Language and Linguistics
Spring & Summer 2023, 18 (35)
DOI: 10.30465/LSI.2023.8499

Original Research
Received: 2022-12-26
Accepted: 2023 - 4 – 16
ISSN: 23223847

Presenting a quality estimation model of English to Farsi machine translator using transfer learning

Jafari Harandi, Mohammad Hossein[1] — Master student of Computer Science, Tehran University, Tehran, Iran
Azadi, Fateme[2] — Ph.D. student of Computer Science, Tehran University, Tehran, Iran
Rafiei, Sepehr[3] — B. A. of Computer Science, Tehran University, Tehran, Iran
Faili, Hesham[4] — Professor of Computer Science, Tehran University, Tehran, Iran
Dousti, Mohammad Javad[5] — Assistant Professor of Computer Science, Tehran University, Tehran, Iran

Abstract

Nowadays the evaluation of machine translation without reference translation is of great importance as one of the research areas of machine translation. One of the challenges in this field, especially for languages with few sources, is the lack of suitable training data. For this purpose, it is possible to use neural network-based methods that have been previously trained on multilingual language models and estimate the translation quality for a pair of new languages using transfer learning. In this article, the quality of an English Persian test set is estimated in this way. Also, a set of educational data for the English Persian language pair has been prepared, and appropriate preprocessing has been done on it, and the existing multilingual model has been fine-tuned with that data. The use of these training data has improved the Pearson correlation with the test set by 29

Keywords: Quality estimation, machine translation, neural network, transfer learning, fine tuning

1. hosein.jafari.h@ut.ac.ir 2. azadi@ut.ac.ir 3. fiei.k@ut.ac.ir
4. hfaili@ut.ac.ir 5. mjdousti@ut.ac.ir

How to cite: jafari harandi, M. H., azadi, F., rafiei, S., faili, H., & dousti, M. J. (2023). Presenting a quality estimation model of English to Farsi machine translator using transfer learning. *Language and Linguistics*, *18*(35), 71-86. doi: 10.30465/lsi.2023.849913:25

زبان و زبان‌شناسی
دوره ۱۸، شماره ۳۵، بهار و تابستان، سال ۱۴۰۱، ۷۱- ۸۶
DOI: 10.30465/LSI.2023.8499

مقاله پژوهشی
تاریخ دریافت: ۱۴۰۱/۱۰/۰۵؛ تاریخ پذیرش: ۱۴۰۲/۰۱/۲۷؛ تاریخ انتشار: بهار ۱۴۰۲
ISSN: 23223847

ارائه یک مدل تخمین کیفیت مترجم ماشینی انگلیسی به فارسی با استفاده از یادگیری انتقالی

جعفری هرندی، محمدحسین — دانشجوی کارشناسی ارشد کامپیوتر دانشگاه تهران، تهران، ایران
آزادی، فاطمه — دانشجوی دکتری، دانشگاه تهران، تهران، ایران
رفیعی، سپهر — کارشناس کامپیوتر، دانشگاه تهران، تهران، ایران
فیلی، هشام — استاد کامپیوتر، دانشگاه تهران، تهران، ایران
دوستی، محمدجواد — استادیار کامپیوتر، دانشگاه تهران، تهران، ایران

چکیده: امروزه، ارزیابی ترجمهٔ ماشینی بدون داشتن ترجمهٔ مرجع، به‌عنوان یکی از حوزه‌های پژوهشی ترجمهٔ ماشینی از اهمیت بالایی برخوردار است. یکی از چالش‌های موجود در این زمینه، مخصوصاً برای زبان‌های کم‌منبع، نبود داده‌های آموزشی مناسب است. برای این منظور می‌توان از روش‌های مبتنی بر شبکهٔ عصبی که قبلاً روی مدل‌های زبانی چندزبانه آموزش‌دیده‌شده استفاده کرده و با استفاده از یادگیری انتقالی کیفیت ترجمه برای یک جفت‌زبان جدید را تخمین زد. در این مقاله کیفیت یک مجموعهٔ آزمایشیِ انگلیسی ـ فارسی به این صورت تخمین زده شده است. همچنین مجموعه‌ای از داده‌های آموزشی برای جفت‌زبان انگلیسی ـ فارسی تهیه شده و مدل چندزبانهٔ موجود با آن تنظیم دقیق شده است. استفاده از این داده‌های آموزشی، همبستگی پیرسون با مجموعهٔ آزمایشی را به میزان ۲۹ درصد بهبود داده است.

کلیدواژه‌ها: تخمین کیفیت[1]، ترجمهٔ ماشینی[2]، شبکه عصبی[3]، یادگیری انتقالی[4]، تنظیم دقیق[5].

1. quality estimation
2. machine translation
3. neural network
4. transfer learning
5. fine-tuning

۱ مقدمه

ترجمهٔ ماشینی امروزه کاربردهای بسیار وسیعی از حوزهٔ مالی گرفته تا پزشکی و علوم انسانی و حتی شبکه‌های اجتماعی پیدا کرده است. با این حال هنوز کیفیت ترجمه در بسیاری از جفت‌زبان‌ها[1] و حوزه‌ها مطلوب نیست. بنابراین داشتن معیارهایی که به‌کمک آن‌ها بتوان ترجمه‌های بد را شناسایی کرد و آن‌ها را برای پساویرایش[2] به مترجمین انسانی سپرد، به یک امر حیاتی تبدیل شده است. این کار باعث می‌شود که در هزینه و وقت نیروی انسانی صرفه‌جویی شود. هدف معیارهای تخمین کیفیت این است که کیفیت ترجمه را بدون داشتن ترجمه‌ها تخمین نتیجه بزند.

