پژوهش‌های زبان‌شناختی در زبان‌های خارجی
شاپای چاپی: ۴۱۲۳-۲۵۸۸ شاپای الکترونیکی:۷۵۲۱-۲۵۸۸
www.jflr.ut.ac.ir

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

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

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 (۳/۶۵درصد)، 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.

شناسه دیجیتال DOI: 10.22059/jflr.2020.309025.751

کرمی، سمیه، 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

* سمیه کرمی، دانشجوی دکتری رشتهٔ ترجمه در دانشگاه اصفهان است. وی به کلیهٔ 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.

JOURNAL OF FOREIGN LANGUAGE RESEARCH
PRINT ISSN: 2588-4123 ONLINE ISSN: 2588-7521
ww.Jflr.ut.ac.ir

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

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 دربارهٔ فراگیر، برای تصمیم‌گیری دربارهٔ پیشرفت وی اطلاق می‌شود (چنگ، راجرز، و هو، Cheng, Rogers, & Hu، ۲۰۰٤ در بیک‌محمدی، علوی، و کیوان‌پناه، ۲۰۲۰) که البته در گسترهٔ پژوهش‌های ترجمه کمتر به این موضوع پرداخته شده است. علت اصلی در حقیقت ماهیت نظری عمل ارزیابی ترجمه است (بوکر، Bowker، ۲۰۰۱، ص. ۳٤۷)، 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، ۲۰۰٤). زیرا کارشناسان، بیشتر بر دانش پیش‌زمینهٔ خود، شواهد روایتی یا نمونه‌ها، لغت‌نامه‌ها، متون همانند و نظرات کارشناسان این گستره تکیه می‌کنند، که آشکارا این منابع، در همهٔ 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، Màrquez، ۲۰۱۳). کارشناسان می‌باید زمان و انرژی بسیاری را صرف ارزیابی و نمره‌دهی یک‌ترجمه کنند که نتیجه آن از یک‌سو بسیار پرهزینه است و از سویی 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 گسترهٔ 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 اندازهٔ 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 ترجمهٔ انسانی بهره گرفته شود. البته با در نظرگرفتن 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 دربارهٔ کیفیت ترجمه منطقی به نظر نمی‌رسد. بنابراین، این پژوهش در صدد پاسخ به این پرسش است که با افزایش مرحله به مرحلهٔ 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 می‌توان 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 سویی دیگر) دست یافت و این ابزارها را به عنوان جایگزینی مناسب برای عامل انسانی در فرایند ترجمه معرفی کرد.

۲. پیشینهٔ 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 گسترهٔ واژگان تمرکز دارند. جدای از روش محاسبهٔ نمره نهایی، تمامی این ابزارها بر اساس قانونی مشترک عمل می‌کنند: محاسبه میزان شباهت میان متون مورد بررسی و متون ترجمه معیار که اصولا 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 فاصلهٔ ویرایش

این ابزارها کیفیت ترجمه را بر اساس فاصلهٔ لونشتاین (لونشتاین، Levenshtein، ۱۹٦٦) یعنی حداقل تعداد

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۰، شماره ۳، پاییز ۱۳۹۹، از صفحه ۶۱۸ تا ۶۲۹

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 ٤ نوع متفاوت است:

۱. TER default- (با در نظر گرفتن ریشهٔ کلمات و مترادف‌ها)
۲. TERbase- (بدون در نظر گرفتن ریشهٔ کلمات، مترادف‌ها یا صورت‌های گوناگون بیان یک مطلب)
۳. TERp- (با در نظر گرفتن ریشهٔ کلمات، مترادف‌ها و صورت‌های گوناگون بیان یک مطلب)
٤. TERpA- (TERp- با درنظرگرفتن عامل بسندگی)

