' Scores Using Automated Translation Quality Evaluation Understudy Metrics. *Journal of Foreign Language Research*, 10 (3), 618-629.
DOI: 10.22059/jflr.2020.309025.751

* سومیه کرمی، دانشجوی دکتری رشته ترجمه در دانشگاه اصفهان است. وی به کلیه حوزه‌های مربوط به فن‌آوری در ترجمه و ارزیابی کیفیت آن علاقمند است.

** داریوش نژادانصاری، دکترای تخصصی آموزش زبان انگلیسی و عضو هیئت علمی رسمی گروه زبان و ادبیات انگلیسی دانشگاه اصفهان است.

*** اکبر حسابی، دکترای تخصصی زبان‌شناسی و عضو هیئت علمی رسمی گروه زبان و ادبیات انگلیسی دانشگاه اصفهان است.



Reliability of Human Translations' Scores Using Automated Translation Quality Evaluation Understudy Metrics



Somayyeh Karami*

PhD Candidate in Translation, Department of Foreign Languages, University of Isfahan, Isfahan, Iran

Email: s.karami@fgn.ui.ac.ir



Dariush Nejadansari**
(corresponding author)

Assistant Professor in Teaching English Language, Department of English Language and Literature, University of Isfahan, Isfahan, Iran

Email: d.nejadansari@fgn.ui.ac.ir



Akbar Hesabi***

Assistant Professor in General Linguistics, Department of English Language and Literature, University of Isfahan, Isfahan, Iran

Email: a.hesabi@fgn.ui.ac.ir

ABSTRACT

Considering the costly nature of translation quality assessment in terms of time, money and energy, it seems logical to benefit from the modern technologies that are introduced in the field of machine translation (MT). Automated Translation Quality Evaluation Understudy Metrics (ATQEUMs) are one of these technologies that have revealed a promising application in assessing the MT output quality. This study, however, attempts to examine the reliability of the scores provided by the lexical ATQEUMs to human translated texts (i.e. the ones provided by 51 senior students of translator training programs in Iran) using 1, 2, ..., 5 reference translations successively and separately. To this end, an empirical applied study is conducted following a quantitative approach to assess the reliability of the lexical ATQEUMs' scores in comparison to the expert scorers' scores. The higher the correlation between the sets of scores (in different stages of using 1, 2, ..., 5 reference translations), the higher the reliability is interpreted to be. The results of the Pearson correlation coefficient analysis revealed that using 5 reference translations had led to the highest correlations in 37.80% of cases, which is more than the number for any other situation considered (i.e. using 4 reference translations (3.65%), 3 reference translations (10.97%), 2 reference translations (31.70%), and 1 reference translation (15.85%)). However, using 2 reference translations achieved the second position in having the highest correlations which contradicted the hypothesis that more reference translations would lead to higher correlations and reliability.

DOI: 10.22059/jflr.2020.309025.751

© 2020 All rights reserved.

ARTICLE INFO

Article history:

Received:

5th, September, 2020

Accepted:

6th, October, 2020

Available online:

Autumn 2020

Keywords:

Translation quality
assessment, Lexical
automated translation
quality evaluation
understudy metrics,
Automated scoring,
Reliability, Reference
translations

Karami, Somayyeh, Nejadansari, Dariush, Hesabi, Akbar (2020). Reliability of Human Translations' Scores Using Automated Translation Quality Evaluation Understudy Metrics. *Journal of Foreign Language Research*, 10 (3), 618-629.

DOI: 10.22059/jflr.2020.309025.751

* Somayyeh Karami is a PhD. candidate in Translation at the University of Isfahan. She is interested in all areas related to technology in translation and its quality evaluation.

** Dariush Nejadansari is an associate professor in Teaching English as a Foreign Language in the Department of English Language and Literature at the University of Isfahan.

*** Akbar Hesabi is an associate professor in Linguistics in the Department of English Language and Literature at the University of Isfahan.

۱. مقدمه

ارزیابی به عنوان عاملی جدایی‌ناپذیر از آموزش به فرایند گردآوری داده‌ها درباره فراگیر، برای تصمیم‌گیری درباره پیشرفت وی اطلاق می‌شود (چنگ، راجرز، و هو، Cheng, Rogers, & Hu, ۲۰۰۴ در بیک‌محمدی، علوی، و کیوان‌پناه، ۲۰۲۰) که البته در گستره پژوهش‌های ترجمه کمتر به این موضوع پرداخته شده است. علت اصلی در حقیقت ماهیت نظری عمل ارزیابی ترجمه است (بوکر، Bowker, ۲۰۰۱، ص. ۳۴۷)، بنابراین ممکن است یک ترجمه یکسان، در شرایط متفاوت، مناسب و قابل قبول یا بکلی غیر قابل قبول انگاشته شود، حتی اگر توسط یک نفر ارزیابی شده باشد (وندرشلدن، Vanderschelden, ۲۰۰۰، در اولوهان، Olohan, ۲۰۰۴). زیرا کارشناسان، بیشتر بر دانش پیش‌زمینه خود، شواهد روایتی یا نمونه‌ها، لغت‌نامه‌ها، متون همانند و نظرات کارشناسان این گستره تکیه می‌کنند، که آشکارا این منابع، در همه موارد، منجر به ارزیابی عینی و بی‌طرفانه نمی‌شود (بوکر، ۲۰۰۱). در عین حال، همانطور که هاوس (House, ۱۹۹۷) نیز بیان می‌کند، ارزیابی ترجمه، نیاز به دامنه‌ای فراتر از داوری‌های نظری یا متعصبانه فردی دارد، بدین معنا که باید بیشتر به سمت عینیت و پایایی رفت (بوکر، ۲۰۰۱). یکی دیگر از اشکال‌هایی که ارزیابی ترجمه همواره با آن رودرروست، ماهیت پرمصرف آن از لحاظ زمان، انرژی و هزینه است (مارکوئز، Marquez, ۲۰۱۳). کارشناسان می‌باید زمان و انرژی بسیاری را صرف ارزیابی و نمره‌دهی یک‌ترجمه کنند که نتیجه آن از یک‌سو بسیار پرهزینه است و از سوی دیگر، به‌عنوان کاری ارزشمند و موثق مورد قبول و تقدیر واقع نمی‌شود. بنابراین، انجام پژوهش‌های بیشتر به منظور حرکت به سمت عینیت و پایایی لازم و منطقی به نظر می‌رسد.

در این باره می‌توان به امکان ابزارهای خودکار جانشین ارزیابی کیفیت ترجمه اشاره کرد که تا حد بسیاری زمان و انرژی مورد نیاز برای ارزیابی مهارت‌های پیچیده‌ای از جمله ترجمه را کاهش می‌دهند (ویگل، Weigle, ۲۰۱۱). پژوهشگران گستره نمره‌دهی خودکار، در بسیاری از موارد، نمره‌های ابزارهای خودکار به ترجمه‌های ماشینی را با نمره‌های کارشناسان به همان متون و با استفاده از تنها یک ترجمه معیار مقایسه کرده و سرانجام به این نتیجه رسیده‌اند که نمره‌های این

ابزارها، کمینه به اندازه نمره‌های کارشناسان پایا و قابل اطمینانند (از جمله پاپینی و همکاران، Papineni, et al., ۲۰۰۲؛ دادینگتون، Doddington, ۲۰۰۲؛ بنرجی و لیوی، Banerjee & Lavie, ۲۰۰۵). اما در این پژوهش سعی بر آن است که از این ابزار برای نمره‌دهی به ترجمه انسانی بهره گرفته شود. البته با در نظر گرفتن ویژگی خلاقیت در مترجمان به هنگام فرایندهای حل مسئله و تصمیم‌گیری برای انتخاب بهترین معادل ممکن در ترجمه استفاده از یک ترجمه معیار به عنوان مبنا برای تصمیم‌گیری درباره کیفیت ترجمه منطقی به نظر نمی‌رسد. بنابراین، این پژوهش در صدد پاسخ به این پرسش است که با افزایش مرحله به مرحله تعداد ترجمه‌های معیار تا پنج نمونه، با هدف پوشش تمامی معادل‌های ممکن و صحیح، چه تغییری در میزان پایایی نمره‌های این ابزارهای خودکار ایجاد می‌شود. به‌عبارت دیگر، هدف این است که مشاهده شود آیا استفاده از این ابزارهای خودکار و افزایش تعداد ترجمه‌های معیار می‌تواند به راه‌حلی برای مشکلات فرایند ارزیابی ترجمه (ماهیت پرمصرف آن از لحاظ زمان، انرژی و هزینه از یک سو و پایایی و عینیت آن از سوی دیگر) دست یافت و این ابزارها را به عنوان جایگزینی مناسب برای عامل انسانی در فرایند ترجمه معرفی کرد.

