TASHKENT UNIVERSITY OF
INFORMATION TECHNOLOGIES
NAMED AFTER MUHAMMAD AL-KHWARIZMI

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

BULLETIN OF TUIT:

MANAGEMENT AND COMMUNICATION TECHNOLOGIES



ISSN:2181-1083 Date:29/12/2024 DOI:
Nazirova Elmira, Abidova Shakhnoza, Uzakova Mamura, Boymurodov Farrukh
№ 4(40)2024

Bulletin of TUIT: Management and Communication Technologies

MATHEMATICAL MODEL FOR EVALUATION OF TRILINGUAL ELECTRONIC TRANSLATION

Nazirova Elmira Shodmonovna
DSc, Department of Multimedia Technologies Tashkent university of information technologies named after Muhammad al-Khwarizmi Tashkent city 100 000, Uzbekistan
Email: elmira_nazirova@mail.ru

Abidova Shakhnoza Bakhodirovna
PhD, Department of Multimedia Technologies Tashkent university of information technologies named after Muhammad al-Khwarizmi Tashkent city 100 000, Uzbekistan
Email: shaxnoza23@mail.ru

Uzakova Mamura Abdurayimovna
PhD student, Department of Multimedia Technologies Tashkent university of information technologies named after Muhammad al-Khwarizmi Tashkent city 100 000, Uzbekistan
Email: marynmm262321@gmail.com

Boymurodov Farrukh Farhod o'g'li
PhD student at the Department of Applied Mathematics, Karshi State University, Karshi city 180100, Uzbekistan
Email: farruxboymurodov96@gmail.com

Abstract. The paper considers the issue of improving the mathematical model for a hybrid method of machine translation evaluation. Today, electronic translation remains a popular tool. The development of technologies and adaptation to linguistic and cultural features makes it indispensable in education, business and everyday life. The scientific article is devoted to the study of improving the model for evaluating electronic translation. The authors paid special attention to the analysis of existing methods for evaluating electronic translation, such as BLEU, METEOR, ROUGE was improved on the basis of which a new mixed model for evaluating electronic translation. An exact, also given example of calculating the assessment with an incoming text in the Uzbek language, which is translated into the Kyrgyz and Karakalpak languages. As a result, assessments were obtained using the METEOR, ROUGE methods. In addition, digital results of evaluating electronic translation based on the hybrid method were obtained. An analysis of the comparison of the translation platforms GOOGLE Translate, PROMT, Yandex and DEEPL after the implementation of the above methods is given.

Keywords: *electronic translation, mathematical model evaluation, BLEU, METEOR, ROUGE, MRS, hb.*

Introduction. Today, electronic translation remains in demand in many areas. The main reasons for its popularity are globalization and intercultural communication, technological development, accessibility for rare languages, high speed and scalability.

Translation quality assessment plays an important role in ensuring the accuracy of translation. The first machine translation systems were assessed manually: experts analyzed grammatical and syntactic errors, checked the accuracy of keywords and the main meaning. With the development of statistical machine translation (SMT), there was a need for automated assessment, since manual verification became too labor-intensive. BLEU (Bilingual Evaluation Understudy) became the first widely accepted metric comparing n-gram matches between machine translation and the standard.

BLEU (Bilingual Evaluation Understudy Score) is a metric for evaluating the quality of machine translation. It compares n-grams (sequences of n words) in the translated text with reference translations. BLEU uses the principle of "preferential accuracy" with a penalty for redundancy (penalty for too long translations). The BLEU value ranges from 0 to 1, where 1 means a complete match with the reference [1].

METEOR (Metric for Evaluation of Translation with Explicit ORdering) - an

assessment of the quality of translation taking into account synonyms, morphological changes and word order [2]. Evaluates the accuracy and completeness of the translation, including the correspondence of synonyms, lexical and syntactic variations. More flexible and sensitive to differences in vocabulary and word order than BLEU. Range: from 0 to 1, where 1 means a perfect translation.

ROUGE (Recall - Oriented Understudy for Gisting Evaluation) - evaluation of the quality of compression systems, such as text summarization. Includes several metrics, such as ROUGE-N (n-gram comparison), ROUGE-L (sequence comparison). Focuses on recall, especially the importance of long fragments of text. Range: from 0 to 1, where 1 means a perfect match with the reference [3].

F1-Score - a general metric for assessing the balance between precision and recall of classifiers. It is calculated as the harmonic of the average precision and recall. A value of 1 indicates a perfect balance between precision and recall, while a value of 0 indicates no balance [4]. Each of these metrics is used in the context of natural language processing to assess the quality of machine translations, summarization, or other text-based tasks.