ب) ابزارهای مبتنی بر دقت

در گسترهٔ ترجمه، دقت به نسبت تعداد واژگان مورد قبول در متون مورد بررسی (یعنی کلماتی که در متون ترجمه معیار نیز استفاده شده‌اند) به‌مجموع 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 (Pl) (گونزالس و گیمنز، Gonzàlez & Giménez، ۲۰۱٤)

ج) ابزارهای مبتنی بر یادآوری

در گسترهٔ 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، ۲۰۰٤) 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)
٤. $ROUGE_{SU}$* (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 (Rl) (گونزالس و گیمنز، ۲۰۱٤)

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:

- تطبیق‌دهندهٔ 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)
٤. $METEOR_{pa}$ (مطابقت دقیق بین کلمات و عبارات همراه با در نظر گرفتن صورت‌های مختلف بیان یک عبارت)

- دقت و یادآوری در سطح واژگان (Fl) (گونزالس و گیمنز، ۲۰۱٤)
- هم‌پوشانی در سطح واژگان (Ol) (گونزالس و

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۰، شماره ۳، پاییز ۱۳۹۹، از صفحه ۶۱۸ تا ۶۲۹

گیمنز، ۲۰۱٤)

چارچوب نظری پژوهش

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.، ۲۰۱٤) 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، مطالعهٔ ویژگی‌ها و مهارت‌های کارشناسان در مقایسه با افراد تازه‌کار است (چی، Chi، ۲۰۰٦). در این رویکرد نسبی فرض بر این است که کارشناسی یکی از سطوح مختلف مهارت است که افراد تازه‌کار نیز می‌توانند به آن دست بیابند.

جدول ۱ سطوح مختلف تخصص بر اساس اصطلاحات رایج در انجمن‌های مهارت‌آموزی (برگرفته از هافمن، ۱۹۹۸ در هافمن و همکاران، ۲۰۱٤)

بی‌تجربه	فردی که هیچ تجربه‌ای در گسترهٔ 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 نشان‌دهندهٔ مهارت شاخص وی در آن گسترهٔ 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 گسترهٔ خویش است که نظراتش قوانین، معیارها و ایده‌آل‌ها را شکل می‌دهد.

با استفاده از طبقه‌بندی سطوح مختلف تخصص هافمن (۱۹۹۸ در هافمن و همکاران، ۲۰۱٤) کیرالی (Kiraly) نیز تلاش کرده است تا آن‌ها را در گسترهٔ مطالعات ترجمه تعریف کند بدین شیوه که «دانشجویانی که در دوره کارشناسی وارد دانشگاه می‌شوند بی‌تجربه‌اند و پس از گذراندن واحدهای مقدماتی به افرادی تازه کار در حوزهٔ ترجمه تبدیل می‌شوند. آن‌ها با گذراندن واحدهای درسی بیشتر به سطح مبتدی می‌رسند و در نهایت با دانش‌آموختگی در دورهٔ کارشناسی به فردی ماهر در ترجمه تبدیل می‌شوند (۲۰۰۰، ص. ۵۹). اما از نظر وی، کارشناسی تنها در سایهٔ 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» (سالدانحا و اوبرین، Saldanha & O'Brien، ۲۰۱٤). در حقیقت، در این پژوهش سعی بر این است که تأثیرات خواسته یا ناخواسته استفاده از یک تا پنج ترجمهٔ معیار، هنگام بکارگیری ابزارهای خودکار برای ارزیابی کیفیت ترجمه انسانی در سطح واژگان مورد بررسی قرار گیرد.