۲. پیشینه پژوهش

ابزارهای خودکار جانشین ارزیابی کیفیت ترجمه

ابزارهای خودکار را می‌توان بر اساس بعدی از زبان و متن که روی آن تمرکز دارند (واژگان، ساختار دستوری، و معنا) دسته‌بندی کرد. در پژوهش حاضر، تمرکز بر آن دسته از ابزارهایی است که تنها بر گستره واژگان تمرکز دارند. جدای از روش محاسبه نمره نهایی، تمامی این ابزارها بر اساس قانونی مشترک عمل می‌کنند: محاسبه میزان شباهت میان متون مورد بررسی و متون ترجمه معیار که اصولاً توسط مترجمان حرفه‌ای و کارشناس انجام می‌شوند. در ادامه تمامی ابزارهایی که در این پژوهش مورد استفاده قرار گرفته‌اند، بر اساس روش محاسبه نمره نهایی معرفی می‌شوند.

الف) ابزارهای مبتنی بر فاصله ویرایش

این ابزارها کیفیت ترجمه را بر اساس فاصله لونشتاین (لونشتاین، Levenshtein, ۱۹۶۶) یعنی حداقل تعداد

جایگزینی‌ها، حذفی‌ها و اضافاتی که برای تبدیل متن مورد بررسی به ترجمه معیار ضروری است، محاسبه می‌کند. از انواع مختلف این ابزارها می‌توان به موارد زیر اشاره کرد:

- میزان خطای لغوی (-WER) (نیسن و همکاران، [Nießen, et al. ۲۰۰۰](#))
- میزان خطای لغوی مستقل از موقعیت (-PER) (تیلمان و همکاران، [Tillmann, et al. ۱۹۹۷](#))
- میزان ویرایش ترجمه (-TER) (اسنور و همکاران، [Snover, et al. ۲۰۰۶](#)) که خود دارای ۴ نوع متفاوت است:
 ۱. TER default- (با در نظر گرفتن ریشه کلمات و مترادفها)
 ۲. TERbase- (بدون در نظر گرفتن ریشه کلمات، مترادفها یا صورت‌های گوناگون بیان یک مطلب)
 ۳. TERp- (با در نظر گرفتن ریشه کلمات، مترادفها و صورت‌های گوناگون بیان یک مطلب)
 ۴. TERpA- (با در نظر گرفتن عامل بسندگی)

ب) ابزارهای مبتنی بر دقت

در گستره ترجمه، دقت به نسبت تعداد واژگان مورد قبول در متون مورد بررسی (یعنی کلماتی که در متون ترجمه معیار نیز استفاده شده‌اند) به مجموع کلمات موجود در متن مورد بررسی اشاره دارد. از انواع مختلف این ابزارها می‌توان به موارد زیر اشاره کرد:

- روش خودکار ارزشیابی دوزبانه جایگزین (BLEU) (پاپینی و همکاران، [۲۰۰۲](#))
- NIST (دادینگتون، [۲۰۰۲](#))
- دقت در سطح واژگان (PI) (گونزالس و گیمنز، [González & Giménez ۲۰۱۴](#))

ج) ابزارهای مبتنی بر یادآوری

در گستره ترجمه، یادآوری به نسبت تعداد واژگان مورد قبول در متون مورد بررسی به مجموع واژگان موجود در متن ترجمه معیار اشاره دارد. به عبارت دیگر، این ابزارها مجموع تعداد واژگانی از متن ترجمه معیار که در متن مورد بررسی استفاده شده‌اند را محاسبه می‌کنند. از انواع مختلف این ابزارها می‌توان به موارد زیر اشاره کرد:

- ROUGE (لین و اوچ، [Lin & Och ۲۰۰۴](#)) که

خود دارای هشت نوع مختلف است، از جمله:

۱. ROUGE_n (محاسبه نمره‌ها برای چندین کلمه با طول مختلف از یک تا چهار کلمه)
 ۲. ROUGE_L (محاسبه طولانی‌ترین ترکیب مشترک)
 ۳. ROUGE_s* (حذف تمامی ترکیب‌های دو کلمه‌ای بدون در نظر گرفتن فاصله میان آن‌ها)
 ۴. ROUGE_{SU}* (حذف تمامی ترکیب‌های دو کلمه‌ای بدون در نظر گرفتن فاصله میان آن‌ها، از جمله تک‌واژه‌ها)
 ۵. ROUGE_w* (محاسبه طولانی‌ترین ترکیب وزنی مشترک با ضریب وزنی ۱/۲)
- یادآوری در سطح واژگان (RI) (گونزالس و گیمنز، [۲۰۱۴](#))

د) ابزارهای مبتنی بر معیار F

ابزارهای مبتنی بر معیار F دو معیار دقت و یادآوری را با یکدیگر تلفیق کرده و نمره‌ای که بیانگر میزان کیفیت ترجمه متن مورد بررسی است، محاسبه می‌کنند. از انواع مختلف این ابزارها می‌توان به موارد زیر اشاره کرد:

- تطبیق‌دهنده عمومی متن (GTMe) (ملامد و همکاران، [Melamed, et al. ۲۰۰۳](#)) (محاسبه نمره‌ها برای چندین کلمه با طول مختلف از یک تا سه کلمه)
- METEOR (بنجری و لیوی، [۲۰۰۵](#)) که خود دارای چهار نوع متفاوت است، از جمله:
 ۱. METEOR_{ex} (مطابقت دقیق بین کلمات و عبارات)
 ۲. METEOR_{st} (مطابقت دقیق بین کلمات و عبارات همراه با در نظر گرفتن ریشه کلمات)
 ۳. METEOR_{sy} (مطابقت دقیق بین کلمات و عبارات همراه با در نظر گرفتن مترادفها)
 ۴. METEOR_{pa} (مطابقت دقیق بین کلمات و عبارات همراه با در نظر گرفتن صورت‌های مختلف بیان یک عبارت)
- دقت و یادآوری در سطح واژگان (FI) (گونزالس و گیمنز، [۲۰۱۴](#))
- هم‌پوشانی در سطح واژگان (OI) (گونزالس و

چارچوب نظری پژوهش

کارشناسی وارد دانشگاه می‌شوند بی‌تجربه‌اند و پس از گذراندن واحدهای مقدماتی به افرادی تازه کار در حوزه ترجمه تبدیل می‌شوند. آن‌ها با گذراندن واحدهای درسی بیشتر به سطح مبتدی می‌رسند و در نهایت با دانش‌آموختگی در دوره کارشناسی به فردی ماهر در ترجمه تبدیل می‌شوند (۲۰۰۰، ص. ۵۹). اما از نظر وی، کارشناسی تنها در سایه تجربه عملی در این گستره و پس از سال‌ها تلاش و پشتکار حاصل می‌شود. در حقیقت هدف از این پژوهش، پاسخ به این سؤال است که چه تفاوتی در نتیجه افزایش تعداد ترجمه‌های معیار در میزان پایایی نمرات ابزارهای خودکار ایجاد می‌شود و آیا این نمرات معیار مناسبی برای تمایز میان افراد کارشناس و افراد کم‌تجربه‌تر از یکدیگرند یا خیر.

۳. روش پژوهش

ماهیت پژوهش حاضر از نوع تجربی و کاربردی است که روی مسائل عملی تمرکز دارد و از نتایج آن می‌توان در زندگی واقعی بهره برد (سالدانها و اوبرین، Saldanha & O'Brien، ۲۰۱۴). در حقیقت، در این پژوهش سعی بر این است که تأثیرات خواسته یا ناخواسته استفاده از یک تا پنج ترجمه معیار، هنگام بکارگیری ابزارهای خودکار برای ارزیابی کیفیت ترجمه انسانی در سطح واژگان مورد بررسی قرار گیرد.