هرساله در کارگاه‌های ترجمهٔ ماشینی[3] یکی از وظیفه[4]های موجود تخمین کیفیت است که در آنجا هر سال شرکت‌کنندگان زیادی حاضر می‌شوند و هدف تخمین کیفیت ترجمه‌ها در چند جفت–زبان مختلف و با استفاده از معیارهای متفاوت است که مجموعه‌دادهٔ متناسب با آن‌ها موجود است و هرقدر همبستگی[5] تخمین‌ها به اعداد واقعی نزدیک‌تر باشد، شرکت‌کنندگان امتیاز بیشتری کسب می‌کنند. نتایج آخرین کارگاهی که برگزار شده (زروا[6] و همکاران، ۲۰۲۲) موجود است.

برای دست‌یابی به این معیارها عموماً به دادگان آموزشی شامل جملات مبدأ، خروجی‌های سیستم ترجمهٔ ماشینی و برچسب کیفیت ترجمه برای آن‌ها نیاز است. برچسب کیفیت معمولاً به دو روش آماده می‌شود: یکی ارزیابی مستقیم که در آن جملات مبدأ و ترجمه‌های آن‌ها توسط نیروی انسانی امتیاز داده می‌شوند، و دیگری با استفاده از معیار نرخ ویرایش ترجمه[7](اسنور[8] و همکاران، ۲۰۰۶ :۲۳۱-۲۲۲). برای محاسبهٔ این معیار ابتدا خروجی‌های ترجمهٔ ماشینی توسط تعدادی مترجم انسانی با کمترین تعداد ویرایش، پساویرایش شده و سپس تعداد ویرایش‌های مورد نیاز برای تبدیل خروجی سیستم به پساویرایش مترجم به‌صورت خودکار محاسبه و به‌عنوان معیار کیفیت ترجمه در نظر گرفته می‌شود. در این روش، هرچقدر که یک جمله تعداد ویرایش کمتری برای تبدیل به یک ترجمهٔ درست نیاز داشته باشد، کیفیت بهتری دارد و درنتیجه امتیاز بیشتری می‌گیرد.

با استفاده از این دادگان آموزشی یک شبکهٔ عصبی آموزش داده می‌شود که بتواند

1. pair languages
2. post-editing
3. WMT (workshop on machine translation)
4. task
5. correlation
6. C. Zerva
7. HTER (human translation edit rate)
8. M. Snover

کیفیت هر جفت‌جملهٔ دیگر را تخمین بزنند. یکی از ابزارهای منبع‌بازی[1] که در حال حاضر وجود دارد، ترنسکوئست[2] (راناسینه[3]، اوراسان[4] و میتکوو[5] ،۲۰۲۰ (5070-5081 :است. این ابزار قبلاً روی دادگان آموزشی ۷ جفت‌زبان[6] آموزش‌دیده است و به‌واسطهٔ مدل زبانی چندزبانه‌ای[7] که در معماری[8] آن وجود دارد، می‌توان آن را با استفاده از یادگیری انتقالی برای هر جفت‌زبان دیگری که از بردارهای تعبیهٔ آن‌ها در این ابزار پشتیبانی شده است، به‌کار گرفت.

۲ چارچوب نظری

در طی دههٔ گذشته پیشرفت‌های زیادی در حوزهٔ تخمین کیفیت ترجمه صورت گرفته است که بیشتر آن‌ها حاصل برگزاری کارگاه‌های ترجمهٔ ماشینی[9] است که از سال ۲۰۱۲ برگزار می‌شود. هر سال تعدادی داده شامل جملات زیادی در جفت‌زبان‌های مختلف که توسط انسان‌ها برچسب‌گذاری[10] شده و کیفیت هر جمله با معیارهای مختلف مشخص شده است، در این کارگاه‌ها در اختیار عموم قرار داده می‌شود. برگزاری این کارگاه‌ها موجب شکل‌گیری ابزارهای منبع‌بازِ بسیار خوبی مثل دیپ‌کوئست[11]) ایو، پلین و اسپسیا[12]، ۲۰۱۸(3146-3157 :، اپن‌کیوی[13] (کپلر[14] و همکاران، ۲۰۱۹ :۱۱۷ (و ترنسکوئست به‌وجود بیایند که با دقت خیلی خوبی می‌توانند کیفیت یک ترجمه را در جفت‌زبان‌های بسیار زیادی پیش‌بینی کنند.

قبل از رواج شبکهٔ عصبی، تخمین کیفیت با استفاده از استخراج تعدادی از ویژگی‌های خاص از مبدأ و مقصد ترجمه و به‌کارگیری روش‌هایی همچون درخت تصمیم تصادفی[15] صورت می‌گرفت. اگرچه آن‌ها نتایج خوبی داشتند ولی امروزه به‌کار نمی‌روند و روش‌های جدیدتر به‌سمت شبکه‌های عصبی روی آورده‌اند.

برای مثال، بهترین سیستم در سال ۲۰۱۷، پستک[16] (کیم، لی و نا[17]، ۲۰۱۷562-568): بود

1. open-source 2. TransQuest 3. T.Ranasinghe
4. C. Orăsan 5. R. Mitkov
6. انگلیسی به رومانیایی ـ استونیانی ـ نپالی ـ سینهالا ـ روسی ـ آلمانی و چینی
7. multilingual 8. architecture 9. WMT (Workshop on Machine Translation)
10. labeling 11. Deep-Quest 12. Ive, J, Blain, F., Specia, L
13. Open-Kiwi 14. F. Kepler 15. Random Decision Tree
16. Postech 17. Kim, H., Lee, J.-H., Na, S.-H.