شرکت‌کنندگان در پژوهش

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

الف) چهار مترجم رسمی با حداقل ده سال سابقه در گستره ترجمهٔ 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 زمینهٔ ترجمه و همچنین آموزش و ارزیابی آن. از این گروه از شرکت‌کنندگان خواسته شده بود که متون ترجمه شده را به دقت بر اساس شیوهٔ تعیین شده توسط نویسندگان مقالهٔ حاضر (مقایسه متون ترجمه شده با متن ترجمهٔ معیار از

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۰، شماره ۳، پاییز ۱۳۹۹، از صفحه ۶۱۸ تا ۶۲۹

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 ٦٢٦، ٦٢٧، ٦٢٨، ٦٢٩، ٦٣٠ و ٦٣١) است که در مجموع مشتمل بر ٦٧٤ کلمه بودند. بایستهٔ یادآوری است که متون فوق، از میان متون مورد استفاده در بخش تخصصی آزمون مترجمان رسمی قوه قضائیه جمهوری اسلامی ایران که در سال ١٣٨٤ 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، ٢٠١٤). 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، ٢٠١٤). آسیه، اساسا 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 شرکت کننده 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 پیش‌شرطی 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 مقایسهٔ 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 به‌صورت مرحله به‌مرحله و جداگانه، به منظور گردآوری دومین مجموعه داده مورد نیاز، برای پاسخ به پرسش پژوهش، ارزیابی کرده و نمره دادند. به‌عبارت 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

رابطهٔ 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 به‌کار گرفته شدند، تا تفاوت میزان پایایی نمره‌های ارائه شده از طریق ابزارهای خودکار، در حالت‌های مختلف مورد بررسی قرار دهند.

٤. نتایج و بحث و بررسی

ابزارهای خودکار مبتنی بر فاصله ویرایش

بررسی رابطه میان نمرات ابزارهای خودکار مبتنی بر فاصله ویرایش و کارشناسان در بارهٔ متن (الف)، نشان می‌دهد که فرضیه پژوهندگان، مبنی بر اینکه افزایش تعداد ترجمه‌های معیار موجب بالارفتن میزان رابطهٔ همبستگی و در نتیجه پایایی بیشتر نمره‌ها می‌شود، در این مورد خاص نقض شده است. چرا که در ٦٦/٦٦ درصد موارد استفاده از دو ترجمه معیار [-WER, -PER, -TER$_{base}$, -TER$_{PA}$] 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 به‌پایین‌ترین میزان همبستگی شده‌اند.

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

	۱ ترجمه معیار	۲ ترجمه معیار	۳ ترجمه معیار	٤ ترجمه معیار	۵ ترجمه معیار
-WER	۰/۷۷۵٦	۰/۷۸۱٤	۰/۷۸۱۱	۰/۷۰٤۳	۰/٦۹٤۷
-PER	۰/۸۷۵۹	۰/۸۸۸۳	۰/۸۸۱۷	۰/۸۰۹۲	۰/۸۰۳۱
-TER	۰/۸۱۷۲	۰/۸۱٦۹	۰/۸۱۵۹	۰/٦۳۸۳	۰/٦۳۸٤
-TER$_{base}$	۰/۸۰۸۹	۰/۸۱۱٦	۰/۸۱۰٦	۰/٦۳۱۰	۰/٦۳۰۲
-TER$_{P}$	۰/۸۳۹۱	۰/۸۳۵۸	۰/۸۳٤۰	۰/٦۵٤۰	۰/٦٦۲٦
-TER$_{PA}$	۰/۸۹٦۹	۰/۹۰۱۱	۰/۸۹۸٦	۰/۷۹۲۷	۰/۸۰٦۲

به‌همان 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 حالی‌که استفاده از سه [-WER]، چهار [-PER] و پنج ترجمه معیار [-TER$_{PA}$] هر یک در ١٦/٦٦ 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 به‌پایین‌ترین میزان همبستگی شده‌اند.

