A HYBRID RULE-BASED AND TRANSFORMER APPROACH FOR UZBEK-ENGLISH MACHINE TRANSLATION: EVALUATION USING METEOR

Elmira Nazirova[1], Nodira Samigova[2], Kamola Usmonova[3]

[1]Professor of Tashkent University of Information Technologies named after Mukhammad Al-Khwrazmi,
[2]Senior lecturer of the Pharmaceutical Education and Research Institute
[3]PhD of the the Department of Multimedia,Technologies, Tashkent University of Information Technologies, named after Al-Khwrazmi
E-mail:kamolausmonova93@gmail.com

KEYWORDS

ABSTRACT

This article presents how hybrid translation tool works. Specifically, it indicates rule-based translation's IDF0 and also gives information about our tranlation system dataset which is done based transformer model. Moreover, it adduces the output evaluation of machine translation quality using the METEOR metric, with comparing the output of the English-Uzbek translation tool against human-produced translations. The study discusses how evaluation methodologies for machine translation systems work in bilingual contexts, emphasizing the importance across diverse domains. With the rapid enhancement of technology, there are different translation tools, each improving translation precision and fluency. This research provides a detailed analysis of translation evaluation metrics, specifically METEOR, applied to sentence-level translation tasks from English to Uzbek. It shows how low-resourced language is translated in our translation and how it is close to human translation.

Introduction

These days of globalization, translation tools play an important role in facilitating communication across linguistic boundaries. Translation tools are utilized in various domains, including education, scientific research, business, and other spheres. In Uzbekistan, a series of governmental decrees have been enacted to promote foreign language acquisition, with a special focus on getting internationally recognized language certificates. Such initiatives encourage all learners significantly and give them the chance for academic advancement, professional development, and integration into the global knowledge economy.

In the field of research, scholars often access valuable materials is often in English, which are almost exclusively scientific literature, conference proceedings, and technical documentation is published in this language. As a result, not being aware of efficient translation mechanisms can create a problem in understanding deeply of any knowledge. Delivering accurate, context-sensitive, and domain-specific translations are very important in any field.

Taking all these considerations, we create a translation tool specifically designed to help bidirectional translation between Uzbek and English and vice versa. With this tool, it can be easy to enhance both the accessibility of English-language resources for Uzbek speakers and the international visibility of Uzbek-language research outputs, in addition, it can play an essential role in global academic exchange.

In this study, we conducted a comparative analysis between the output of our translation system and human translation, assessing the degree of similarity between the two. The evaluation, carried out using the METEOR[1] metric, yielded an accuracy score of 86%, which can be considered a satisfactory result.

© Elmira Nazirova, Nodira Samigova, Kamola Usmonova

Method

As ALPAC 1966 and ARPA mentioned evaluation matrix had its important role on the field large-scale [2]. ALPAC, which stands for Automatic Language Processing Advisory Committee, studies the comparison output of human translation with machine translation and concludes the quality of human and machine translations is highly reliable [3]. There are 5 attributes of good automatic metric. They are reliability, consistency, generality, correlation and sensitivity. In fact if there is a high correlation with human assessment, which has similar result on similar text then it is considered good matric. Matric should work with different text domains in various machine translation tasks [4]. There are different evaluation matric such as BLEU, ROUGE, METEOR, NIST, LEPOR and others. BLUE correlates highly with human assessment. This is one of the most well-known. BLEU compares n-gram. Behind this metric lies it would be better if translation is close to human translation. At corpus level it is known its highly correlation with human assessment of quality [5]. It works best for language with similar word order [6]. However, BLEU has some limitations which is it is sensitive to exact wording. That means only small changes effect drastically its score even meaning is preserved. Moreover, it ignores word order and synonyms [7]. Meanwhile, ROUGE places more on recall beneficial capturing all meaning than exact precision as well as it is less sensitive to fluency and exact grammar. While the study of claims that METEOR addresses the limitations found in the BLEU approach [8]. METEOR matric considers synonyms, stems and word order better for morphologically rich languages. Yet, it has also limitation such as it requires language resources. Among these matric we chose to use meteor taking into consideration that Uzbek has morphological richness, its tolerance of synonym, precision and recall balance and its better correlation with human assessment. Meteor is a matrix to show translation tools accuracy. Meteor is created in order to address the BLUE weaknesses. Meteor computes score between translation and reference translation word to word explicitly. On condition that there are more than one reference translation the given translation is calculated against each reference independently, after scoring the best solution is introduced. After comparing a pair translation between two strings are created an alignment. Moreover, Precision and Recall are computed: P is calculated how many match words in the reference translation and it is divided by the total number of words in the system translation. While recall is counted how many words match words in reference translation then is divided by total number of words in the reference translation. After finding Precision and recall F_{mean} is calculated based on this formula. $F_{mean} = \frac{10*\square*\square}{\square+9\square}$ .Precision, recall, and F-mean are based on single-word (unigram) matches.in case there are longer word matches, METEOR adds a penalty. First, the matched words in the system translation are put into the smallest possible number of groups (chunks). In each chunk, the words must be next to each other in both the system translation and the reference translation. If the match has long word groups, there will be fewer chunks. If the whole system translation is exactly the same as the reference translation, there will be only one chunk. If there are no matches longer than one word, there will be as many chunks as matched words. The penalty is then calculated using a special formula.