که یک مدل رمزگذار–رمزگشا[1] بازگشتی به‌عنوان پیش‌بینی‌کننده[2] ساخته بود که که با یک مدل بازگشتی دیگر با عنوان تخمین‌گر[3] که وظیفهٔ تولید تخمین‌های کیفیت را به‌عهده داشت، تجمیع شده بود .مدل پیش‌بینی‌کننده کلمات را دریافت می‌کرد و یک بازنمایی محتوامحور از کلمات ارائه می‌داد و این بازنمایی‌ها به‌عنوان ورودی به تخمین‌گر داده می‌شدند تا بتواند کیفیت ترجمه را تخمین بزند .مدل پیش‌بینی‌کننده برای آموزش به مقدار خیلی زیادی دادهٔ موازی احتیاج دارد .معماری پستک بعدها در چارچوب دیپ‌کوئست پیاده‌سازی شد.

اپن‌کیوی یک ابزار منبع‌باز دیگر است که توسط آنبابل[4] در سال ۲۰۱۹ ارائه شد که چهار معماری مبتنی بر شبکه عصبی را پیاده‌سازی کرده است. یکی از این معماری‌ها معماری پیش‌بینی‌کننده–تخمین‌گر[5] پستک است که مشابه معماری است و به دادگان موازی اضافی وابسته است و از بقیهٔ معماری‌ها بهتر کار می‌کند.

برای اینکه این وابستگی از بین برود، می‌توانیم از تعبیه‌های بین زبانی[6] استفاده کنیم که از قبل برای انعکاس ویژگی‌های بین زبان‌ها تنظیم دقیق[7] شده است. در سال‌های اخیر کارهای خوبی در این زمینه انجام شده است. از زمان معرفی برت[8]) دولین[9] و همکاران، ۲۰۱۹: ۴۱۷۱–۴۱۸۶) این مدل به‌صورت موفقی در وظایف مختلف پردازش زبان‌های طبیعی به‌کار گرفته شده است. این مدل به‌صورت چندزبانه نیز ارائه شده است (پیرس، اشلینگر و گرت[10]، ۲۰۱۹5001-4996).

همچنین مدل ایکس‌ال‌ام–آر[11] (کانو[12] و همکاران، ۲۰۲۰: ۸۴۵۱–۸۴۴۰) ارائه شد که روی یک مجموعه‌دادهٔ عظیم چندزبانه آموزش داده شده و نتایج بسیار خوبی هم در وظایف بین زبانی داشته است که در این مقاله از آن استفاده شده است.

۳ روش پژوهش

در این مقاله مجموعه‌دادهٔ انگلیسی–فارسی‌ای که حاصل کار مترجمین فرازین[13] است، تهیه شده است. مترجم فرازین یک ابزار ترجمهٔ انگلیسی–فارسی است که توسط پارک علم و

1. encoder-decoder
2. predictor
3. estimator
4. Unbable (https://developers.unbabel.com/)
5. Predictor-Estimator
6. Cross-lingual Embedding
7. Fine-tune
8. BERT
9. J. Devlin
10. Pires, T., Schlinger, E., Garrette, D.
11. XLM-R
12. A. Conneau
13. https://www.faraazin.ir/

فناوری دانشگاه تهران ساخته شده و به‌روش یادگیری عمیق[1] کار می‌کند (شمس‌فرد، بی‌جن‌خان و احمدی فرد، ۱۴۰۱: ۳۶۰).

شیوهٔ ساخت این مجموعه‌داده به این صورت بوده که ابتدا تعدادی جمله به مترجم ماشینی داده شده و سپس خروجی آن توسط مترجمین انسانی پساویرایش شده است. پس از آن، امتیاز هر جمله توسط معیار نرخ ویرایش ترجمهٔ انسانی[2] محاسبه شده است.

در ابتدا مدل چندزبانهٔ موجود را که درواقع برای جفت‌زبان انگلیسی به هر زبانی (هر زبانی که در مدل تعبیهٔ کلمات پشتیبانی می‌شود) آموزش‌دیده است را به‌عنوان خط پایه[3] برای تخمین کیفیت انگلیسی ـ فارسی روی دادگان آزمایشی ارزیابی کرده و همبستگی پیرسون[4] بین امتیازهای به‌دست‌آمده از مدل و امتیازهای مرجع را محاسبه کرده‌ایم. سپس مدل چندزبانهٔ موجود را با استفاده از دادگان آموزشیِ تهیه‌شده تنظیم دقیق کرده و همبستگی بین این امتیازهای جدید و امتیازهای مرجع را نیز به‌دست آورده‌ایم.