شرکت‌کنندگان در پژوهش

سه گروه متفاوت در فرایند انجام این پژوهش مشارکت داشتند، از جمله:

الف) چهار مترجم رسمی با حداقل ده سال سابقه در گستره ترجمه متون حقوقی. از این گروه از شرکت‌کنندگان خواسته شده بود که متون ارائه شده را از فارسی به انگلیسی ترجمه کنند تا به عنوان ترجمه معیار در این پژوهش استفاده شوند.

ب) پنج کارشناس ارزیابی کیفیت ترجمه با کمینه ده سال تجربه در زمینه ترجمه و همچنین آموزش و ارزیابی آن. از این گروه از شرکت‌کنندگان خواسته شده بود که متون ترجمه شده را به دقت بر اساس شیوه تعیین شده توسط نویسندگان مقاله حاضر (مقایسه متون ترجمه شده با متن ترجمه معیار از

چارچوب نظری پژوهش حاضر مبتنی بر سطوح مختلف کارشناسی بر اساس اصطلاحات رایج در انجمن‌های مهارت‌آموزی است که از نظریه هافمن (Hoffman، ۱۹۹۸) در هافمن و همکاران، Hoffman, et al. (۲۰۱۴) گرفته شده است. تخصص به ویژگی‌ها، مهارت‌ها، و دانشی اشاره دارد که کارشناسان را از افراد تازه‌کار و کم‌تجربه متمایز می‌کند. یکی از رویکردهای پژوهش در زمینه کارشناسی، مطالعه ویژگی‌ها و مهارت‌های کارشناسان در مقایسه با افراد تازه‌کار است (چی، Chi، ۲۰۰۶). در این رویکرد نسبی فرض بر این است که کارشناسی یکی از سطوح مختلف مهارت است که افراد تازه‌کار نیز می‌توانند به آن دست یابند.

جدول ۱ سطوح مختلف تخصص بر اساس اصطلاحات رایج در انجمن‌های مهارت‌آموزی (برگرفته از هافمن، ۱۹۹۸ در هافمن و همکاران، ۲۰۱۴)

| | |
|----------|--|
| بی‌تجربه | فردی که هیچ تجربه‌ای در گستره مورد نظر ندارد. |
| تازه‌کار | فردی که به‌تازگی وارد یک گستره شده است و هنوز نمی‌توان وی را عضوی ثابت در آن گستره در نظر گرفت. |
| مبتدی | فردی که در آغاز راه است و آموزش‌های مقدماتی را آغاز کرده است. |
| کارآموز | فردی که در حال یادگیری است و فراتر از سطح مقدماتی پیش رفته است. به طور معمول، کارآموزان در کنار فردی در یکی از سطوح بالاتر آموزش می‌بینند و زمان این دوره از یک تا دوازده سال در گستره‌های مختلف متفاوت است. |
| ماهر | فردی که به سطحی از تجربه و توانایی رسیده است که می‌تواند کارها را به تنهایی و بدون نظارت فردی دیگر، البته در یک مجموعه مدیریت شده که راهکارهای لازم ابلاغ می‌شود، انجام دهد. |
| متخصص | فردی دارای مهارت‌های برجسته که مورد احترام همکاران خویش است و نظریاتش به طور خاصی صحیح و موثق است و کارکردش نشان‌دهنده مهارت شاخص وی در آن گستره خاص است و می‌تواند به‌راحتی مسائل و مشکلات جدید را برطرف کند. |
| استاد | به طور معمول، استاد فرد، ماهر یا کارشناسی است که شایستگی لازم برای آموزش در سطوح پایین‌تر را نیز دارد. استاد یکی از کارشناسان برجسته در گستره خویش است که نظراتش قوانین، معیارها و ایده‌آل‌ها را شکل می‌دهد. |

با استفاده از طبقه‌بندی سطوح مختلف تخصص هافمن (۱۹۹۸ در هافمن و همکاران، ۲۰۱۴) کیرالی (Kiryaly) نیز تلاش کرده است تا آن‌ها را در گستره مطالعات ترجمه تعریف کند بدین شیوه که «دانشجویانی که در دوره

پیش تعیین شده) ارزیابی کرده و نمره دهند.

ج) پنجاه و یک نفر از دانشجویان ترم شش یا بالاتر رشته مترجمی زبان انگلیسی در دانشگاه‌های ایران. از این گروه از شرکت‌کنندگان خواسته شده بود که متون ارائه شده را از فارسی به انگلیسی ترجمه کنند. آن‌ها هیچ گونه پیش شرطی برای شرکت در این پژوهش نداشتند و تنها می‌بایست متون فوق را در قالب تکلیف کلاسی درس ترجمه اسناد و مدارک در مدت حداکثر دو هفته ترجمه می‌کردند.

ابزار مورد استفاده در پژوهش

در فرایند انجام این پژوهش، به منظور گردآوری داده‌های مورد نیاز، ابزار زیر مورد استفاده قرار گرفته‌اند:

الف) آزمون ترجمه مشتمل بر دو متن متفاوت (متن الف و ب). متن (الف) شامل سه ماده از قانون مدنی جمهوری اسلامی ایران (ماده‌های شماره ۹۷۶، ۹۷۷ و ۹۷۸) و متن (ب) شامل شش ماده از آیین‌نامه دادرسی کیفری جمهوری اسلامی ایران (ماده‌های شماره ۶۲۶، ۶۲۷، ۶۲۸، ۶۲۹، ۶۳۰ و ۶۳۱) است که در مجموع مشتمل بر ۶۷۴ کلمه بودند. بایسته یادآوری است که متون فوق، از میان متون مورد استفاده در بخش تخصصی آزمون مترجمان رسمی قوه قضائیه جمهوری اسلامی ایران که در سال ۱۳۸۴ برگزار شد، گرفته شده است.

ب) مجموعه ابزارهای خودکار آسیه (Asiya, Repository of Automated Metrics)

مجموعه آسیه با دسترسی آزاد و رایگان در <http://asiya.lsi.upc.edu> امکان بهرمندی از مجموعه‌ای از ابزارهای خودکار مبتنی بر ترجمه معیار و همچنین ابزارهای تخمین کیفیت ترجمه را برای همگان فراهم می‌کند (گونزالس و گیمنز، ۲۰۱۴). آسیه مجموعه‌ای غنی از ابزارهای ارزیابی کیفیت ترجمه بر اساس معیارهای مشابهت در سطوح مختلفی از متن است و امکان تلفیق نتایج ابزارها و همچنین سازوکارهایی برای تعیین مجموعه‌ای بهینه از ابزارهای موجود را فراهم می‌کند (گونزالس و گیمنز، ۲۰۱۴). آسیه، اساساً

برای ارزیابی کیفیت خروجی ترجمه ماشینی طراحی شده است، اما پژوهندگان این پژوهش سعی کرده‌اند که از آن برای ارزیابی کیفیت ترجمه انسانی نیز استفاده کنند.

فرایند انجام پژوهش

ابتدا از پنجاه و یک نفر از دانشجویان شرکت‌کننده در این پژوهش خواسته شد تا متون مورد نظر را به انگلیسی ترجمه کنند. از آنجا که فرایند نمره‌دهی متون ترجمه شده، تنها عامل مورد بررسی در پژوهش حاضر است، هیچ پیش شرطی برای انجام این ترجمه تعیین نشد. از آنجا که یک نمونه ترجمه رسمی از این متون، به زبان انگلیسی، در دسترس همگان بود، از چهار مترجم رسمی دیگر نیز خواسته شد که همان متون را به انگلیسی ترجمه کرده و در اختیار پژوهشگران قرار دهند تا از آن‌ها به‌عنوان ترجمه معیار استفاده کنند.