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

	۱ ترجمه معیار	۲ ترجمه معیار	۳ ترجمه معیار	٤ ترجمه معیار	۵ ترجمه معیار
-WER	۰/۸۳۸٦	۰/۸۵۳٦	۰/۸۵۳۷	۰/۸۵۳۱	۰/۸۵۲۱
-PER	۰/۸۹۱۸	۰/۹۰۳۱	۰/۹۰۵۸	۰/۹۰٦۰	۰/۹۰۵۷
-TER	۰/۸۵٤۳	۰/۸۷۱٦	۰/۸۵۷۷	۰/۸۵۷۹	۰/۸۵۰۵
-TER$_{base}$	۰/۸۵٦۲	۰/۸۷٤۳	۰/۸۵۹۲	۰/۸۵۹۸	۰/۸۵۲۸
-TER$_{P}$	۰/۸۸٦۷	۰/۹۰۳۲	۰/۸۹۲٤	۰/۸۹۳۳	۰/۸۸٦۵
-TER$_{PA}$	۰/۸۷۲۲	۰/۹۱۸۷	۰/۹۲۳۲	۰/۹۲۵۳	۰/۹۲٦۷

ابزارهای خودکار مبتنی بر دقت

نتایج بررسی رابطه میان نمرات ابزار خودکار مبتنی بر دقت و کارشناسان، نشان می‌دهد که در متن (الف) در ۵۲/۹٤ درصد موارد استفاده از دو ترجمه معیار [BLEU-1, BLEU-2, BLEU-3, BLEU-4, NIST-3, NIST-4, NIST-5, NISTi-4, Pl] 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 ٤١/١٧ درصد موارد [BLEUi-2, BLEUi-3, BLEUi-4, NIST-1, NIST-2, NISTi-2, NISTi-5] منجر به‌بالاترین میزان همبستگی شده است که این نتایج تا حدی فرضیه پژوهش را تایید می‌کند. اما استفاده از چهار ترجمه معیار در ٧٦/٤٧ درصد موارد [BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, Pl] منجر به پایین‌ترین میزان همبستگی شده است. اما در ١٧/٦٤ درصد موارد استفاده از یک ترجمه معیار [NISTi-3, NISTi-4, NISTi-5] و تنها در ۵/۸۸ درصد موارد استفاده از سه ترجمه معیار [BLEUi-4] منجر به پایین‌ترین میزان همبستگی شده‌اند.

جدول ٤ رابطهٔ همبستگی پیرسون میان نمره‌های ابزار خودکار مبتنی بر دقت و کارشناسان (متن الف)

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۰، شماره ۳، پاییز ۱۳۹۹، از صفحه ٦١٨ تا ٦٢٩

	۱ ترجمه معیار	۲ ترجمه معیار	۳ ترجمه معیار	٤ ترجمه معیار	۵ ترجمه معیار
BLEU-1	۰/۹۰۸۲	۰/۹۰۹۰	۰/۹۰٤٦	۰/۸۳۸۱	۰/۸٦۳۸
BLEU-2	۰/۸٤۷۵	۰/۸۵۳۹	۰/۸٤۹۲	۰/۷۸٦۹	۰/۸۱۷۸
BLEU-3	۰/۸۰۳۰	۰/۸۱۰۹	۰/۸۰۷۸	۰/۷۵٦۲	۰/۷۸۵۰
BLEU-4	۰/۷۷۳۲	۰/۷۷۹۹	۰/۷۷۵۵	۰/۷۳٦۱	۰/۷٦۱٤
BLEUi-2	۰/٦۹٤۵	۰/۷۰۰۵	۰/٦۹۷۷	۰/٦۹۰۷	۰/۷۱۸۰
BLEUi-3	۰/٦۸٦٤	۰/٦۸۸۹	۰/٦۸۳۸	۰/٦۸۳۱	۰/۷۰۵۸
BLEUi-4	۰/٦۷۸۷	۰/٦۷۸٦	۰/٦۷۳۰	۰/٦۷۸۰	۰/٦۹۵۱
NIST-1	۰/۹۱٤۵	۰/۹۱٦۹	۰/۹۱٤٤	۰/۹۰۷٤	۰/۹۱۸۹
NIST-2	۰/۹۰۳۳	۰/۹۰۷۰	۰/۹۰۳٤	۰/۸۸۹۹	۰/۹۰۷۱
NIST-3	۰/۸۹۹۱	۰/۹۰۳۳	۰/۸۹۹۵	۰/۸۸٤۰	۰/۸۹۹۹
NIST-4	۰/۸۹۷۷	۰/۹۰۱٦	۰/۸۹۷۸	۰/۸۸۱٤	۰/۸۹٦۹
NIST-5	۰/۸۹٦۳	۰/۸۹۹۷	۰/۸۹٦۱	۰/۸۷۸۵	۰/۸۹۵۳
NISTi-2	۰/۸۱۷۵	۰/۸۲٤۳	۰/۸۲۱۳	۰/۷۹٦٦	۰/۸۲٦۵
NISTi-3	۰/۷۱٦۱	۰/۷۵۳۸	۰/۷۵٤٦	۰/۷۲٤۰	۰/۷۲۸۹
NISTi-4	۰/٦۷۷۳	۰/۷۲۲۰	۰/۷۲۱۵	۰/٦۹۰۸	۰/۷۰۵۷
NISTi-5	۰/٦٤۳٦	۰/٦۵۲٤	۰/٦۵۰۰	۰/٦٤۳۹	۰/٦۹۳٦
Pl	۰/٦۳۳٦	۰/٦۷۹۹	۰/٦۷۷۳	۰/٦۲۰۸	۰/٦۲۰۹