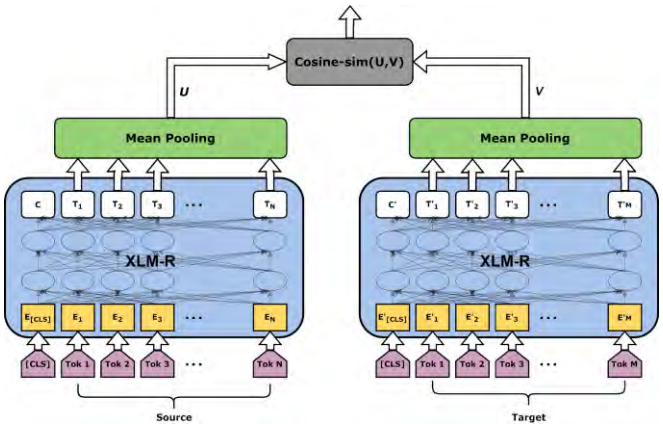
در این بخش ابتدا معماری کلی ترنسکوئست را بررسی می‌کنیم، سپس روش‌های پیشنهادی در این پژوهش را شرح می‌دهیم و در نهایت شیوهٔ ارزیابی روش‌های ارائه‌شده را بیان می‌کنیم.

۱-۳ معماری ترنسکوئست

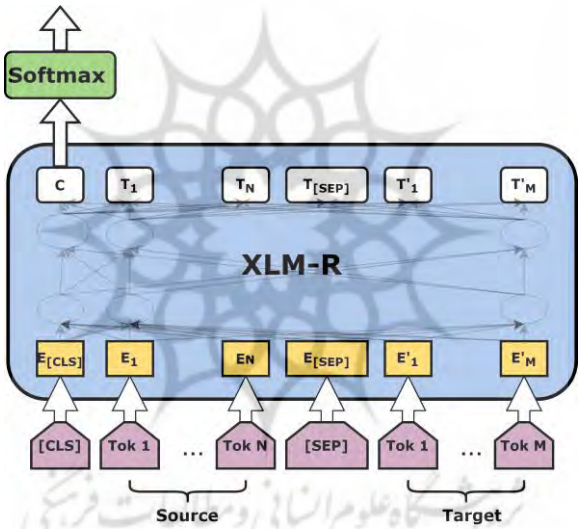
ترنسکوئست یک ابزار منبع‌باز برای تخمین کیفیت ترجمهٔ ماشینی است. این ابزار داری دو معماری مختلف است؛ در معماری اول که در شکل ۱ آمده است، از معماری ایکس‌ال‌ام-آر استفاده شده است که ابتدا با استفاده از یک توکن[5] اس‌ای‌پی[6]، جملهٔ اصلی و ترجمهٔ آن را جدا کرده و به این صورت دو جمله را به‌هم پیوند داده است. سپس ترنسکوئست با استفاده از توان خروجی سی‌ال‌اس[7]، میزان شباهت دو جمله را مدل کرده است.

در معماری دوم که در شکل ۲ آمده است، جملهٔ اصلی و ترجمهٔ آن را به‌طور جداگانه به ایکس‌ال‌ام-آر داده است و سپس با یک لایهٔ ادغام[8] هرکدام را به‌صورت بردار بازنمایی کرده و سپس شباهت کسینوسی[9] این بردارها را محاسبه کرده است.

1. deep learning
2. human translation edit rate (HTER)
3. baseline
4. Pearson correlation
5. token
6. SEP (separator)
7. CLS (class)
8. Pooling
9. Cosine-Similarity



شکل ۱- معماری اول ترنسکوئست (راناسینه، اوراسان و میتکوو، ۲۰۲۰: ۵۰۸۱-۵۰۷۰)



شکل ۲- معماری دوم ترنسکوئست (راناسینه، اوراسان و میتکوو، ۲۰۲۰: ۵۰۸۱-۵۰۷۰)

۳-۲ روش‌های پیشنهادی

در این مقاله سه روش برای تخمین کیفیت ترجمهٔ انگلیسی ـ فارسی ارائه شده است:

در روش اول از یک روش غیرنظارتی[1] برای تخمین کیفیت ترجمه استفاده کردیم به این صورت که فاصلهٔ کسینوسی[2] بردار تعبیهٔ جمله‌ای[3] خروجی ترجمهٔ ماشینی و خروجیِ

1. unsupervised 2. cosine similarity 3. sentence embedding vector

پساویرایش‌شده توسط انسان را به‌عنوان امتیاز کیفیت ترجمه به‌دست می‌آوریم. این بردار تعبیهٔ جمله‌ای از مقالهٔ لابسه[1] (فنگ[2] و همکاران ۲۰۲۲: ۸۹۱–۸۷۸) که توسط گوگل منتشر شده آمده است.

در روش دوم از یادگیری انتقالی برای تخمین کیفیت ترجمه استفاده کرده‌ایم. در این روش از یکی از مدل‌های ازپیش‌آموزش‌دیدهٔ ترنسکوئست (مربوط به معماری اول[3]) که بر روی دادگان آموزشی ۷ جفت‌زبان انگلیسی به رومانیایی، استونیایی، نپالی، سینهالا، روسی، آلمانی و چینی آموزش داده شده است، به همان صورت برای امتیازدهی به جملات مجموعهٔ آزمایشی انگلیسی ـ فارسی استفاده کرده، و یک خط پایه برای تخمین کیفیت ترجمه در زبان‌های انگلیسی ـ فارسی ایجاد کرده‌ایم.

در روش سوم به‌منظور بهبود این خط پایه، با ایجاد دادگان آموزشی برای جفت‌زبان انگلیسی ـ فارسی با استفاده از پساویرایش‌های مترجمین انسانی و معیار نرخ ویرایش ترجمه انسانی، مدل پیش‌آموزش‌دیدهٔ ترنسکوئست را برای جفت‌زبان انگلیسی ـ فارسی تنظیم دقیق کرده‌ایم، به این صورت که آموزش مدل ترنسکوئست را برای چند قدم آموزشی[4] بر روی دادگان ایجادشده، ادامه داده‌ایم.