The development of metrics for machine translation is the result of the work of many scientists and research groups. Each new metric represents a step forward in our understanding of how to automatically evaluate complex language tasks, getting closer to human perception of translation quality. This is especially important for languages with fewer speakers, such as Uzbek, Kyrgyz, or Karakalpak, since machine translation helps preserve and disseminate linguistic heritage. Although there are no scientists in Uzbekistan focused exclusively on developing metrics for evaluating machine translation, local researchers are actively working on improving translation for the Uzbek language. They use international metrics and adapt them to the specifics of local languages. As interest in NLP grows in Central Asia, we can expect more active participation of Uzbek scientists in this field.

Universities in Uzbekistan, such as the National University of Uzbekistan (NUU) and the Tashkent University of Information Technologies (TUIT), conduct research aimed at creating and improving translation systems for the Uzbek language. Researchers working in this field, such as Pulatov A.K., Khakimov M., Abdurakhmanova N., Turaeva Sh. and others, specialize in computational linguistics. Their works consider the problems of machine translation and propose ways to solve them [5]. Research in the field of combining metrics is carried out by such scientists as Graham Newman, Lucia Spano, Katie Lofus. They work to improve the automatic evaluation of machine translation, developing methods for combining different metrics to more accurately match human assessments. Their research also concerns the adaptation of metrics to low-resource languages and the creation of hybrid models that take into account morphological, syntactic and semantic features [6].

The WMT (Workshop on Machine Translation) team holds annual machine translation competitions, where combined metrics are developed. For example, the COMET model, which combines BLEU, METEOR and other metrics using neural networks. In foreign countries, scientists Kishore Papineni (BLEU), Alon Lavi (METEOR), Matthew Snover (ROUGE-L), Chin-Yew Lin, Thibault Sellam, Dipanjan Das, Ankur Parikh are the authors of metrics for machine translations [7-9].

Main Part. Let's consider the most popular methods, also known as metrics, for assessing the quality of machine translation, such as BLEU, METEOR, ROUGE-1. The formula for calculating BLEU looks like this [10]:

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

here :

BP (Brevity Penalty) — a penalty for brevity, applied to ensure that the translated text should not be too short compared to the reference. It prevents translations that are too short, which can lead to an inflated score if the system translates only part of

the original text, and is calculated as follows*(2):*

$$BP = \begin{cases} 1, \text{ if translation length} \geq \text{reference length} \\ \exp\left(1 - \dfrac{\text{translation length}}{\text{reference length}}\right), \text{ if translation length} \prec \text{reference length} \end{cases}$$

p_n is the precision for n-grams (n from 1 to N), where pn is calculated as:

$$p_n = \frac{\sum_{translation} number\ of\ matches\ n-gramm}{\text{total number of matches } n-gramm \text{ in translation}} \quad (3)$$

Precision of n-grams measures how often n-grams from the translated text appear in the references. This is an important aspect, since BLEU rewards n-gram matches.

BLEU takes values from 0 to 1, where 1 means a perfect match of the translation with the reference texts. It takes into account the following aspects:

- Precision — the proportion of matching words between the translated text and references.
- Recall — the proportion of words from the reference text that matched the translated text.
- Synonyms and morphology - takes into account, for example, synonyms or different forms of words.
- Word order penalty - a penalty is imposed for discrepancies in word order between the translation and the reference.

Formule METEOR:

$$METEOR = GM \times (1-\alpha) \quad (4)$$

GM(Harmonic Mean) – harmonic mean between precision (P) and recall (R) [11]:

$$GM = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \quad (5)$$

Гармоническое среднее (GM) помогает учесть как точность, так и полноту, чтобы сбалансировать эти два показателя, избегая переоценки одного из них[12].

-P –precision:

$$P = \frac{\text{number of matching words}}{\text{total number of words in translation}}$$

- R - recall:

$$R = \frac{\text{number of matches}}{\text{number of words in the sample}}$$

The word order penalty is imposed to take into account how important it is for the correct order of words in a sentence to more accurately reproduce the meaning.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a series of metrics proposed by Chin-Yew Lin in 2004 to evaluate the quality of the solution to the text summarization problem.

Like BLEU, ROUGE is based on counting the coincidences of words and phrases in machine translation and in the standard. However, ROUGE takes into account not only precision, but also recall, as well as the BLUE parameter. This allows one to do without a brevity penalty.

The formula for ROUGE-1 (where 1-grams are compared between generated and reference text) estimates recall, precision, based on 1-gram (i.e. individual word) matches [13].

