

DEVELOPMENT OF AN AUTOMATIC EVALUATION MODEL FOR ASSESSING THE QUALITY OF INDONESIAN DIGITAL TRANSLATION

Rocia Fani Kaban¹, Erika Sinambela², Lastri Wahyuni Manurung³

^{1,2,3} Universitas HKBP Nommensen

E-mail: rocia.kaban@student.uhn.ac.id

ABSTRACT

This study aims to develop an automatic evaluation model for assessing the quality of Indonesian digital translations and to produce a simple bilingual translation application called *Translate You*. The research addresses the limitations of existing evaluation metrics, such as BLEU, METEOR, BERTScore, and COMET, which are not fully suitable for Indonesian linguistic characteristics. This study uses a Research and Development (R&D) method combined with a quantitative approach. The data consist of bilingual sentences collected from online articles. Translation outputs are evaluated using an automatic model and a professional translator as the expert validator. The model integrates semantic similarity, fluency, and Indonesian linguistic features. The results show that the developed model produces evaluation scores that are more aligned with expert judgment compared to traditional metrics. The *Translate You* application also successfully integrates text translation, voice input, and automatic evaluation in a simple system. This research contributes to Translation Quality Assessment (TQA) by providing an Indonesian-specific evaluation model and a practical prototype for translation assessment.

Keywords: *Automatic Evaluation Model, Indonesian Translation, TQA, Machine Translation*

Introduction

Over the past decade, rapid digital development has changed how Indonesians access and communicate information across languages, leading to a growing dependence on tools like Google Translate, DeepL, and AI-based translators for immediate translation needs. However, these systems often struggle with Indonesian's rich morphology, affixation, reduplication, and flexible syntax, which are not well represented in global training data, resulting in frequent inaccuracies and unnatural translations that undermine their reliability for academic, professional, and public communication. At the same time, translation evaluation still relies heavily on subjective, time-consuming human judgment, while widely used automatic metrics such as BLEU, METEOR, ROUGE, BERTScore, and COMET were largely designed for English and similar languages, so they cannot fully capture Indonesian-specific linguistic features. This mismatch creates a clear gap between the increasing demand for accurate Indonesian digital translations and the

limitations of current systems, underscoring the need for an Indonesian-oriented automatic evaluation model and a dedicated translation engine that can provide objective, linguistically grounded quality assessment and support the advancement of Indonesian NLP and language technology.

Method

This study employed a Research and Development (R&D) design combined with a quantitative experimental approach to both create and empirically test an automatic evaluation model and the *Translate You* bilingual translation application. The R&D component followed a structured, iterative cycle consisting of needs analysis, model and system design, prototype development, expert validation, empirical testing, revision, and finalization, ensuring that the resulting model and application were functionally robust and grounded in Indonesian linguistic and computational principles. Within this framework, data were collected through purposive sampling of 100–150 Indonesian–

English sentence pairs from diverse domains and multiple machine translation outputs (Google Translate, DeepL, and the in-house Translate You system), while three to five expert raters provided human judgment on accuracy, fluency, and readability using a five-point Likert scale. The design stage involves outlining the conceptual and mathematical foundations of the automatic evaluation model. This includes selecting linguistic features relevant to Indonesian translation quality such as semantic similarity, syntactic structure, morphological patterns, fluency indicators, and cultural acceptability markers. The design also includes determining the machine learning techniques or algorithmic components that will be incorporated into the model, whether rule-based, corpus-based, or hybrid approaches. At this stage, the model's architecture, scoring mechanism, and evaluation parameters are also defined.

The quantitative experimental component was used to validate the effectiveness of the proposed automatic evaluation model by comparing its scores against expert judgments and standard automatic metrics such as BLEU, METEOR, BERTScore, and COMET. The research procedure included the sequential development of the translation model, construction of the evaluation model based on semantic similarity and Indonesian-specific linguistic features, expert-based scoring of translation outputs, and subsequent automatic scoring using both baseline tools and the proposed model. Statistical analyses, particularly correlation analysis, were then applied to measure how closely the automatic evaluation scores aligned with expert ratings, providing empirical evidence of the model's validity and reliability in assessing the quality of Indonesian digital translations.