$$Penalty=0{,}5\left(\frac{chunks}{matches}\right)^3$$

Finally, the METEOR Score can be computed as follows: Score = Fmean 1(* − Penalty.For one system translation, METEOR calculates the score for each reference translation. Then, it chooses the highest score as the final score for that translation.

The total METEOR score for the whole system is found using all sentences in the test set. This is similar to how BLEU works. We first find the total precision, total recall, and total penalty for all sentences. Then, we put them together using the same formula that is used for single sentences.

Discussion

Operation Sequence of the Translation Tool

In our rule-based layer the process follows a structured sequence. A portion of the input text is selected and sent for translation. First, the type of sentence is determined based on punctuation marks. Each word is then loaded into a sentence vector, and tokenization is performed.Next, each

(cc) BY

word is divided into word so tokinization occurs and represented as an X vector. This WORD vector contains elements such as the prefix, root, derivational affix and inflectional affix The WORD vector is represented as a binary sequence of 1s and 0s.Following this, morphological analysis is conducted to identify the root and affixes of each word, as well as to determine its part of speech. Then, syntactic analysis establishes the position of the word within the sentence and identifies its grammatical components.The subsequent step is semantic analysis, which interprets the meaning of the word in context and resolves homonymy by considering the frequency of occurrence in the text.After analysis, the WORD vector is replaced with its corresponding English equivalent. If the comparison probability between the original and translated WORD vectors is high, the translation is accepted. If the probability is low, synonyms or equivalent phrases are selected, and the comparison process is repeated. Finally, the fully translated text is displayed on the screen.

Photo 1 indicates of UzMarine IDF 0. In order create Uzmarine translation system we use hybride approach. The parallel corpus and dictionary is used in this system. Hence, in preprocessing stage sentences are normalized by omitting unnecessary word and signs. In Morphological Analyze and rule based translation stages after sentence divided into words each word is analyzed both languages morphology rules and as mentioned above. Each word is stemmed and divided affixes and signed word vector. In our research thirty-two possible sentence patterns and their corresponding equivalents in Uzbek and English were examined in this study. Based on this sentences we developed mathemathical model and system used these. After that, it order to achieve high accuracy we decided to use LLM (large language model). We used our corpus for fine-tuning in Neural Machine Translation System and we got output of target language.