Recall for ROUGE-1 (recall):

$$ROUGE-1 \cdot Recall = \frac{\text{number of matching 1-grams}}{\text{total number of 1-gram in the standard}}$$

This measures how well the generated text covers all 1-grams of the reference text.Precision for ROUGE-1:

$$ROUGE-1 \cdot \text{Pr}ecision = \frac{\text{number of matching 1-grams}}{\text{total number of 1-grams in generated text}}$$

This measures how accurately the reference text uses the same 1-grams as the reference text[14].

$$ROUGE-1 = 2 \cdot \frac{ROUGE-1 \cdot \text{Pr}ecision \cdot ROUGE-1 \cdot Recall}{ROUGE-1 \cdot \text{Pr}ecision + ROUGE-1 \cdot Recall}$$

Statistics on machine translation evaluation metrics can be seen in the following diagram:

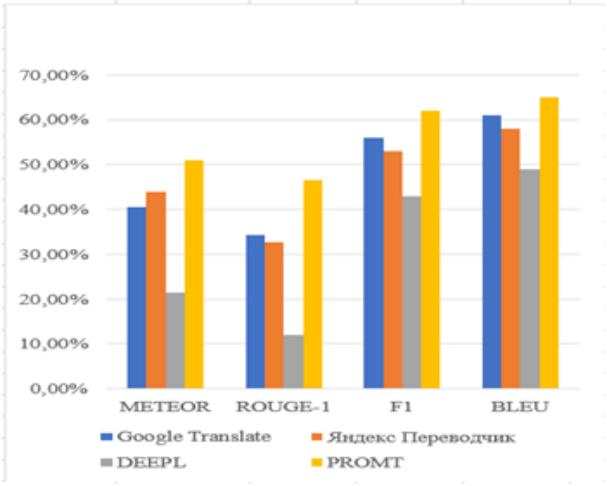


Fig. 1. Statistics of machine translation evaluation metrics

As can be seen from the diagram (Figure 1), the PROMT translator showed high results for all metrics. The lowest indicator is observed for DeepL. In Google and Yandex translators, the translation score is presented as an average value. This means that the translation performed using DeepL is of higher quality and accuracy. Combining the METEOR and ROUGE-1 metrics allows you to get a more complete assessment of the quality of machine translation. Each of these metrics focuses on different aspects of translation, and their combination allows you to take into account the accuracy, completeness, semantic equivalence and structural features of the text. Consider how this can be implemented.

To obtain an overall score, these metrics can be combined using a weighted average or normalized approach [15].

Combining formulas (1) and (2) yields the following combined formula:

METEORE+ROUGE-1

$$\frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} + 2\frac{P \cdot R}{P+R} \quad (6)$$

In expression *(6)* enter the Suffix parameter and obtain the following model:

$$\frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} + 2\frac{P \cdot R}{P+R} + Suffix_{PR} \quad (7)$$

here, Suffix – suffixes of natural languages. PR-Precision and Recall of Suffix. Explanation $Suffix_{PR}$:

$$Suffix_{PR} = \begin{cases} Suffix_{PRuzb} \\ Suffix_{PRkk} \\ Suffix_{PRkg} \end{cases}$$

$$Suffix_{uz} = \oplus\downarrow V_{z1} \oplus\downarrow F_{y1} \oplus\downarrow I_{f1}$$
$$Suffix_{kg} = \oplus\downarrow V_{z2} \oplus\downarrow F_{y2} \oplus\downarrow I_{f2}$$
$$Suffix_{kk} = \oplus\downarrow V_{z3} \oplus\downarrow F_{y3} \oplus\downarrow I_{f3}$$

here,

V_z-word-forming suffix,

F_y-formative suffix,

I_f-inflectional suffix,

$\oplus$ - connecting sign,

$\downarrow$ - sign may or may not be.

The Suffix parameter provides a more accurate assessment of suffixes in words and phrases. It is assessed through morphological analysis. As is known, morphological analysis is an analysis of a word that includes determining its part of speech, morphological features (gender, number, case, tense, etc.) and syntactic role in a sentence.

Computational examples:

Calculation example: Metrics normalization: All metrics are brought to the same value range (e.g. from 0 to 1). Weight assignment: Each metric is assigned a weight depending on its importance for a specific task [16-20].