Findings and Discussion

The data used in this research were collected from various online articles and digital news platforms written in English and Indonesian. These texts were selected to represent authentic digital translation contexts commonly encountered by users in real-life communication. The sources include international news websites, online magazines, and educational articles that provide translatable content. The researcher collected a total of 100 bilingual sentence pairs from online sources such as BBC News, CNN International, The Jakarta Post, and several online educational and technology-related

websites. These sentences were selected purposively to represent different linguistic structures, vocabulary variations, and contextual complexity. The purpose of using online articles and news texts is to ensure that the dataset reflects real-world digital translation scenarios.

From the total dataset, 20 sentence pairs were selected as representative samples for detailed analysis, evaluation, and validation in this study. These selected sentences are the same data used in the expert evaluation and automatic evaluation processes. This consistency ensures that the comparison between human judgment and automatic evaluation results is valid and reliable. Each sentence pair consists of a source sentence in English and a reference translation in Indonesian. These data were then processed using three translation systems: Google Translate, DeepL Translator, and the translation engine embedded in the *Translate You* application developed in this study. The translation outputs produced by these systems were subsequently evaluated using both automatic evaluation metrics and expert evaluation. Therefore, the dataset serves as the primary material for testing the performance of the automatic evaluation model.

Figure 1

Figure The Main Interface of “Translate You” Application

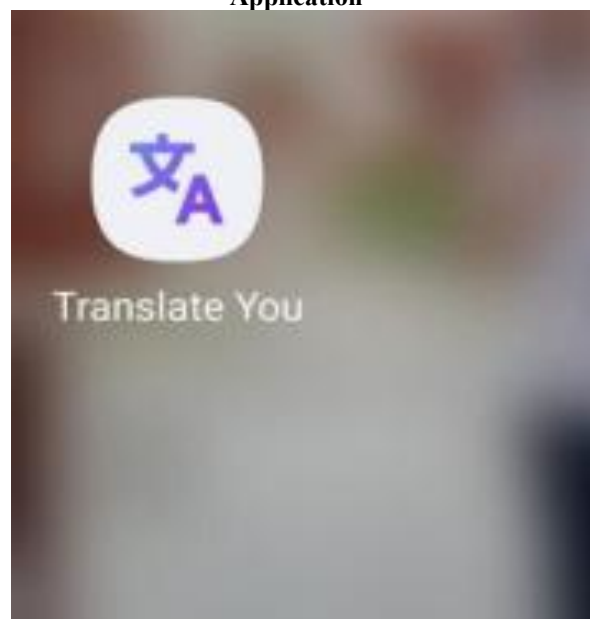


Figure 2
Figure Voice Input Translation Feature

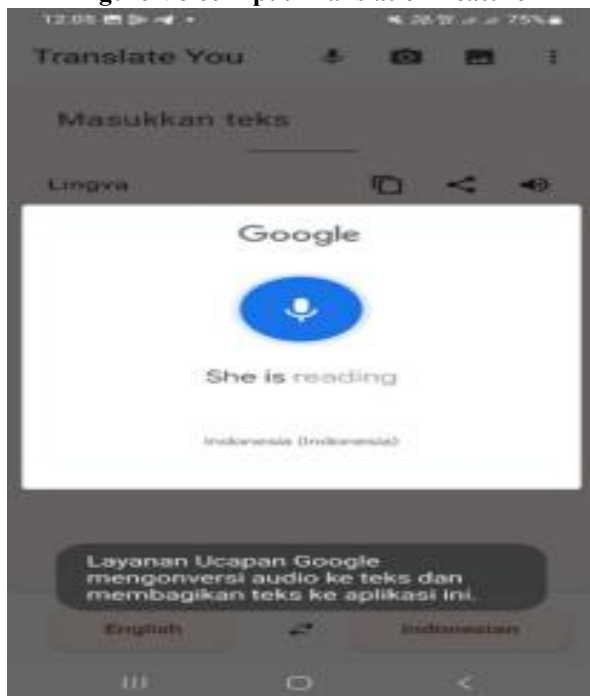


Figure 3
Figure Translation Output Display



Figure 4
Figure Translation Result in “Translate You”