سپس از پنج کارشناس ارزیابی کیفیت ترجمه خواسته شد تا ترجمه‌های متون (الف) و (ب) را جداگانه و بر اساس ترجمه معیار که پژوهشگران در اختیار آنان گذاشتند، تصحیح کرده و در مقیاس ۰ تا ۲۰ نمره دهند. این روش مقایسه متون ترجمه شده با متن ترجمه صحیح که از ابتدا در اختیار مصحح‌ها قرار داده می‌شود، به دقت همان روشی است که در نمره‌دهی متون ترجمه شده توسط ابزارهای خودکار جانشین ارزیابی کیفیت ترجمه در سطح واژگان و همچنین در نمره‌دهی آزمون مترجمان رسمی قوه قضائیه استفاده می‌شود. در مرحله بعد، پژوهندگان با بکارگیری ابزارهای خودکار، متون ترجمه شده را در حالت‌های مختلف استفاده از ۱، ۲، ... تا ۵ ترجمه معیار به صورت مرحله به مرحله و جداگانه، به منظور گردآوری دومین مجموعه داده مورد نیاز، برای پاسخ به پرسش پژوهش، ارزیابی کرده و نمره دادند. به عبارت دیگر، پژوهندگان، متون ترجمه شده دانشجویان را طی پنج مرحله جداگانه و هر بار با استفاده از ۱، ۲، ... تا ۵ ترجمه معیار که به ترتیب به ترجمه‌های معیار قبلی افزوده می‌شدند، با استفاده از ابزارهای خودکار ارزیابی کرده و نمره دادند.

روش تحلیل داده

رابطه میان نمرات ابزارهای خودکار و نمرات کارشناسان به ترجمه‌های متون (الف) و (ب) با استفاده از ۱، ۲، ... تا ۵ ترجمه معیار به صورت مرحله به مرحله و جداگانه

با رویکردی کمی و از طریق ضریب پیوستگی پیرسون با فاصله اطمینان ۹۵ درصد با استفاده از روش بازنمونه‌گیری بوت استرپ محاسبه شده است. مجموعه ابزارهای خودکار جانشین متفاوت (مبتنی بر فاصله ویرایش، دقت، یادآوری، و معیار F) برای هر دو متن (الف) و (ب) به طور جداگانه بر اساس حالت‌های مختلف استفاده از ۱، ۲، ... تا ۵ ترجمه معیار به کار گرفته شدند، تا تفاوت میزان پایایی نمره‌های ارائه شده از طریق ابزارهای خودکار، در حالت‌های مختلف مورد بررسی قرار دهند.

۴. نتایج و بحث و بررسی

ابزارهای خودکار مبتنی بر فاصله ویرایش

بررسی رابطه میان نمرات ابزارهای خودکار مبتنی بر فاصله ویرایش و کارشناسان در باره متن (الف)، نشان می‌دهد که فرضیه پژوهندگان، مبنی بر اینکه افزایش تعداد ترجمه‌های معیار موجب بالارفتن میزان رابطه همبستگی و در نتیجه پایایی بیشتر نمره‌ها می‌شود، در این مورد خاص نقض شده است. چرا که در ۶۶/۶۶ درصد موارد استفاده از دو ترجمه معیار [-WER, -PER, -TER_{base}, -TER_P] و در ۳۳/۳۳ درصد موارد استفاده از یک ترجمه معیار [-TER, -TER_P] منجر به بالاترین میزان همبستگی شده‌اند. جالب اینکه استفاده از چهار [-TER, -TER_P, -TER_{PA}] و پنج ترجمه معیار [-WER, -PER, -TER_{base}] هر یک در ۵۰ درصد موارد منجر به پایین‌ترین میزان همبستگی شده‌اند.

جدول ۲ رابطه همبستگی پیرسون میان نمره‌های ابزار خودکار مبتنی بر فاصله ویرایش و کارشناسان (متن الف)

| | ۱ ترجمه معیار | ۲ ترجمه معیار | ۳ ترجمه معیار | ۴ ترجمه معیار | ۵ ترجمه معیار |
|----------------------|---------------|---------------|---------------|---------------|---------------|
| -WER | ۰/۷۷۵۶ | ۰/۷۸۱۴ | ۰/۷۸۱۱ | ۰/۷۰۴۳ | ۰/۶۹۴۷ |
| -PER | ۰/۸۷۵۹ | ۰/۸۸۸۳ | ۰/۸۸۱۷ | ۰/۸۰۹۲ | ۰/۸۰۳۱ |
| -TER | ۰/۸۱۷۲ | ۰/۸۱۶۹ | ۰/۸۱۵۹ | ۰/۶۳۸۳ | ۰/۶۳۸۴ |
| -TER _{base} | ۰/۸۰۸۹ | ۰/۸۱۱۶ | ۰/۸۱۰۶ | ۰/۶۳۱۰ | ۰/۶۳۰۲ |
| -TER _P | ۰/۸۳۹۱ | ۰/۸۳۵۸ | ۰/۸۳۴۰ | ۰/۶۵۴۰ | ۰/۶۶۲۶ |
| -TER _{PA} | ۰/۸۹۶۹ | ۰/۹۰۱۱ | ۰/۸۹۸۶ | ۰/۷۹۲۷ | ۰/۸۰۶۲ |

به همان نسبت، در مورد متن (ب) نیز در ۵۰ درصد موارد استفاده از دو ترجمه معیار [-TER, -TER_{base}, -TER_P] منجر به بالاترین میزان همبستگی شده است، در

حالی که استفاده از سه [-WER]، چهار [-PER] و پنج ترجمه معیار [-TER_{PA}] هر یک در ۱۶/۶۶ درصد موارد منجر به بالاترین میزان همبستگی شده‌اند. در عین حال، استفاده از یک [-WER, -PER, -TER_{PA}] و پنج ترجمه معیار [-TER, -TER_{base}, -TER_P] هر یک در ۵۰ درصد موارد منجر به پایین‌ترین میزان همبستگی شده‌اند.

جدول ۳ رابطه همبستگی پیرسون میان نمره‌های ابزار خودکار مبتنی بر فاصله ویرایش و کارشناسان (متن ب)

| | ۱ ترجمه معیار | ۲ ترجمه معیار | ۳ ترجمه معیار | ۴ ترجمه معیار | ۵ ترجمه معیار |
|----------------------|---------------|---------------|---------------|---------------|---------------|
| -WER | ۰/۸۳۸۶ | ۰/۸۵۳۶ | ۰/۸۵۳۷ | ۰/۸۵۳۱ | ۰/۸۵۲۱ |
| -PER | ۰/۸۹۱۸ | ۰/۹۰۳۱ | ۰/۹۰۵۸ | ۰/۹۰۶۰ | ۰/۹۰۵۷ |
| -TER | ۰/۸۵۴۳ | ۰/۸۷۱۶ | ۰/۸۵۷۷ | ۰/۸۵۷۹ | ۰/۸۵۰۵ |
| -TER _{base} | ۰/۸۵۶۲ | ۰/۸۷۴۳ | ۰/۸۵۹۲ | ۰/۸۵۹۸ | ۰/۸۵۲۸ |
| -TER _P | ۰/۸۸۶۷ | ۰/۹۰۳۲ | ۰/۸۹۲۴ | ۰/۸۹۳۳ | ۰/۸۸۶۵ |
| -TER _{PA} | ۰/۸۷۲۲ | ۰/۹۱۸۷ | ۰/۹۲۳۲ | ۰/۹۲۵۳ | ۰/۹۲۶۷ |

ابزارهای خودکار مبتنی بر دقت

نتایج بررسی رابطه میان نمرات ابزار خودکار مبتنی بر دقت و کارشناسان، نشان می‌دهد که در متن (الف) در ۵۲/۹۴ درصد موارد استفاده از دو ترجمه معیار [BLEU-1, BLEU-2, BLEU-3, BLEU-4, NIST-3, NIST-4, NIST-5, NISTi-4, PI] و در ۵/۸۸ درصد موارد استفاده از سه ترجمه معیار [NISTi-3] منجر به بالاترین میزان همبستگی شده‌اند. این در حالی است که استفاده از پنج ترجمه معیار در ۴۱/۱۷ درصد موارد [BLEUi-2, BLEUi-3, BLEUi-4, NIST-1, NIST-2, NISTi-2, NISTi-5] منجر به بالاترین میزان همبستگی شده است که این نتایج تا حدی فرضیه پژوهش را تایید می‌کند. اما استفاده از چهار ترجمه معیار در ۷۶/۴۷ درصد موارد [BLEU-1, BLEU-2, BLEU-3, BLEUi-2, BLEUi-3, NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, PI] منجر به پایین‌ترین میزان همبستگی شده است. اما در ۱۷/۶۴ درصد موارد استفاده از یک ترجمه معیار [NISTi-3] استفاده از سه ترجمه معیار [BLEUi-4] منجر به پایین‌ترین میزان همبستگی شده‌اند.