۳-۳ نحوهٔ ارزیابی

برای ارزیابی روش‌های ارائه‌شده از یک مجموعه آزمایشی با ۸۷۷ جفت‌جمله و امتیاز نرخ ویرایش ترجمهٔ انسانی آن‌ها استفاده شده است. این مجموعه آزمایشی توسط هریک از مدل‌های ارائه‌شده امتیازدهی شده و سپس همبستگی پیرسون و همبستگی اسپیرمن[5] بین امتیازهای محاسبه‌شده و امتیازی که توسط معیار نرخ ویرایش ترجمه انسانی به‌دست آمده است را محاسبه می‌کنیم. فرمول همبستگی پیرسون در رابطه (۱) و فرمول همبستگی اسپیرمن در رابطه (۲) آمده است.

$$r(pearson) = \frac{\sum (x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\sum (x_i - \underline{x})^{\square} \sum (y_i - \underline{y})}} \quad (1)$$

1. LABSE (language agnostic BERT sentence embedding)
2. F. Feng
3. https://huggingface.co/TransQuest/monotransquest-hter-en_any
4. epoch
5. Spearman correlation

$$r(spearman) = \frac{\sum x_i y_i}{\sqrt{\sum (x_i)^2 \sum (y_i)^2}} \qquad (2)$$

در این دو رابطه، رابطهٔ $\overline{x}$ و $\overline{y}$ به‌ترتیب میانگین امتیازات به‌دست‌آمده از مدل و میانگین امتیازات نرخ ویرایش ترجمه انسانی بر روی پیکره آزمایشی بوده و x_i و y_i نیز به‌ترتیب امتیازات مدل و برچسب نرخ ویرایش ترجمهٔ انسانی برای هریک از جملات هستند. درنهایت، همبستگی‌های به‌دست‌آمده با استفاده از دو روش پیشنهادی را با هم مقایسه کرده‌ایم.

۴ بحث و تحلیل

در این بخش ابتدا در مورد مجموعه دادهٔ مورد استفاده صحبت می‌کنیم، سپس نتایج آزمایش‌های مربوط به دو روش پیشنهادی را ارائه می‌دهیم و با هم مقایسه و تحلیل می‌کنیم.

۴-۱ مجموعه داده

مجموعه دادگان موجود تعدادی جفت‌جملهٔ انگلیسی ـ فارسی است که به‌صورت مقابل تهیه شده است: ابتدا تعدادی جملهٔ انگلیسی در دامنه‌های مختلف به مترجم فرازین داده شده و سپس خروجی مترجم توسط تعدادی مترجم انسانی پساویرایش شده تا ترجمهٔ با کیفیت به‌دست آید. سپس فاصلهٔ ویرایشی[1] بین این جفت‌جملات با خروجی‌های مترجم ماشینی با استفاده از روش نرخ ویرایش ترجمهٔ انسانی محاسبه شده است که نشان‌دهندهٔ امتیاز کیفیت این ترجمه است.

سپس این دادگان به سه قسمت آموزشی[2]، ارزیابی[3] و آزمایشی[4] تقسیم شده است. به‌این صورت که ۶۳ درصد دادگان برای آموزش، ۱۵ درصد دادگان برای ارزیابی و ۲۲ درصد دادگان برای آزمایش در نظر گرفته‌شده‌است که نمونه‌ای از این دادگان در جدول ۱ و آمارهای دقیق آن در جدول ۲ آمده است.

1. edit-distance
2. train
3. validation
4. test

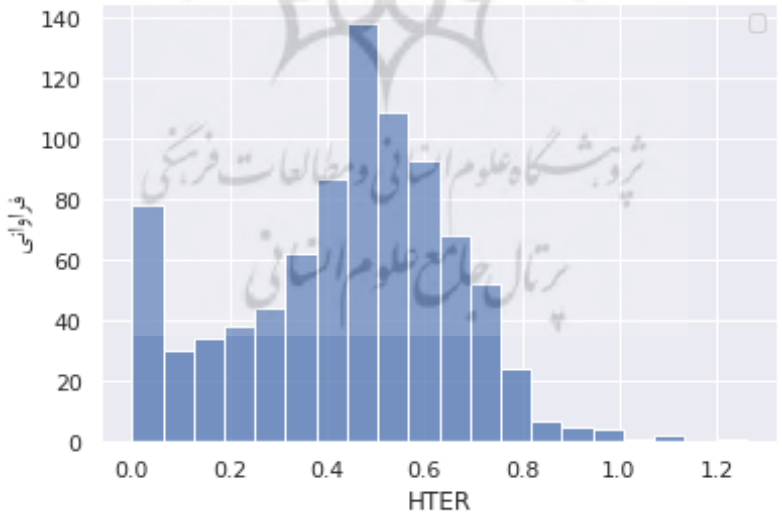
جدول ۱- نمونه‌ای از دادگان آزمایشی

ورودی	In this paper, we propose ComBIM, a community-based solution approach for solving the BIM problem.
خروجی فرازین	در این مقاله، ما ComBIM، یک روش راه‌حل مبتنی بر جامعه برای حل مساله BIM را پیشنهاد می‌کنیم.
خروجی ویرایش شده توسط انسان	در این مقاله، ComBIM، یک روش حل مبتنی بر جامعه برای حل مساله BIM را پیشنهاد می‌کنیم.
امتیاز نرخ ویرایش ترجمه انسانی	77/0

جدول ۲- آمارهای مجموعه داده

	تعداد جفت‌جملات	تعداد کلمات فارسی	تعداد کلمات انگلیسی
آموزشی	2804	77,067	77,267
ارزیابی	701	20,143	19,355
آزمایشی	877	23,646	24,740

در شکل ۳ فراوانی برچسب‌های نرخ ویرایش ترجمه انسانی داده‌های آموزشی، نمایش داده شده است. این نمودار نشان می‌دهد که داده‌های آموزشی تهیه‌شده، حاوی نمونه‌هایی از جملاتی با کیفیت‌های ترجمهٔ متنوع بوده، که می‌تواند به آموزش سیستم تخمین کیفیت کمک کند. چراکه از همه طیف اعداد در آن دیده می‌شود.