در بارهٔ 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 ۱۱/۷٦ درصد موارد استفاده از دو ترجمه معیار [NIST-1, Pl] 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 ۹٤/۱۱ 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 [Pl] منجر به پایین‌ترین میزان همبستگی شده‌اند.

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

	1ترجمه معیار	۲ ترجمه معیار	۳ ترجمه معیار	٤ ترجمه معیار	۵ ترجمه معیار
BLEU-1	۰/۹۰۵۳	۰/۹٤۵۳	۰/۹٤٦٦	۰/۹٤۳۹	۰/۹٤٤٦
BLEU-2	۰/۸۵۰۳	۰/۹۰٦٦	۰/۹۳۳۱	۰/۹۳۵۱	۰/۹۳۹٦
BLEU-3	۰/۸۰٤۵	۰/۸۵۹۸	۰/۸۹۱۵	۰/۸۹۵٦	۰/۸۹۹٦
BLEU-4	۰/۷۷۱۱	۰/۸۲۲۰	۰/۸۵۲۸	۰/۸۵٤٦	۰/۸۵۹۷
BLEUi-2	۰/۷۷۰۷	۰/۸۱۹۰	۰/۸۳۲۵	۰/۸۳٤۷	۰/۸۳۵۱
BLEUi-3	۰/۷۲۸۷	۰/۷۷۰٦	۰/۷۸۸۰	۰/۷۹۰۱	۰/۷۹۰۷
BLEUi-4	۰/۷۰۳۰	۰/۷٤۱٦	۰/۷۵۳۰	۰/۷۵۲۸	۰/۷۵۳۸
NIST-1	۰/۹۱۳۷	۰/۹٤۸٤	۰/۹٤٤۲	۰/۹٤۰٤	۰/۹٤۱۸
NIST-2	۰/۹۰۱۰	۰/۹۳۹۳	۰/۹٤۷۵	۰/۹٤٦٤	۰/۹٤۸۵
NIST-3	۰/۸۹۸۷	۰/۹۳٦۵	۰/۹٤٦٤	۰/۹٤٦۳	۰/۹٤۸۲
NIST-4	۰/۸۹۷۷	۰/۹۳۵۲	۰/۹٤۵۹	۰/۹٤۵۷	۰/۹٤۷۷
NIST-5	۰/۸۹۷۲	۰/۹۳٤٦	۰/۹٤۵۵	۰/۹٤٤۹	۰/۹٤۷۰
NISTi-2	۰/۸۱۱۷	۰/۸٦۸٦	۰/۸۹۵۸	۰/۸۹۷۱	۰/۸۹۷۰
NISTi-3	۰/۷٦۱٦	۰/۸۱۵۹	۰/۸٤٦۵	۰/۸٤۲۲	۰/۸۳۵۸
NISTi-4	۰/۷٤۱٦	۰/۷۸۷۵	۰/۸۲۲۳	۰/۸۱۲٤	۰/۸۱۱۵
NISTi-5	۰/۷۰۷۹	۰/۷٦۱۵	۰/۷۹۳۲	۰/۷۷۷۰	۰/۷۷٦۱
Pl	۰/۷۱۱۲	۰/۷۲۰۳	۰/٦۳٤۷	۰/٦۳۱۲	۰/٦۲۷۳