جدول ۴ رابطه همبستگی پیرسون میان نمره‌های ابزار خودکار مبتنی بر دقت و کارشناسان (متن الف)

| | | | | | |
|--------|--------|--------|--------|--------|---------|
| ۰/۸۵۹۷ | ۰/۸۵۴۶ | ۰/۸۵۲۸ | ۰/۸۲۲۰ | ۰/۷۷۱۱ | BLEU-4 |
| ۰/۸۳۵۱ | ۰/۸۳۴۷ | ۰/۸۳۲۵ | ۰/۸۱۹۰ | ۰/۷۷۰۷ | BLEUi-2 |
| ۰/۷۹۰۷ | ۰/۷۹۰۱ | ۰/۷۸۸۰ | ۰/۷۷۰۶ | ۰/۷۲۸۷ | BLEUi-3 |
| ۰/۷۵۳۸ | ۰/۷۵۲۸ | ۰/۷۵۳۰ | ۰/۷۴۱۶ | ۰/۷۰۳۰ | BLEUi-4 |
| ۰/۹۴۱۸ | ۰/۹۴۰۴ | ۰/۹۴۴۲ | ۰/۹۴۸۴ | ۰/۹۱۳۷ | NIST-1 |
| ۰/۹۴۸۵ | ۰/۹۴۶۴ | ۰/۹۴۷۵ | ۰/۹۳۹۳ | ۰/۹۰۱۰ | NIST-2 |
| ۰/۹۴۸۲ | ۰/۹۴۶۳ | ۰/۹۴۶۴ | ۰/۹۳۶۵ | ۰/۸۹۸۷ | NIST-3 |
| ۰/۹۴۷۷ | ۰/۹۴۵۷ | ۰/۹۴۵۹ | ۰/۹۳۵۲ | ۰/۸۹۷۷ | NIST-4 |
| ۰/۹۴۷۰ | ۰/۹۴۴۹ | ۰/۹۴۵۵ | ۰/۹۳۴۶ | ۰/۸۹۷۲ | NIST-5 |
| ۰/۸۹۷۰ | ۰/۸۹۷۱ | ۰/۸۹۵۸ | ۰/۸۶۸۶ | ۰/۸۱۱۷ | NISTi-2 |
| ۰/۸۳۵۸ | ۰/۸۴۲۲ | ۰/۸۴۶۵ | ۰/۸۱۵۹ | ۰/۷۶۱۶ | NISTi-3 |
| ۰/۸۱۱۵ | ۰/۸۱۲۴ | ۰/۸۲۲۳ | ۰/۷۸۷۵ | ۰/۷۴۱۶ | NISTi-4 |
| ۰/۷۷۶۱ | ۰/۷۷۷۰ | ۰/۷۹۳۲ | ۰/۷۶۱۵ | ۰/۷۰۷۹ | NISTi-5 |
| ۰/۶۲۷۳ | ۰/۶۳۱۲ | ۰/۶۳۴۷ | ۰/۷۲۰۳ | ۰/۷۱۱۲ | PI |

ابزارهای خودکار مبتنی بر یادآوری

بررسی رابطه میان نمره ابزارهای خودکار مبتنی بر یادآوری و کارشناسان درباره متن (الف) نشان می‌دهد که در ۸۸/۸۸ درصد موارد استفاده از پنج ترجمه معیار -ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-S*, ROUGE-SU*, ROUGE-W منجر به بالاترین میزان همبستگی شده است که این خود فرضیه پژوهش را تایید می‌کند. درحالی‌که استفاده از دو ترجمه معیار تنها در ۱۱/۱۱ درصد موارد [RI] به بالاترین میزان همبستگی منجر شده است. اما در ۸۸/۸۸ درصد موارد استفاده از چهار ترجمه معیار [ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-S*, ROUGE-SU*, ROUGE-W] و در ۱۱/۱۱ درصد موارد استفاده از پنج ترجمه معیار [RI] منجر به پایین‌ترین میزان همبستگی شده‌اند.

جدول ۶ رابطه همبستگی پیرسون میان نمره‌های ابزار خودکار مبتنی بر یادآوری و کارشناسان (متن الف)

| | | | | | |
|---------------|---------------|---------------|---------------|---------------|-----------|
| ۵ ترجمه معیار | ۴ ترجمه معیار | ۳ ترجمه معیار | ۲ ترجمه معیار | ۱ ترجمه معیار | |
| ۰/۸۵۰۵ | ۰/۸۳۹۹ | ۰/۸۴۶۱ | ۰/۸۴۴۲ | ۰/۸۴۷۴ | ROUGE-1 |
| ۰/۷۸۸۱ | ۰/۷۶۹۵ | ۰/۷۷۳۵ | ۰/۷۷۵۵ | ۰/۷۷۴۳ | ROUGE-2 |
| ۰/۷۳۷۱ | ۰/۷۰۹۵ | ۰/۷۱۲۲ | ۰/۷۱۸۰ | ۰/۷۱۴۳ | ROUGE-3 |
| ۰/۷۰۹۸ | ۰/۶۸۷۶ | ۰/۶۹۰۰ | ۰/۶۹۵۹ | ۰/۶۹۳۱ | ROUGE-4 |
| ۰/۸۲۹۶ | ۰/۸۱۴۴ | ۰/۸۲۰۰ | ۰/۸۱۹۲ | ۰/۸۲۲۰ | ROUGE-L |
| ۰/۷۹۵۸ | ۰/۷۸۲۳ | ۰/۷۸۶۵ | ۰/۷۸۷۴ | ۰/۷۸۸۰ | ROUGE-S* |
| ۰/۸۱۱۵ | ۰/۷۹۵۴ | ۰/۷۹۷۵ | ۰/۷۹۹۶ | ۰/۸۰۱۹ | ROUGE-SU* |
| ۰/۸۰۴۶ | ۰/۷۸۶۶ | ۰/۷۸۹۲ | ۰/۷۹۱۰ | ۰/۷۹۳۱ | ROUGE-W |
| ۰/۷۸۰۳ | ۰/۷۸۰۷ | ۰/۸۴۸۵ | ۰/۸۵۳۵ | ۰/۸۴۶۵ | RI |