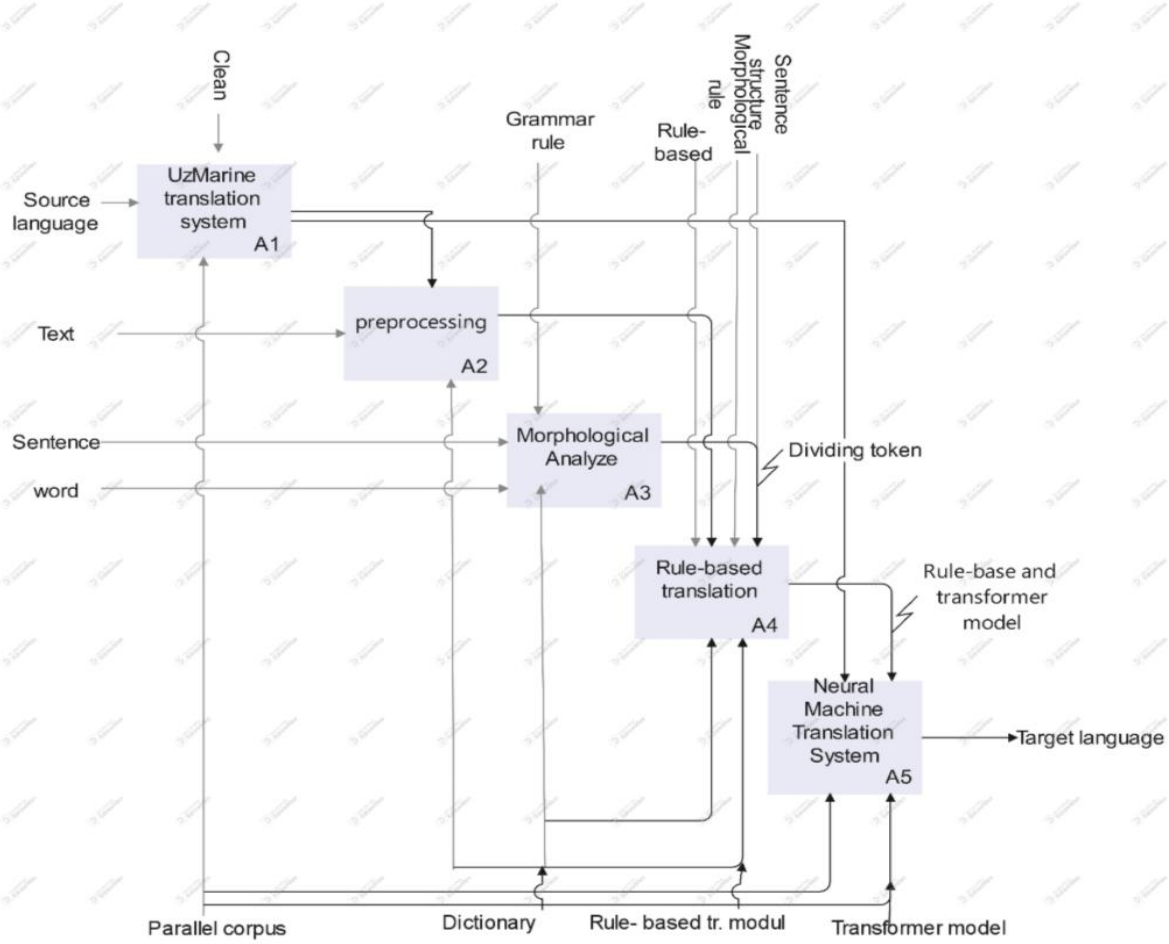


Photo 1. IDF0 of UzMarine translation system

(cc) BY © Elmira Nazirova, Nodira Samigova, Kamola Usmonova

Analyzing morphology, both languages is very important to achieve precise translation our rule- based translator is done with Django (photo 2). IndexView layer works with user, it helps to open main page for user. In translate, user send the text in ApiView layer. Logic layer helps to identifying sentence types and to work with the verb types. As it is clear there are different types of verbs compound verb, adverbial and others. First, it identifies the sentence types and if sentence is complex it sends to ComplexPredicteHandler, otherwise it checks verbal, nominal or extential.