For example:

- METEOR: grammatical accuracy.
- ROUGE-1: lexical matches

To calculate METEOR, ROUGE-1 for Uzbek text using the following sentence as an example:

«Компьютернинг ишлаш тезлиги тезлашган»

(“Computer speed has increased”)

in Kyrgyz and Karakalpak languages, consider several hypothetical examples for the generated text and reference texts. METEOR (in Kyrgyz and Karakalpak languages) assume that:

Generated text(uzb)	*Компьютернинг ишлаш тезлиги тезлашган*
Reference text(kg)	*Компьютердин иштеш ылдамдыгы тездеген*
Reference text(kk)	*Компютердиң ислев тезлиги тезленган*

To calculate METEOR we will use the following steps: Precision and Recall. Calculate the number of matching words between the generated and reference text.

Matching words for Kyrgyz	Matching words for Karakalpak
«Компьютердин» «иштеш» «ылдамдыгы»	*«Компютердиң» «ислев» «тезлиги»*

Now will be calculating:
For kyrgyz text: 3 out of 4 words matched.
Precision=3\4=0.75
Recall=3\4=0.75
For karakalpak text: 3 out of 4 words matched.
Precision=3\4=0.75
Recall=3\4=0.75
Harmonic Mean (GM)

$$GM = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \cdot 0{,}75 \cdot 0{,}75}{0{,}75 + 0{,}75} = 0{,}75$$

Word order penalty. If the word order is the same, the word order penalty is minimal. METEOR=0.75
For ROUGE-1 Recall metric (1-gram comparison):
Matched 1-grams:

for Kyrgyz	for Karakalpak
«Компьютердин» «иштеш» «ылдамдыгы»	*«Компютердиң» «ислев» «тезлиги»*

Total number of 1-grams in reference texts: 4 words in both references.

Language / Metrika	For kyrgyz language	For karakalpak language
ROUGE-1	3\4=0.75	3\4=0.75

Now to calculation $Suffix_{PR}$

Kyrgyz language:
«Компьютердин иштеш ылдамдыгы тездеди»
here:

Uzbek	*Kyrgyz*
-нинг	*-дин*
-ш	*-ш*
-ган	*-ди*

Karakalpak language:
«Компьютердин ислев тезлиги тезленген.»
here:

Uzbek	*Karakalpak*
-нинг	*-дин*
-ш	*-в*
-ган	*-ген*

Analysis by suffixes:
To assess the quality of the translation, the following morphological elements can be compared:
Genitive case:

In Uzbek	*- нинг*
In Kyrgyz	*- нин*
In Karakalpak	*- дин*

The coincidence confirms the correct rendering of the case.
Process (verbal action)partially preserved through the root:

In Uzbek	- ш
In Kyrgyz	- ш
In Karakalpak	- ев

Past tense:

In Uzbek	*-ган*
In Kyrgyz	*-ди*
In Karakalpak	*-ген*

This is taken into account as a correct translation based on equivalent grammatical features.

$Suffix_{PR}$: For automatic translation, you can calculate the accuracy and completeness of the comparison. For example:

Out of 3 suffixes, 2 completely matched, and one (verbal action) matched partially.

Suffix $_{Precision} = 2.5 / 3 \approx 0.83$ (83%)

Suffix $_{Recall} = 2.5 / 3 \approx 0.83$ (83%)

$Suffix_{PR} = 2 \times (0.83 \times 0.83) / (0.83 + 0.83) \approx 0.83$ (83%)

This approach helps to assess how well the morphological features of the text are conveyed when translated into agglutinative languages.

Results:

language / metrica	Uzbek→Karakalpak→ Kyrgyz
METEOR	0.75
ROUGE-1	0.75
Suffix	0.83
MRS_{hb}	***0.782***

MRS_{hb} – METEORE, ROUGE-1, Suffix

hb - hybrid

MRS_{hb}= METEORE + ROUGE-1+Suffix

Taking into account the weights:

$$MRS_{hb} = 0.4 \cdot 0.83 + 0.4 \cdot 0.75 + 0.2 \cdot 0.75 = 0.782$$

The final score of 0.782 reflects the overall quality of the translation, taking into account morphological accuracy, semantic equivalence and lexical content.

The advantages of combining are that different aspects of translation quality are taken into account and a more complete assessment is achieved compared to using a single metric.

Thus, the METEOR, ROUGE-1 and Suffix metrics equally assess the similarity between machine and reference texts, and the result for both languages was 0.75, indicating a high similarity between the machine and reference texts.