| | | | | | |
|---------------|---------------|---------------|---------------|---------------|---------|
| ۵ ترجمه معیار | ۴ ترجمه معیار | ۳ ترجمه معیار | ۲ ترجمه معیار | ۱ ترجمه معیار | |
| ۰/۸۶۳۸ | ۰/۸۳۸۱ | ۰/۹۰۴۶ | ۰/۹۰۹۰ | ۰/۹۰۸۲ | BLEU-1 |
| ۰/۸۱۷۸ | ۰/۷۸۶۹ | ۰/۸۴۹۲ | ۰/۸۵۳۹ | ۰/۸۴۷۵ | BLEU-2 |
| ۰/۷۸۵۰ | ۰/۷۵۶۲ | ۰/۸۰۷۸ | ۰/۸۱۰۹ | ۰/۸۰۳۰ | BLEU-3 |
| ۰/۷۶۱۴ | ۰/۷۳۶۱ | ۰/۷۷۵۵ | ۰/۷۷۹۹ | ۰/۷۷۳۲ | BLEU-4 |
| ۰/۷۱۸۰ | ۰/۶۹۰۷ | ۰/۶۹۷۷ | ۰/۷۰۰۵ | ۰/۶۹۴۵ | BLEUi-2 |
| ۰/۷۰۵۸ | ۰/۶۸۳۱ | ۰/۶۸۳۸ | ۰/۶۸۸۹ | ۰/۶۸۶۴ | BLEUi-3 |
| ۰/۶۹۵۱ | ۰/۶۷۸۰ | ۰/۶۷۳۰ | ۰/۶۷۸۶ | ۰/۶۷۸۷ | BLEUi-4 |
| ۰/۹۱۸۹ | ۰/۹۰۷۴ | ۰/۹۱۴۴ | ۰/۹۱۶۹ | ۰/۹۱۴۵ | NIST-1 |
| ۰/۹۰۷۱ | ۰/۸۸۹۹ | ۰/۹۰۳۴ | ۰/۹۰۷۰ | ۰/۹۰۳۳ | NIST-2 |
| ۰/۸۹۹۹ | ۰/۸۸۴۰ | ۰/۸۹۹۵ | ۰/۹۰۳۳ | ۰/۸۹۹۱ | NIST-3 |
| ۰/۸۹۶۹ | ۰/۸۸۱۴ | ۰/۸۹۷۸ | ۰/۹۰۱۶ | ۰/۸۹۷۷ | NIST-4 |
| ۰/۸۹۵۳ | ۰/۸۷۸۵ | ۰/۸۹۶۱ | ۰/۸۹۹۷ | ۰/۸۹۶۳ | NIST-5 |
| ۰/۸۲۶۵ | ۰/۷۹۶۶ | ۰/۸۲۱۳ | ۰/۸۲۴۳ | ۰/۸۱۷۵ | NISTi-2 |
| ۰/۷۲۸۹ | ۰/۷۲۴۰ | ۰/۷۵۴۶ | ۰/۷۵۳۸ | ۰/۷۱۶۱ | NISTi-3 |
| ۰/۷۰۵۷ | ۰/۶۹۰۸ | ۰/۷۲۱۵ | ۰/۷۲۲۰ | ۰/۶۷۷۳ | NISTi-4 |
| ۰/۶۹۳۶ | ۰/۶۴۳۹ | ۰/۶۵۰۰ | ۰/۶۵۲۴ | ۰/۶۴۳۶ | NISTi-5 |
| ۰/۶۲۰۹ | ۰/۶۲۰۸ | ۰/۶۷۷۳ | ۰/۶۷۹۹ | ۰/۶۳۳۶ | PI |

در باره متن (ب) نیز در ۵۸/۸۲ درصد موارد استفاده از دو ترجمه معیار [BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, BLEUi-4, NIST-2, NIST-3, NIST-4, NIST-5] و در ۲۳/۵۲ درصد موارد استفاده از سه ترجمه معیار [BLEU-1, NISTi-3, NISTi-4, NISTi-5] منجر به بالاترین میزان همبستگی شده‌اند. اما در ۱۱/۷۶ درصد موارد استفاده از دو ترجمه معیار [NIST-1, PI] و در ۵/۸۸ درصد موارد استفاده از چهار ترجمه معیار [NISTi-2] منجر به بالاترین میزان همبستگی شده‌اند. افزون بر این، استفاده از یک ترجمه معیار در ۹۴/۱۱ درصد موارد [BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, BLEUi-4, NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, NISTi-3, NISTi-4, NISTi-5] و استفاده از پنج ترجمه معیار تنها در ۵/۸۸ درصد موارد [PI] منجر به پایین‌ترین میزان همبستگی شده‌اند.

جدول ۵ رابطه همبستگی پیرسون میان نمره‌های ابزار خودکار مبتنی بر دقت و کارشناسان (متن ب)

| | | | | | |
|---------------|---------------|---------------|---------------|---------------|--------|
| ۵ ترجمه معیار | ۴ ترجمه معیار | ۳ ترجمه معیار | ۲ ترجمه معیار | ۱ ترجمه معیار | |
| ۰/۹۴۴۶ | ۰/۹۴۳۹ | ۰/۹۴۶۶ | ۰/۹۴۵۳ | ۰/۹۰۵۳ | BLEU-1 |
| ۰/۹۳۹۶ | ۰/۹۳۵۱ | ۰/۹۳۳۱ | ۰/۹۰۶۶ | ۰/۸۵۰۳ | BLEU-2 |
| ۰/۸۹۹۶ | ۰/۸۹۵۶ | ۰/۸۹۱۵ | ۰/۸۵۹۸ | ۰/۸۰۴۵ | BLEU-3 |

در مورد متن (ب) نیز در ۵۵/۵۵ درصد موارد [ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W], S*, ROUGE-W] استفاده از فقط یک ترجمه معیار منجر به بالاترین میزان همبستگی شده است. اما استفاده از دو [ROUGE-SU*, RI] و پنج ترجمه معیار, ROUGE-3] ROUGE-4] هر یک در ۲۲/۲۲ درصد موارد منجر به بالاترین میزان همبستگی شده‌اند. این در حالی است که استفاده از پنج ترجمه معیار در ۶۶/۶۶ درصد موارد [ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W], S*, ROUGE-SU*, ROUGE-W] استفاده از یک ترجمه معیار [ROUGE-3, ROUGE-4, RI] در ۳۳/۳۳ درصد موارد منجر به پایین‌ترین میزان همبستگی شده‌اند.

جدول ۷ رابطه همبستگی پیرسون میان نمره‌های ابزار خودکار مبتنی بر یادآوری و کارشناسان (متن ب)

| | ۱ ترجمه معیار | ۲ ترجمه معیار | ۳ ترجمه معیار | ۴ ترجمه معیار | ۵ ترجمه معیار |
|-----------|---------------|---------------|---------------|---------------|---------------|
| ROUGE-1 | ۰/۸۷۵۳ | ۰/۸۷۱۳ | ۰/۸۴۰۷ | ۰/۸۴۴۸ | ۰/۸۳۹۴ |
| ROUGE-2 | ۰/۸۲۲۰ | ۰/۸۱۱۶ | ۰/۷۸۲۳ | ۰/۷۷۴۷ | ۰/۷۶۱۳ |
| ROUGE-3 | ۰/۷۶۶۳ | ۰/۸۲۳۴ | ۰/۸۶۱۱ | ۰/۸۵۸۷ | ۰/۸۶۵۱ |
| ROUGE-4 | ۰/۷۲۷۷ | ۰/۷۷۳۱ | ۰/۸۰۳۰ | ۰/۸۰۴۳ | ۰/۸۱۲۱ |
| ROUGE-L | ۰/۸۷۱۷ | ۰/۸۷۱۱ | ۰/۸۴۴۹ | ۰/۸۴۷۲ | ۰/۸۴۱۶ |
| ROUGE-S* | ۰/۸۲۴۶ | ۰/۸۱۴۲ | ۰/۷۸۵۵ | ۰/۷۸۳۵ | ۰/۷۶۸۳ |
| ROUGE-SU* | ۰/۸۵۰۹ | ۰/۸۵۶۶ | ۰/۸۳۲۵ | ۰/۸۲۹۹ | ۰/۸۲۱۷ |
| ROUGE-W | ۰/۸۳۲۹ | ۰/۸۲۵۶ | ۰/۸۰۴۱ | ۰/۷۹۶۸ | ۰/۷۸۷۸ |
| RI | ۰/۸۷۲۵ | ۰/۸۸۵۱ | ۰/۸۷۷۷ | ۰/۸۷۹۰ | ۰/۸۸۰۴ |

ابزارهای خودکار مبتنی بر معیار F

بررسی رابطه میان ابزارهای خودکار مبتنی بر معیار F و کارشناسان در زمینه متن (الف) نشان می‌دهد که در ۶۶/۶۶ درصد موارد استفاده از فقط یک ترجمه معیار [GTM-2, GTM-3, METEOR-ex, METEOR-st, METEOR-sy, METEOR-pa] و در ۲۲/۲۲ درصد موارد استفاده از دو ترجمه معیار [FI, OI] منجر به بالاترین میزان همبستگی شده‌اند. افزون بر این، استفاده از پنج ترجمه معیار تنها در ۱۱/۱۱ درصد موارد [GTM-1] منجر به بالاترین میزان همبستگی شده است که این خود با فرضیه پژوهش تناقض دارد. در حالی که استفاده از پنج ترجمه معیار در ۴۴/۴۴ درصد موارد [METEOR-ex, METEOR-st, METEOR-sy, METEOR-pa] و استفاده از سه [GTM-2, GTM-3] و چهار ترجمه معیار [FI, OI] هر یک در ۲۲/۲۲ درصد موارد

منجر به پایین‌ترین میزان همبستگی شده‌اند. افزون بر این، استفاده از یک ترجمه معیار تنها در ۱۱/۱۱ درصد موارد [GTM-1] منجر به پایین‌ترین میزان همبستگی شده است.