ابزارهای خودکار مبتنی بر یادآوری

بررسی رابطه میان نمره ابزارهای خودکار مبتنی بر یادآوری و کارشناسان دربارهٔ متن (الف) نشان می‌دهد که در ۸۸/۸۸درصد 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 ۱۱/۱۱درصد موارد [Rl] به بالاترین میزان همبستگی منجر شده است. اما در ۸۸/۸۸درصد موارد استفاده از چهار ترجمه معیار [ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-S*, ROUGE-SU*, ROUGE-W] و در ۱۱/۱۱درصد موارد استفاده از پنج ترجمه معیار [Rl] منجر به پایین‌ترین میزان همبستگی شده‌اند.

جدول ٦ رابطه همبستگی پیرسون میان نمرهای ابزار خودکار مبتنی بر یادآوری و کارشناسان (متن الف)

	۱ ترجمه معیار	۲ ترجمه معیار	۳ ترجمه معیار	٤ ترجمه معیار	۵ ترجمه معیار
ROUGE-1	۰/۸٤۷٤	۰/۸٤٤۲	۰/۸٤٦۱	۰/۸۳۹۹	۰/۸۵۰۵
ROUGE-2	۰/۷۷٤۳	۰/۷۷۵۵	۰/۷۷۳۵	۰/۷٦۹۵	۰/۷۸۸۱
ROUGE-3	۰/۷۱٤۳	۰/۷۱۸۰	۰/۷۱۲۲	۰/۷۰۹۵	۰/۷۳۷۱
ROUGE-4	۰/٦۹۳۱	۰/٦۹۵۹	۰/٦۹۰۰	۰/٦۸۷٦	۰/۷۰۹۸
ROUGE-L	۰/۸۲۲۰	۰/۸۱۹۲	۰/۸۲۰۰	۰/۸۱٤٤	۰/۸۲۹٦
ROUGE-S*	۰/۷۸۸۰	۰/۷۸۷٤	۰/۷۸٦۵	۰/۷۸۲۳	۰/۷۹۵۸
ROUGE-SU*	۰/۸۰۱۹	۰/۷۹۹٦	۰/۷۹۷۵	۰/۷۹۵٤	۰/۸۱۱۵
ROUGE-W	۰/۷۹۳۱	۰/۷۹۱۰	۰/۷۸۹۲	۰/۷۸٦٦	۰/۸۰٤٦
Rl	۰/۸٤٦۵	۰/۸۵۳۵	۰/۸٤۸۵	۰/۷۸۰۷	۰/۷۸۰۳