Conclusion. Translation quality, especially when using hybrid metrics, should take into account several aspects, including accuracy, grammar, contextuality, naturalness, and relevance to the specifics of the target language. In case of using the hybrid method in evaluating electronic translation, it is also important to consider the use of different methods, such as metrics based on comparisons with reference translations (e.g. BLEU, METEOR, ROUGE-1). These metrics can provide a quantitative assessment of the accuracy of the translation, but it is important to remember that they do not always fully reflect the quality of the translation from a human perspective. These subjective assessments provide an idea of how the translation is perceived in the context and how well it meets the expectations of users. Context-sensitive assessment means that the translation should not only be accurate, but also logical within the entire text, preserving its stylistic and cultural features. In mixed machine translation evaluations, it is important to use multiple methods to achieve a more balanced and reliable evaluation. It is necessary to consider not only the number of matches with the reference translation, but also aspects such as the naturalness of the text, the logic of the translation, and its relevance to the target audience. Automatic metrics such as BLEU can be useful for an initial evaluation, but for a complete picture, it is important to involve human experts who can identify nuances that machine translation may not capture. Therefore, BLEU was not chosen for the hybrid method.

REFERENCES

[1] Papineni K., Roukos S, Ward T., Zhu W. J., BLEU: a method for automatic evaluation of machine translation. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pp. 311–318.

[2] Coughlin D., "Correlating Automated and Human Assessments of Machine Translation Quality" in MT Summit IX, New Orleans, USA pp. 23–27. 2003

[3] Callison-Burch C., Osborne M., Koehn P. "Re-evaluating the Role of BLEU in Machine Translation Research" in 11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006. pp. 249–256.

[4] Lin C., Oching F. "Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics" in Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics. 2004.

[5] Banerjee S., Lavie A. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" in Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics, Ann Arbor, Michigan, June 2005.
[6] Переходько И.В., Мячин Д.А., «Оценка качества машинного перевода», 2017. 260с.
[7] Kishore Papineni"BLEU: a Method for Automatic Evaluation of Machine Translation" (2002).
[8] Alon Lavie "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" (2004).
[9] Matthew Snover"A Study of Translation Edit Rate with Targeted Human Annotation" (2006).
[10] Chin-Yew Lin "ROUGE: A Package for Automatic Evaluation of Summaries" (2004).
[11] Thibault Sellam, Dipanjan Das, Ankur Parikh "BLEU might be Guilty: A Study of Bias in Evaluation Metrics for Machine Translation" (2019).
[12] Banerjee S., Lavie A. «METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments» in Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005
[13] Lavie, A., Sagae, K. and Jayaraman, S. (2004) «The Significance of Recall in Automatic Metrics for MT Evaluation» in Proceedings of AMTA 2004, Washington DC. September 2004 Banerjee, S. and Lavie, A. (2005)
[14] Alon Lavie, Abhaya Agarwal. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments // Proceedings of the Second Workshop on Statistical Machine Translation. — 2007. — С. 228—231. 26 октября 2020 года.
[15] Michael Denkowski, Alon Lavie. [https://aclanthology.org/N10-1031.pdf Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level] // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL. — С. 250–253. 2022.
[16] Elmira Nazirova, Shaxnoza Abidova, Shaxnoza Obidova Mathematical model and algorithm of simple words and simple sentences in the Uzbek language // 2021 International Conference on Information Science and Communications Technologies (ICISCT) (Scopus). Date Added to IEEE Xplore: 17 January 2022. DOI: 10.1109/ICISCT52966.2021.9670202
[17] Nazirova E.Sh., Abidova Sh.B., Hojiev S.A., Jamolov N.F. Construction of a mathematical model of complex word formation in the Uzbek-Karakalpak language. // Scientific horizon in the context of social crises. Proceedings of the 7th International Scientific and Practical Conference. Tokyo-Japan, 6-8.02.2021 y. pp.1117-1123.
[18] Пулатов А.К., Джураева Н. Разработка формальной модели грамматики узбекского языка // Узбекский математический журнал. – Ташкент, 2002. – №1. – С. 47-54.
[19] Коваль С.А. О сравнимости и эквивалентности компьютерных представлений морфологии // Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. конференции Диалог'2003 (Протвино, 11–16 июня 2003 г.) / Под ред. И.М. Кобозевой, Н.И. Лауфер, В.П. Селегея. М.: Наука, 2003. С. 305–311.
[20] Venkata Sireesha Nagineni, Neerudi Bhoopal, DSNMRAO, Kasireddy Idamakanti, Devineni Gireesh Kumar, Bhanu Teja Bacha “Quality Evaluation and Assessment of Machine Translations for English to Turkish (Google Translations Vs DeepL Translations)” Global Conference on Information Technologies and Communications (GCITC) Karnataka, India. Dec 1-3, 2023