جدول ۸ رابطه همبستگی پیرسون میان نمره‌های ابزار خودکار مبتنی بر معیار F و کارشناسان (متن الف)

| | ۱ ترجمه معیار | ۲ ترجمه معیار | ۳ ترجمه معیار | ۴ ترجمه معیار | ۵ ترجمه معیار |
|-----------|---------------|---------------|---------------|---------------|---------------|
| GTM-1 | ۰/۸۷۶۸ | ۰/۸۸۸۳ | ۰/۸۸۳۴ | ۰/۸۸۳۴ | ۰/۹۱۱۶ |
| GTM-2 | ۰/۷۱۶۱ | ۰/۷۰۷۲ | ۰/۶۹۳۱ | ۰/۷۰۲۵ | ۰/۷۱۳۴ |
| GTM-3 | ۰/۶۹۵۹ | ۰/۶۸۷۳ | ۰/۶۷۱۴ | ۰/۶۸۵۱ | ۰/۶۹۴۴ |
| METEOR-ex | ۰/۸۴۵۶ | ۰/۸۳۴۷ | ۰/۸۰۱۱ | ۰/۷۵۶۴ | ۰/۷۳۹۹ |
| METEOR-st | ۰/۸۴۵۸ | ۰/۸۳۶۴ | ۰/۸۰۲۷ | ۰/۷۵۵۴ | ۰/۷۴۱۰ |
| METEOR-sy | ۰/۸۵۱۵ | ۰/۸۴۱۹ | ۰/۸۰۶۶ | ۰/۷۶۰۳ | ۰/۷۴۵۷ |
| METEOR-pa | ۰/۸۶۶۶ | ۰/۸۵۶۲ | ۰/۸۱۶۴ | ۰/۷۷۰۷ | ۰/۷۵۴۰ |
| FI | ۰/۸۲۸۹ | ۰/۸۳۹۶ | ۰/۸۳۶۳ | ۰/۷۴۴۸۵ | ۰/۷۴۴۸۸ |
| OI | ۰/۷۹۴۳ | ۰/۸۰۵۳ | ۰/۷۹۹۱ | ۰/۷۱۸۴ | ۰/۷۲۲۴ |

در خصوص متن (ب) نیز استفاده از دو [METEOR-GTM-1, GTM-3, pa, FI, OI] و سه ترجمه معیار [GTM-1, GTM-3, METEOR-ex] هر یک در ۳۳/۳۳ درصد موارد منجر به بالاترین میزان همبستگی شده‌اند. در حالی که استفاده از پنج ترجمه معیار تنها در ۲۲/۲۲ درصد موارد [GTM-2, METEOR-st] و استفاده از چهار ترجمه معیار در ۱۱/۱۱ درصد موارد [METEOR-sy] منجر به بالاترین میزان همبستگی شده‌اند. جالب آن‌که استفاده از یک ترجمه معیار در ۱۰۰ درصد موارد منجر به پایین‌ترین میزان همبستگی شده است.

جدول ۹ رابطه همبستگی پیرسون میان نمره‌های ابزار خودکار مبتنی بر معیار F و کارشناسان (متن ب)

| | 1 ترجمه معیار | ۲ ترجمه معیار | ۳ ترجمه معیار | ۴ ترجمه معیار | ۵ ترجمه معیار |
|-----------|---------------|---------------|---------------|---------------|---------------|
| GTM-1 | ۰/۸۸۷۰ | ۰/۹۴۸۷ | ۰/۹۵۳۷ | ۰/۹۵۰۴ | ۰/۹۵۱۷ |
| GTM-2 | ۰/۷۴۲۰ | ۰/۷۷۶۱ | ۰/۷۸۳۴ | ۰/۷۸۲۱ | ۰/۷۸۴۲ |
| GTM-3 | ۰/۷۲۴۲ | ۰/۷۵۴۴ | ۰/۷۶۱۱ | ۰/۷۵۸۷ | ۰/۷۵۹۹ |
| METEOR-ex | ۰/۷۹۹۴ | ۰/۸۱۸۹ | ۰/۸۲۶۲ | ۰/۸۲۶۰ | ۰/۸۲۵۹ |
| METEOR-st | ۰/۸۰۳۹ | ۰/۸۲۳۰ | ۰/۸۳۰۸ | ۰/۸۳۰۷ | ۰/۸۳۱۲ |
| METEOR-sy | ۰/۸۱۹۰ | ۰/۸۳۸۴ | ۰/۸۴۲۵ | ۰/۸۴۳۵ | ۰/۸۴۳۲ |
| METEOR-pa | ۰/۸۲۹۸ | ۰/۸۵۲۰ | ۰/۸۵۰۵ | ۰/۸۵۰۷ | ۰/۸۵۱۶ |
| FI | ۰/۸۶۰۸ | ۰/۸۷۶۴ | ۰/۸۶۶۴ | ۰/۸۶۶۶ | ۰/۸۶۵۱ |
| OI | ۰/۸۰۷۲ | ۰/۸۳۲۰ | ۰/۸۱۲۲ | ۰/۸۱۱۲ | ۰/۸۱۰۱ |

پژوهشگران این مقاله، فرض را بر این گذاشته‌اند که استفاده از ترجمه‌های معیار بیشتر حین بکارگیری ابزارهای خودکار به عنوان معیاری برای تصمیم‌گیری در مورد کیفیت ترجمه‌های انسانی منجر به نمراتی عینی‌تر و پایاتر در مقایسه با نمره‌های کارشناسان می‌شود. این فرضیه بر اساس این منطق استوار است که مترجمان انسانی به هنگام ترجمه از خلاقیت خود در فرایندهای حل مسئله و تصمیم‌گیری در گزینش معادل استفاده می‌کنند. بنابراین، تصمیم‌گیری در باره کیفیت متون ترجمه شده بر اساس تنها یک ترجمه معیار که تمامی معادل‌های ممکن به یک اندازه صحیح‌اند، در بر نمی‌گیرد، چندان صحیح و منصفانه به نظر نمی‌رسد. البته افزایش تعداد ترجمه‌های معیار در این پژوهش در تمامی موارد منجر به افزایش میزان همبستگی میان دو مجموعه نمره مورد مطالعه نشده است.

در باره متن (الف)، استفاده از تنها یک ترجمه معیار در ۱۹/۵۱ درصد موارد منجر به بالاترین میزان همبستگی شده است. این عدد، با اضافه شدن ترجمه معیار دوم به ۳۹/۰۲ درصد افزایش پیدا کرده است که این نتیجه در راستای فرضیه پژوهش بوده و آن را تایید می‌کند. اما اضافه شدن ترجمه معیار سوم نتیجه‌ای عکس در پی داشت و استفاده از سه ترجمه معیار تنها در ۲/۴۳ درصد موارد منجر به بالاترین میزان همبستگی شد. اضافه شدن ترجمه معیار چهارم نیز این عدد را به صفر کاهش داد. در عین حال، اضافه شدن پنجمین ترجمه معیار، همه این کاستی‌ها را جبران کرد و استفاده از پنج ترجمه معیار در ۳۹/۰۲ درصد موارد منجر به بالاترین میزان همبستگی شد. به عبارت دیگر، نتیجه حاصل از استفاده از پنج و دو ترجمه معیار با یکدیگر در باره متن (الف) یکسان بود و هر دو در ۳۹/۰۲ درصد موارد به بالاترین میزان همبستگی منتهی شدند.

در باره متن (ب)، استفاده از یک ترجمه معیار در ۱۲/۱۹ درصد موارد منجر به بالاترین میزان همبستگی شد. با اضافه شدن دومین ترجمه معیار این عدد به ۲۴/۳۹ درصد افزایش یافت که این خود هم‌راستا با فرضیه پژوهش بوده و آن را تایید می‌کند. اما اضافه شدن سومین ترجمه معیار، میزان همبستگی را به ۱۹/۵۱ درصد کاهش داد. این سیر کاهشی با

اضافه شدن چهارمین ترجمه معیار نیز ادامه داشت و به ۷/۳۱ درصد رسید. اما، اضافه شدن پنجمین ترجمه معیار؛ این سیر کاهشی میزان همبستگی را جبران کرد و استفاده از پنج ترجمه معیار در ۳۶/۵۸ درصد موارد منجر به بالاترین میزان همبستگی شد.