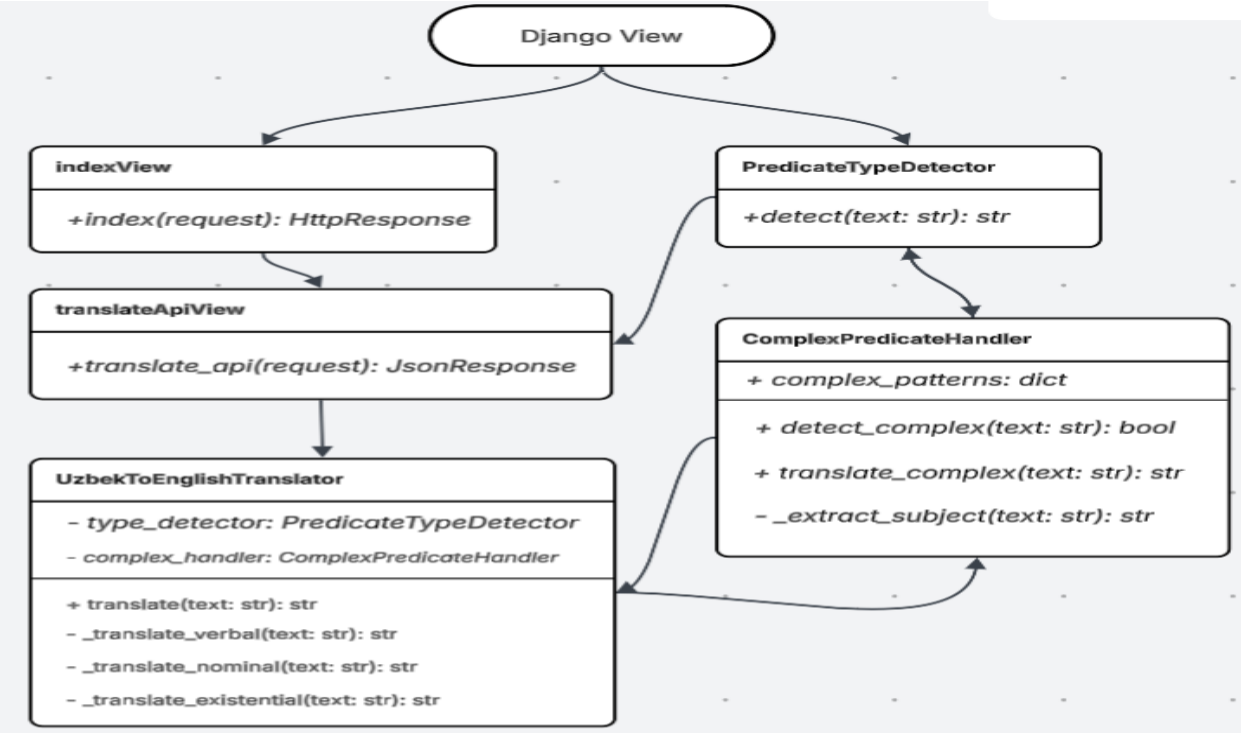


Photo 2. Diagram of what rule-based translation does during the translation process

Furthermore, using transformer model, we have created our own model, we use 142838 paired- parallel sentences, and our dataset is introduced (photo 3).