شکل ۳- نمودار توزیع امتیازات نرخ ویرایش ترجمه انسانی برای داده‌های آموزشی

۲-۴ آزمایش‌ها و تحلیل نتایج

در روش یادگیری انتقالی ما از مدل چندزبانهٔ ازپیش‌آموزش‌دیده‌شدهٔ ترنسکوئست استفاده کرده و تعدادی جفت‌جملهٔ موجود در دادگان آزمایشی را ارزیابی کرده‌ایم. سپس همبستگی پیرسون بین این ارزیابی‌ها و امتیاز واقعی آن‌ها که حاصل کار مترجم‌های انسانی است را محاسبه کردیم.

در ادامه، برای بهبود مدل موجود آن را که با استفاده از دادگان آموزشی تهیه شده تنظیم دقیق کردیم. برای این کار فرآیند آموزش مدل پیش‌آموزش‌دیدهٔ ترنسکوئست را با پارامترهایی که در جدول ٤ آمده است، ادامه داده و سپس با مدل جدید دادگان آزمایشی را ارزیابی کرده و همبستگی پیرسون و اسپیرمن بین این ارزیابی‌های جدید و امتیاز واقعی آن‌ها را محاسبه کردیم.

نتایج این آزمایش‌ها در جدول ۳ آمده است. مشاهده می‌کنیم که روش دوم نسبت به روش اول ۱۷ درصد بهبود در همبستگی پیرسون و ۱۷ درصد بهبود در همبستگی اسپیرمن را به‌همراه داشته است. با توجه به اینکه روش اول یک روش بدون نظارت بوده و روی همهٔ زبان‌ها قابل اجراست می‌توان نتیجه گرفت که یادگیری انتقالی به‌خوبی کار کرده است.

مقایسهٔ روش سوم و روش دوم نشان می‌دهد که تنظیم دقیق کردن مدل روی دادگان آموزشی تهیه‌شده، همبستگی پیرسون مدل تخمین کیفیت با برچسب‌های نرخ ویرایش ترجمهٔ انسانی را ۲۹ درصد و همبستگی اسپیرمن آن‌ها را ۲۵ درصد بهبود داده است.

جدول ۳- همبستگی پیرسون و اسپیرمن در روش‌های ما

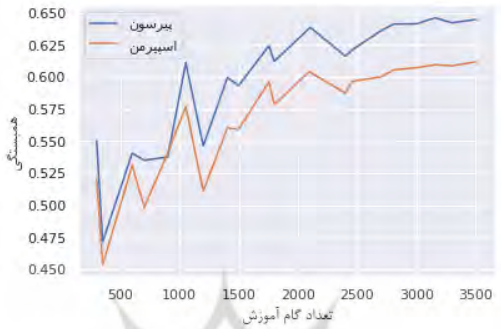
	همبستگی پیرسون	همبستگی اسپیرمن
روش اول (غیر نظارت شده)	**18/0**	**19/0**
روش دوم (یادگیری انتقالی)	**35/0**	**36/0**
روش سوم (تنظیم دقیق شده و پیشنهادی ما)	**64/0**	**61/0**

جدول ٤- پارامتر های مدل

تعداد قدم های آموزشی[1]	۱۰
اندازهٔ دسته آموزشی[2]	۸
نرخ آموزش[3]	۵×۱۰^{-۵}
بیشترین طول دنباله‌ها[4]	۸۰

1. epochs size
2. training batch size
3. learning rate
4. maximum sequence length

همچنین نمودار تعداد گام آموزشی برحسب همبستگی پیرسون و اسپیرمن داده‌های ارزیابی و برچسب‌های نرخ ویرایش ترجمهٔ انسانی، در شکل ۴ آمده است. در این شکل مشاهده می‌شود که در هر گام همبستگی روی داده‌های ارزیابی در حال افزایش است و این نشان می‌دهد که داده‌ها به‌خوبی در مدل جا گرفته‌اند و باعث بهبود مدل شده‌اند.



شکل ۴- نمودار تعداد گام آموزشی برحسب همبستگی پیرسون و اسپیرمن داده‌های ارزیابی و برچسب‌های نرخ ویرایش ترجمه انسانی

۴-۳ تحلیل نتایج

نمونه‌ای از نتایج دادگان آزمایشی روی دو روش ما در جدول ۵ آمده است. امتیاز نرخ ویرایش ترجمهٔ انسانی این نمونه که از خروجی پساویرایش‌شده توسط انسان به‌دست آمده است، ۰/۵۱ است. روش اول ما این امتیاز را ۰/۳۷ تخمین زده است و روش دوم ما ۰/۵ تخمین زده است که به مقدار واقعی نزدیک‌تر است. این نشان می‌دهد که عمل تنظیم دقیقی که توسط دادگان آموزشی ما انجام شده است باعث شده است تا کیفیت مدل افزایش یابد و مدل تخمین دقیق‌تری ارائه دهد.