در مورد متن (ب) نیز در ۵۵/۵۵ درصد موارد [ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-S*, ROUGE-W] استفاده از فقط یک ترجمه معیار منجر به بالاترین میزان همبستگی شده است. اما استفاده از دو [ROUGE-SU*, Rl] 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 ٦٦/٦٦ درصد موارد [ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-S*, ROUGE-SU*, ROUGE-W] و استفاده از یک ترجمه معیار [ROUGE-3, ROUGE-4, Rl] 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	٤ ترجمه معیار	٥ ترجمه معیار
ROUGE-1	۰/۸۷۵۳	۰/۸۷۱۳	۰/۸٤۰۷	۰/۸٤٤۸	۰/۸۳۹٤
ROUGE-2	۰/۸۲۲۰	۰/۸۱۱٦	۰/۷۸۲۳	۰/۷۷٤۷	۰/۷٦۱۳
ROUGE-3	۰/۷٦٦۳	۰/۸۲۳٤	۰/۸٦۱۱	۰/۸۵۸۷	۰/۸٦۵۱
ROUGE-4	۰/۷۲۷۷	۰/۷۷۳۱	۰/۸۰۳۰	۰/۸۰٤۳	۰/۸۱۲۱
ROUGE-L	۰/۸۷۱۷	۰/۸۷۱۱	۰/۸٤٤۹	۰/۸٤۷۲	۰/۸٤۱٦
ROUGE-S*	۰/۸۲٤٦	۰/۸۱٤۲	۰/۷۸۵۵	۰/۷۸۳۵	۰/۷٦۸۳
ROUGE-SU*	۰/۸۵۰۹	۰/۸۵٦٦	۰/۸۳۲۵	۰/۸۲۹۹	۰/۸۲۱۷
ROUGE-W	۰/۸۳۲۹	۰/۸۲۵٦	۰/۸۰٤۱	۰/۷۹٦۸	۰/۷۸۷۸
Rl	۰/۸۷۲۵	۰/۸۸۵۱	۰/۸۷۷۷	۰/۸۷۹۰	۰/۸۸۰٤

ابزارهای خودکار مبتنی بر معیار F

بررسی رابطه میان ابزارهای خودکار مبتنی بر معیار F و کارشناسان در زمینهٔ متن (الف) نشان می‌دهد که در ٦٦/٦٦ 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 [Fl, Ol] 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 حالی‌که استفاده از پنج ترجمه معیار در ٤٤/٤٤ درصد موارد [METEOR-ex, METEOR-st, METEOR-sy, METEOR-pa] و استفاده از سه [GTM-2, GTM-3] و چهار ترجمه معیار [Fl , Ol] 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	٤ ترجمه معیار	٥ ترجمه معیار
GTM-1	۰/۸۷٦۸	۰/۸۸۸۳	۰/۸۸۳٤	۰/۸۸۳٤	۰/۹۱۱٦
GTM-2	۰/۷۱٦۱	۰/۷۰۷۲	۰/٦۹۳۱	۰/۷۰۲۵	۰/۷۱۳٤
GTM-3	۰/٦۹۵۹	۰/٦۸۷۳	۰/٦۷۱٤	۰/٦۸۵۱	۰/٦۹٤٤
METEOR-ex	۰/۸٤۵٦	۰/۸۳٤۷	۰/۸۰۱۱	۰/۷۵٦٤	۰/۷۳۹۹
METEOR-st	۰/۸٤۵۸	۰/۸۳٦٤	۰/۸۰۲۷	۰/۷۵۵٤	۰/۷٤۱۰
METEOR-sy	۰/۸۵۱۵	۰/۸٤۱۹	۰/۸۰٦٦	۰/۷٦۰۳	۰/۷٤۵۷
METEOR-pa	۰/۸٦٦٦	۰/۸۵٦۲	۰/۸۱٦٤	۰/۷۷۰۷	۰/۷۵٤۰
Fl	۰/۸۲۸۹	۰/۸۳۹٦	۰/۸۳٦۳	۰/۷٤٤۸۵	۰/۷٤٤۸۸
Ol	۰/۷۹٤۳	۰/۸۰۵۳	۰/۷۹۹۱	۰/۷۱۸٤	۰/۷۲۲٤