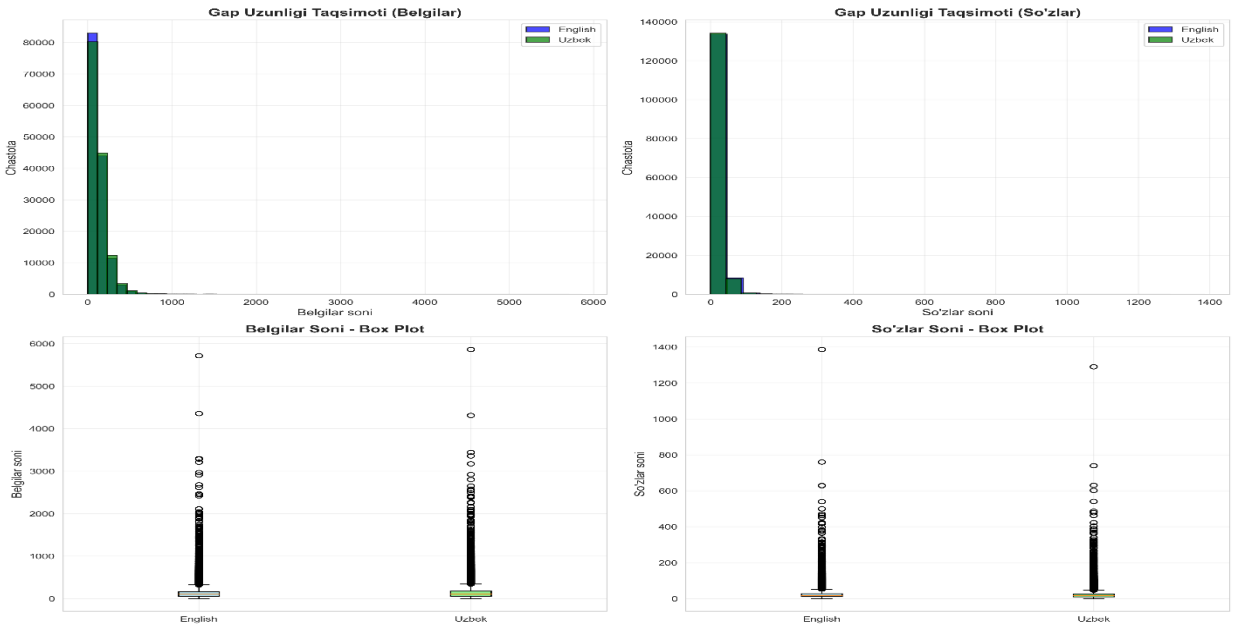
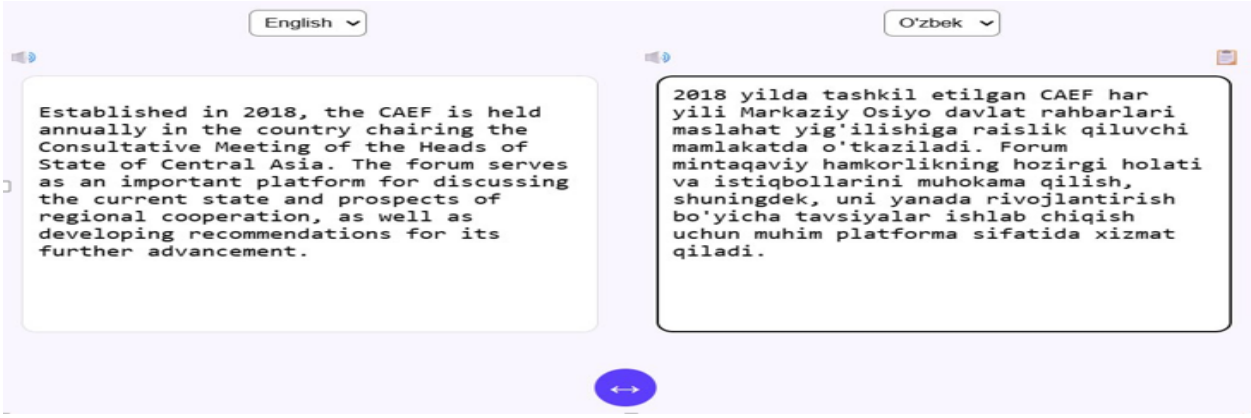


Photo 3. How many words are used in translation dataset

(cc) BY © Elmira Nazirova, Nodira Samigova, Kamola Usmonova

We entered the text to our translation tool, the sentence is: *Established in 2018, the CAEF is held annually in the country chairing the Consultative Meeting of the Heads of State of Central Asia. The forum serves as an important platform for discussing the current state and prospects of regional cooperation, as well as developing recommendations for its further advancement.*



UzMarin translation tool

Our translation tool translated : **Candidate translation:**

2018 yilda tashkil etilgan CAEF har yili Markaziy Osiyo davlat rahbarlari maslahat yig'ilishiga raislik qiluvchi mamlakatda o'tkaziladi. Forum mintaqaviy hamkorlikning hozirgi holati va istiqbollarini muhokama qilish, shuningdek, uni yanada rivojlantirish bo'yicha tavsiyalar ishlab chiqish uchun muhim platforma sifatida xizmat qiladi.

Reference translation:

2018-yilda tashkil etilgan MOIF har yili Markaziy Osiyo davlatlari rahbarlarining Maslahat uchrashuviga raislik qilayotgan mamlakatda o‘tkaziladi. Forum mintaqaviy hamkorlikning bugungi holati va istiqbollarini muhokama qilish, uni yanada rivojlantirish bo‘yicha tavsiyalar ishlab chiqish uchun muhim maydon bo‘lib xizmat qiladi.

In order to know how precise the tool translate the sentences we calculate in meteor matrix: To do this we have to compare word level: Exact matches, Stem matches, Synonym matches.