جدول ۵- نمونه‌ای از نتایج دادگان آزمایشی روی دو روش ما

ورودی	The solution at least in part must be metacities that employ the best new principles of urban planning, the best new technology, and best invoke the spirit of urban pride.
خروجی فرازین	راه‌حل حداقل در بخشی باید metacities باشد که بهترین اصول جدید برنامه‌ریزی شهری، بهترین فناوری جدید را به کار گرفته و به بهترین نحو روح غرور شهری را طلب می‌کند.
خروجی ویرایش شده توسط انسان	راه حل حداقل تا حدودی باید استعاره‌هایی باشد که از بهترین اصول جدید برنامه ریزی شهری، بهترین فناوری جدید استفاده می‌کنند و به بهترین وجه از روح غرور شهری استفاده می‌کنند.
امتیاز نرخ ویرایش ترجمه انسانی	51/0
نتیجه روش دوم	37/0
نتیجه روش سوم	5/0

۴-۴ تحلیل دسته‌بندی‌شدهٔ اصلاحات

ما نمونه‌های مجموعه داده را از این‌جهت که ویرایشگران به چه صورت آن‌ها را اصلاح کردند مورد بررسی قرار دادیم و به این نتیجه رسیدیم که این نوع خطاها بر پنج دسته هستند.

۱. اصلاحات نگارشی: اصلاحاتی که ویرایشگر جملات را ازنظر نگارشی اصلاح کرده است؛ مثلاً نیم‌فاصله را به فاصله تبدیل کرده است یا بالعکس.

۲. اصلاحات صحیح: اصلاحاتی که واقعاً اشتباه هستند و ویرایشگر آن‌ها را اصلاح کرده است.

۳. اصلاحات غلط: ویرایشگر به‌اشتباه متن را اصلاح کرده است یعنی ترجمهٔ ماشین درست بوده. این نوع نمونه‌داده‌ها پرسروصدا هستند.

۴. اصلاحات ساختاری: ویرایشگر ترتیب کلمات را عوض کرده است تا از نظر ساختاری درست شود.

۵. اصلاحات بهبودی: اصلاحاتی که ترجمهٔ ماشین درست بوده ولی ویرایشگر برای اینکه ترجمه روان‌تر شود و بهتر مفهوم را برساند اصلاح کرده است.

نمونه ای از این اصلاحات برای نمونه‌ای از دادگان که در جدول ٦ آمده است، در جدول ۷ یافت می‌شود.

جدول ٦- نمونه‌ای ازدادگان

ورودی	Stakeholders' are used different devices to deliver different medicinal requests (tasks) easily through cloud environment to get a set of medicinal services such as diagnosis of diseases (CKD) as an example of healthcare services.
خروجی فرازین	سهامداران از وسایل مختلفی برای تحویل درخواست‌های دارویی مختلف (وظایف) به راحتی از طریق محیط ابری برای دریافت مجموعه‌ای از خدمات پزشکی مانند تشخیص بیماری‌ها (CKD) به عنوان نمونه از خدمات بهداشتی استفاده می‌کنند.
خروجی ویرایش شده توسط انسان	ذینفعان از دستگاه‌های مختلفی برای ارائه درخواست‌های مختلف دارویی (وظایف) به راحتی از طریق محیط ابر استفاده می‌کنند تا مجموعه‌ای از خدمات دارویی مانند تشخیص بیماری‌ها (CKD) را به عنوان نمونه خدمات درمانی ارائه دهند.

جدول ۷- اصلاحات نمونه‌دادهٔ موجود در جدول ٦

ترجمه شده	ویرایش شده	معادل انگلیسی	نوع اصلاح
سهامدارن	ذینفعان	Stakeholders	بهبودی
وسایل	دستگاه ها	devices	صحیح
تحویل	ارائه	deliver	بهبودی
مختلف دارویی	دارویی مختلف	different medicinal	ساختاری
ابری	ابر	cloud	غلط
پزشکی	دارویی	medicinal	بهبودی
بهداشتی	درمانی	healthcare	غلط
استفاده می‌کنند.	ارائه دهند.	get	صحیح

ما بعد از این‌که حدود ۳۰ نمونه از دادگان را بررسی کردیم، به این نتیجه رسیدیم که مدل ما (روش سوم) وقتی که اصلاحات ویرایشی زیاد باشند به‌خوبی نمی‌تواند امتیاز نرخ ویرایش ترجمهٔ انسانی را پیش‌بینی کند. نمونه‌ای از این دادگان در جدول ۸ آمده است.

همچنین به این نتیجه رسیدیم که وقتی تعداد اصلاحات صحیح و بهبودی زیاد باشند به‌خوبی می‌تواند خطا را پیش‌بینی کند. نمونه‌ای از این دادگان در جدول ۹ آمده است.