در خصوص متن (ب) نیز استفاده از دو [METEOR-pa, Fl, Ol] و سه ترجمه معیار [GTM-1, GTM-3, METEOR-ex] هر یک در ۳۳/۳۳ درصد موارد منجر به بالاترین میزان همبستگی شده‌اند. درحالی‌که استفاده از پنج ترجمه معیار تنها در ۲۲/۲۲درصد 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)

	1ترجمه معیار	۲ ترجمه معیار	۳ ترجمه معیار	٤ ترجمه معیار	٥ ترجمه معیار
GTM-1	۰/۸۸۷۰	۰/۹٤۸۷	۰/۹۵۳۷	۰/۹۵۰٤	۰/۹۵۱۷
GTM-2	۰/۷٤۲۰	۰/۷۷٦۱	۰/۷۸۳٤	۰/۷۸۲۱	۰/۷۸٤۲
GTM-3	۰/۷۲٤۲	۰/۷۵٤٤	۰/۷٦۱۱	۰/۷۵۸۷	۰/۷۵۹۹
METEOR-ex	۰/۷۹۹٤	۰/۸۱۸۹	۰/۸۲٦۲	۰/۸۲٦۰	۰/۸۲۵۹
METEOR-st	۰/۸۰۳۹	۰/۸۲۳۰	۰/۸۳۰۸	۰/۸۳۰۷	۰/۸۳۱۲
METEOR-sy	۰/۸۱۹۰	۰/۸۳۸٤	۰/۸٤۲۵	۰/۸٤۳۵	۰/۸٤۳۲
METEOR-pa	۰/۸۲۹۸	۰/۸۵۲۰	۰/۸۵۰۵	۰/۸۵۰۷	۰/۸۵۱٦
Fl	۰/۸٦۰۸	۰/۸۷٦٤	۰/۸٦٦٤	۰/۸٦٦٦	۰/۸٦۵۱
Ol	۰/۸۰۷۲	۰/۸۳۲۰	۰/۸۱۲۲	۰/۸۱۱۲	۰/۸۱۰۱

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۰، شماره ۳، پاییز ۱۳۹۹، از صفحه ٦۱۸ تا ٦۲۹

بحث و بررسی

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 بارهٔ 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 به‌نظر 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 بارهٔ 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 به۳۹/۰۲ 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 ۲/٤۳ 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 ترجمهٔ معیار، همهٔ 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 ترجمهٔ معیار با یکدیگر در بارهٔ متن (الف) یکسان بود و هر دو در ۳۹/۰۲ درصد موارد به بالاترین میزان همبستگی منتهی شدند.

در بارهٔ 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 به۲٤/۳۹ درصد افزایش یافت که این خود هم‌راستا با فرضیه پژوهش بوده و آن را تایید می‌کند. اما اضافه شدن سومین ترجمهٔ معیار، میزان همبستگی را به۱۹/۵۱ درصد کاهش داد. این سیر کاهشی با اضافه‌شدن چهارمین ترجمه معیار نیز ادامه داشت و به۷/۳۱ درصد رسید. اما، اضافه‌شدن پنجمین ترجمهٔ معیار؛ این سیر کاهشی میزان همبستگی را جبران کرد و استفاده از پنج ترجمه معیار در ۳٦/۵۸ درصد موارد منجر به بالاترین میزان همبستگی شد.

بایستهٔ 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. به‌عبارت 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 نمرهٔ مورد بررسی، به‌منزلهٔ 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 به‌بالاترین میزان همبستگی شده است که این عدد از هر حالت مورد بررسی دیگر در پژوهش حاضر (چهار ترجمه معیار (۳/٦۵ درصد)، سه ترجمه معیار (۱۰/۹۷

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 را می‌توان با مسائلی از قبیل کیفیت ترجمه‌های معیار، نوع ترجمهٔ آن‌ها (مثلا آزاد یا لغت به‌لغت و ...)، و پوشش مفاهیمی 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.

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

Gonzàlez, 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.

https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2011.tb02260.x