بایسته یادآوری است که دقت بررسی میزان رابطه همبستگی میان نمره‌ها ۰/۰۰۰۱ است، که این میزان تفاوت بسیار ناچیزتر از آن است که بتوان بر اساس آن منطق استفاده از ترجمه‌های معیار بیشتر را زیر سؤال برد، یا به کلی آن را رد کرد. اما می‌توان این مسئله را با این موضوع نیز توجیه کرد که برخی از ابزارهای خودکار مفاهیمی از قبیل مترادف، بیان به شیوه‌های متفاوت، و ریشه‌های یکسان را پوشش می‌دهند، از جمله TER- و METEOR و انواع آن‌ها. به عبارت دیگر، در نظر گرفتن مفاهیمی از این دست خود نیاز به استفاده از ترجمه‌های معیار بیشتر را تا حدی کاهش می‌دهد و این موضوع خود می‌تواند علت عدم افزایش میزان همبستگی میان نمره‌ها با افزایش تعداد ترجمه‌های معیار باشد.

۵. نتیجه‌گیری

هدف از این پژوهش یافتن پاسخ این سؤال است که افزایش تعداد ترجمه‌های معیار به‌عنوان مبنایی برای سنجش کیفیت ترجمه انسانی به هنگام بکارگیری ابزارهای خودکار جانشین ارزیابی کیفیت ترجمه در سطح واژگان چه تاثیری در میزان پایایی نمره‌های این ابزارها دارد. به منظور ارزیابی میزان پایایی نمره‌های این ابزارها، میزان همبستگی آن‌ها با نمره‌های کارشناسان محاسبه شده است و فرض بر این است که رابطه مستقیم و مثبتی میان میزان همبستگی میان دو مجموعه نمره، مورد بررسی و میزان پایایی نمره‌های ابزار خودکار وجود دارد. به عبارت دیگر، میزان همبستگی بالاتر میان دو مجموعه نمره مورد بررسی، به‌منزله پایایی بیشتر نمره‌های ابزار خودکار تفسیر شده است. در مجموع، نتایج حاصل از ارزیابی کیفیت ترجمه‌های هر دو متن (الف) و (ب) در حالت‌های مختلف استفاده از ۱، ۲، ... تا ۵ ترجمه معیار نشان می‌دهد که استفاده از پنج ترجمه معیار در ۳۷/۸۰ درصد موارد منجر به بالاترین میزان همبستگی شده است که این عدد از هر حالت مورد بررسی دیگر در پژوهش حاضر (چهار ترجمه معیار (۳/۶۵ درصد)، سه ترجمه معیار (۱۰/۹۷

را می‌توان با مسائلی از قبیل کیفیت ترجمه‌های معیار، نوع ترجمه آن‌ها (مثلا آزاد یا لغت به لغت و ...)، و پوشش مفهیمی نظیر مترادف، بیان به عبارتی دیگر، و ریشه یکسان توسط برخی از ابزارهای خودکار توضیح داد. در عین حال، پژوهش‌هایی با دامنه‌ای گسترده‌تر می‌بایست برای یافتن استانداردی در مورد تعداد ترجمه‌های معیار که در نهایت منجر به پایاترین نمره‌های ممکن شوند، انجام شود.

درصد)، دو ترجمه معیار (۳۱/۷۰ درصد) و یک ترجمه معیار (۱۵/۸۵ درصد) بیشتر است. این موضوع خود فرضیه پژوهش را تایید می‌کند، بدان معنا که استفاده از ترجمه‌های معیار بیشتر منجر به پایایی بیشتر نمرات می‌شود وقتی مقایسه میان حالت‌های استفاده از ۱، ۲ و ۵ ترجمه معیار است. اما اضافه شدن سومین و چهارمین ترجمه‌های معیار به روشنی با فرضیه پژوهش در تناقض است و آن را رد می‌کند. البته این موضوع

منابع

- Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.
<https://www.aclweb.org/anthology/W05-0909.pdf>
- Beikmohammadi, Maryam, Alavi, Seyyed-Mohammad, Kaivanpanah, Shiva (2020). Learning-oriented Assessment of Reading: A Mixed Methods Study of Iranian EFL University Instructors' Perceptions and Practices. *Journal of Foreign Language Research*, 10 (2), 316-329.
https://jflr.ut.ac.ir/article_77098_en.html
- Bowker, L. (2001). Towards a methodology for a corpus-based approach to translation evaluation. *Meta: Translators' journal*, 46(2), pp. 345-364.
<https://www.erudit.org/fr/revues/meta/2001-v46-n2-meta159/002135ar.pdf>
- Chi, M. T. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman, *The Cambridge handbook of expertise and expert performance*, (pp. 21-29).
<https://learnlab.org/uploads/mypslc/publications/chi%20two%20approaches%20chapter%202006.pdf>
- Doddington, G. (2002). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology*, (pp. 138-145).
<https://dl.acm.org/doi/10.5555/1289189.1289273>
- González, M., & Giménez, J. (2014). *Asiya: An open toolkit for automatic machine translation (meta-)evaluation*, Technical Manual, Version 3.0. Retrieved from TALP Research Center Project Management.
http://asiya.lsi.upc.edu/Asiya_technical_manual_v3.0.pdf
- Hoffman, R., Ward, P., Feltovich, P. J., Dibello, L., Fiore, S. M., & Andrews, D. H. (2014). *Accelerated expertise, training for high proficiency in a complex world*. New York: Taylor & Francis.
<http://gen.lib.rus.ec/book/index.php?md5=8A76C54D4A31FC7377F3C9E9AB6882C3>
- House, J. (1997). Quality of translation. In: M. Baker, ed., *The Routledge encyclopedia of translation studies* (pp. 197-200). London and New York: Routledge.
<https://www.routledge.com/Routledge-Encyclopedia-of-Translation-Studies/Baker-Saldanha/p/book/9781138933330>
- Kiraly, D. (2000). *A social constructivist approach to translator education, empowerment from theory to practice*. London and New York: St. Jerome Publishing.

<https://www.goodreads.com/book/show/3842881-a-social-constructivist-approach-to-translator-education>

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 8(10), 707–710.

<https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf>

Lin, C.-Y., & Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics.

<https://www.aclweb.org/anthology/P04-1077/>

Màrquez, L. (2013). Automatic evaluation of machine translation quality. *Invited talk at Dialogue 2013*. Bekasovo Resort, Russia: TALP Research Center, Technical University of Catalonia (UPC). <http://ufal.mff.cuni.cz/pbml/94/art-gimenez-marques-evaluation.pdf>

Melamed, I. D., Green, R., & Turian, J. (2003). Precision and recall of machine translation. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics*.

<https://www.aclweb.org/anthology/N03-2021/>

Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). An evaluation tool for machine translation: Fast evaluation for MT research. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

<http://www.lrec-conf.org/proceedings/lrec2000/pdf/278.pdf>

Olohan, M. (2004). *Introducing corpora in translation studies*. New York: Routledge. <https://www.taylorfrancis.com/books/9780203640005>

Papineni, K., Roukos, S., Ward, T., & 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*, (pp. 311-318). Philadelphia. <https://www.aclweb.org/anthology/P02-1040.pdf>

Saldanha, G., & O'Brien, S. (2014). *Research methodologies in translation studies*. London and New York: Routledge, Taylor and Francis Group.

<http://gen.lib.rus.ec/book/index.php?md5=7a1453cac7fc114e796bd75be079006a>

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, (pp. 223–231).

<https://www.cs.umd.edu/~snover/pub/amta06/ter-amta.pdf>

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based search for statistical translation. *Proceedings of European Conference on Speech Communication and Technology*. <https://www-i6.informatik.rwth-aachen.de/publications/download/203/TillmannC.VogelS.NeyH.SawafH.ZubiagaA.--AcceleratedDP-basedSearchforStatisticalTranslation--1997.pdf>

Weigle, S. C. (2011). Validation of automated scores of TOEFL iBT® tasks against nontest indicators of writing ability. *TOEFL iBT® Research Report*. ETS, Georgia State University, Atlanta.