جدول ۸- نمونه‌ای از نتایج دادگان آزمایشی که اختلاف زیادی بین امتیاز نرخ ویرایش ترجمهٔ انسانی و نتیجهٔ روش سوم وجود دارد

ورودی	Thomas Friedman, "The Mean Season," New York Times, September 9, 1999, A23.
خروجی فرازین	توماس فریدمن، "فصل میانگین"، نیویورک‌تایمز، ۹ سپتامبر ۱۹۹۹، ۲۳A.
خروجی ویرایش شده توسط انسان	توماس فریدمن، «فصل متوسط»، نیویورک تایمز، ۹ سپتامبر ۱۹۹۹، ۲۳.۶A.
امتیاز نرخ ویرایش ترجمه انسانی	۰/۷۸
نتیجه روش سوم	۰/٤۹
اختلاف بین امتیاز نرخ ویرایش ترجمه انسانی و نتیجه روش سوم	۰/۲۹

جدول ۹- نمونه‌ای از نتایج دادگان آزمایشی که اختلاف کمی بین امتیاز نرخ ویرایش ترجمهٔ انسانی و نتیجهٔ روش سوم وجود دارد

ورودی	When truly inefficient buildings that are over 100 stories high are authorized and built when safety and security are taken into consideration there should be a smart analysis completed.
خروجی فرازین	هنگامی که ساختمان‌های واقعاً ناکارآمد بیش از ۱۰۰ طبقه داشته باشند، مجاز و ساخته می‌شوند زمانی که امنیت و امنیت در نظر گرفته می‌شوند باید یک آنالیز هوشمند تکمیل شود.
خروجی ویرایش شده توسط انسان	هنگامی که ساختمان‌های واقعاً ناکارآمد که بیش از ۱۰۰ طبقه ارتفاع دارند مجاز و ساخته شده اند وقتی امنیت مورد توجه قرار می‌گیرد باید یک تحلیل هوشمند کامل شود.
امتیاز نرخ ویرایش ترجمه انسانی	۰/۵۶
نتیجه روش سوم	۰/۵۷
اختلاف بین امتیاز نرخ ویرایش ترجمه انسانی و نتیجه روش سوم	۰/۰۱

۵ نتیجه‌گیری

در این مقاله، ما ابتدا از مدل پیش‌آموزش‌دیدهٔ چندزبانهٔ ترنسکوئست استفاده کرده و همبستگی امتیازات نتیجهٔ کیفیت این مدل روی دادگان آزمایشی خود را با امتیازهای معیار نرخ ویرایش ترجمهٔ انسانی اندازه‌گیری کردیم، که به‌عنوان یک سیستم خط پایه برای نتیجهٔ کیفیت ترجمهٔ انگلیسی- فارسی قابل استفاده است. سپس این مدل را با استفاده از دادگان آموزشی تهیه‌شده در این مقاله تنظیم دقیق کرده و نشان دادیم که همبستگی ۲۹ درصد بهبود پیدا کرده است. با توجه به این بهبود می‌توان گفت که دادگان آموزشی تهیه‌شده از کیفیت مناسبی برای بهبود مدل نتیجه کیفیت ترجمه برخوردار است و می‌تواند در پژوهش‌های آتی مورد استفاده قرار گیرد.

نوآوری‌های موجود در مقاله به‌شرح زیر است:

۱) ایجاد یک خط پایه برای تخمین کیفیت انگلیسی ـ فارسی

۲) بهبود این خط پایه با استفاده از داده‌های آموزشیِ تهیه‌شده

منابع

شمس‌فرد مهنوش، محمود بی‌جن‌خان و مهدی احمدی (۱۴۰۱). پردازش متن و گفتار فارسی: مروری بر مبانی نظری و آخرین یافته‌های پژوهشی. تهران: سمت.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. "Unsupervised Cross-lingual Representation Learning" at Scale. Presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451. arXiv preprint arXiv:1911.02116.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186. 2019. arXiv preprint arXiv:1810.04805.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W., 2022. "Language-agnostic BERT Sentence Embedding". Presented at the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 878–891. arXiv preprint arXiv:2007.01852.

Ive, J., Blain, F., Specia, L., 2018. "DeepQuest: a framework for neural-based quality estimation". Presented at the Proceedings of the 27th International Conference on Computational Linguistics, pp. 3146–3157. Association for Computational Linguistics

Kepler, F., Trénous, J., Treviso, M., Vera, M., Martins, A.F., 2019. OpenKiwi: An Open Source Framework for Quality Estimation. ACL 2019 117. arXiv preprint arXiv:1902.08646.

Kim, H., Lee, J.-H., Na, S.-H., 2017. "Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation". Presented at the Proceedings of the Second Conference on Machine Translation, pp. 562–568. WMT 2017, 562.

Pires, T., Schlinger, E., Garrette, D., 2019. "How Multilingual is Multilingual BERT? " Presented at the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4996–5001. arXiv preprint arXiv:1906.01502.

Ranasinghe, T., Orăsan, C., Mitkov, R., 2020. "TransQuest: Translation Quality Estimation with Cross-lingual Transformers". Presented at the Proceedings of the 28th International Conference on Computational Linguistics, pp. 5070–5081. arXiv preprint arXiv:2011.01536.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J., 2006a. "A study of translation edit rate with targeted human annotation." Presented at the Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, pp. 223–231. arXiv preprint arXiv:2102.04020.

Zerva, Chrysoula, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José GC de Souza, Steffen Eger, Diptesh Kanojia et al. "Findings of the wmt 2022 shared task on quality estimation." In Proceedings of the Seventh Conference on Machine Translation. 2022. Findings of the WMT 2022 Shared Task on Quality Estimation.

استناد به این مقاله: جعفری هرندی، محمد حسین؛ آزادی، فاطمه؛ رفیعی، سپهر؛ فیلی، هشام و دوستی، محمدجواد (۱۴۰۱). ارائه یک مدل تخمین کیفیت مترجم ماشینی. *زبان و زبان‌شناسی ۱۸* (۳۵)، ۸۶ـ۷۱. doi: 10.30465/lsi.2023.8499