N	Candidate tokens (tool)	Reference tokens (human)
1	2018	2018-yilda +
2	yilda	Tashkil+
3	tashkil	Etilgan+
4	etilgan	MOIF -
5	CAEF	Har+
6	har	Yili+
7	yili	Markaziy+
8	Markaziy	Osiyo+
9	Osiyo	Davlatlari +
10	davlat	Rahbarlarining +
11	rahbarlari	Maslahat +
12	maslahat	Uchrashuviga -
13	yig'ilishiga	*Raislik*+
14	raislik	*Qilayotgan* -
15	qiluvchi	*Mamlakatda* +
16	mamlakatda	*o‘tkaziladi* +
17	o'tkaziladi.	*Forum* +
18	Forum	*Mintaqaviy* +
19	mintaqaviy	*Hamkorlikning* +
20	hamkorlikning	*Bugungi* -
21	hozirgi	*holati* +
22	holati	*Va* +
23	va	*Istiqbollarini* +
24	istiqbollarini	*Muhokama* +
25	muhokama	*Qilish* +
26	qilish	*Uni* +
27	shuningdek	*Yanada* +
28	uni	*Rivojlantirish* +
29	yanada	*bo‘yicha* +
30	rivojlantirish	*Tavsiyalar* +
31	bo'yicha	*Ishlab* +
32	tavsiyalar	*Chiqish* +
33	ishlab	*Uchun* +
34	chiqish	Muhim +
35	uchun	*Maydon* -
36	muhim	*bo‘lib* -
37	platforma	*Xizmat* +
38	sifatida	*Qiladi* +
39	xizmat	
40	qiladi	

Exact matches are 33

(cc) BY © *Elmira Nazirova, Nodira Samigova, Kamola Usmonova*

$$Precision = \frac{M}{candidate\ length} = \frac{33}{40} = 0{,}825 \ ,Recall = \frac{M}{reference\ length} = \frac{33}{38} = 0{,}8684$$

$$Penalty = 0{,}5\left(\frac{chunks}{matches}\right)^3 \quad Penalty = 0{,}0000625$$

$$\alpha = 0{,}5,\ Meteor = 0{,}8638 * (1 - 0.0000625) = 0.8637$$

Final score was 86%

Conclusion

This study assessed the performance of our translation tool by comparing its output with a human-produced translation for a selecting any text. The evaluation was conducted using the METEOR metric, which considers precision, recall, and alignment penalties to provide a more comprehensive measurement. Our tool achieved a score of **0.86**, indicating a high level of similarity to human translation quality.

In translation system we use hybride method which includes rule-based and model based on transformer model. The analysis revealed that the system demonstrates strong performance in translating news-related materials, where sentence structures and vocabulary tend to be more standardized. However, performance in more specialized domains-such as technical, scientific, and literary texts - remains an area for further development. Future work will focus on enhancing domain adaptability through the integration of larger and more diverse training datasets, refinement of linguistic rules, and the application of advanced neural translation architectures.

By continually refining both the linguistic and computational components of the system, we aim to create a translation tool that delivers consistently high-quality results across a wide variety of content types, thereby supporting more effective cross-linguistic communication.

Reference

1. S. Banerjee and A. Lavie, *"METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,"* in *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA, June 2005, pp. 65–72.
2. White, J., O'Connell, T. and O'Mara, F. (1994) "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches". *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas. Columbia, MD* pp. 193–20.
3. ALPAC (1966) "Languages and machines: computers in translation and linguistics". A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council, 1966.
4. Banerjee, S. and Lavie, A. (2005) "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.*
5. Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "BLEU: a method for automatic evaluation of machine translation" in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311–318
6. N. Peng, L. Liu, Y. Wang, L. Wang, Y. Wei, and Y. Wang, "Quality Evaluation Model of Automatic Machine Translation Based on Deep Learning Algorithm," in *Proc. 2023 4th Int. Conf. for Emerging Technology (INCET)*, Belgaum, India, May 26–28, 2023.
7. M. Labied, A. Belangour, and M. Banane, "Assessing Speech-to-Text Translation Quality: An Overview of Key Metrics," in *Proc. 2024 Int. Conf. on Decision Aid Sciences and Applications (DASA)*, 2024, doi: 10.1109/DASA63652.2024.10836447.
8. S. Banerjee, and A. Lavie, "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, pp. 65-72, 2005.

(cc) BY © *Elmira Nazirova, Nodira Samigova, Kamola Usmonova*