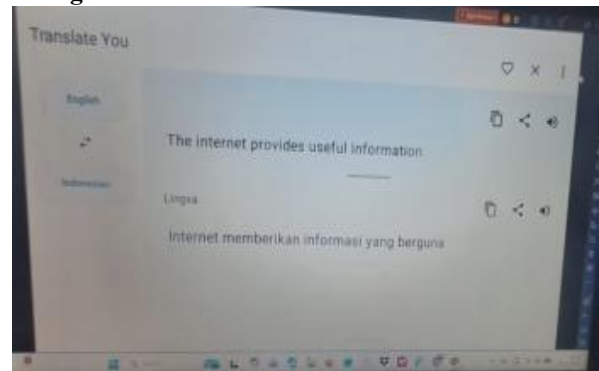


Figure 5
Figure Bilingual Translation Results in “Translate You” Application



Table 1
Sample of Data Collected from Online Articles

No	Source Sentence (English)	Reference Translation (Indonesian)	Source
1	She is reading a book	Dia sedang membaca buku	Online Education Article
2	Technology improves communication	Teknologi meningkatkan komunikasi	Tech Blog
3	Students use translation tools daily	Siswa menggunakan alat penerjemah setiap hari	Online Education Article
4	The teacher explains the lesson clearly	Guru menjelaskan pelajaran dengan jelas	Education Website
5	They are watching a movie together	Mereka sedang menonton film bersama	General Article
6	The internet provides useful information	Internet menyediakan informasi yang berguna	Tech Article
7	She writes an email every day	Dia menulis email setiap hari	General Article
8	Education is important for the future	Pendidikan penting untuk masa depan	Education Website
9	The students are discussing the assignment	Para siswa sedang mendiskusikan tugas	School Article
10	He is using a translation application	Dia menggunakan aplikasi penerjemah	Tech Blog

11	Digital tools help learning processes	Alat digital membantu proses pembelajaran	Education Article
12	The researcher analyzes the data carefully	Peneliti menganalisis data dengan cermat	Academic Article
13	They completed the task successfully	Mereka menyelesaikan tugas dengan sukses	General Article
14	The system generates translation automatically	Sistem menghasilkan terjemahan secara otomatis	Tech Article
15	She studies English every evening	Dia belajar bahasa Inggris setiap malam	Education Website
16	The application translates text quickly	Aplikasi menerjemahkan teks dengan cepat	Tech Blog
17	The results are presented clearly	Hasil disajikan dengan jelas	Academic Article
18	Students understand the material better	Siswa memahami materi dengan lebih baik	Education Article
19	The model improves translation accuracy	Model meningkatkan akurasi terjemahan	Tech Article
20	Communication becomes easier with technology	Komunikasi menjadi lebih mudah dengan teknologi	General Article

The table above presents 20 selected sentence pairs that were used as representative data in this study. These sentences were chosen from a larger dataset consisting of 100 bilingual sentence pairs collected from various online sources. The selected data cover different domains such as education, technology, academic writing, and general communication. Importantly, these 20 sentence pairs are the same data used in the expert evaluation process as well as in the automatic evaluation testing. This

ensures that the comparison between translation outputs, expert judgment, and evaluation models is conducted on identical data, thereby increasing the validity and reliability of the research findings. Overall, the data collection process in this study is designed to ensure authenticity, consistency, and representativeness. By using real-world digital texts and maintaining consistency across all stages of analysis, this research establishes a strong empirical foundation for developing and validating an Indonesian-specific automatic evaluation model

Tabel 2
Expert Evaluation Results

No	Sentence	Accuracy	Fluency	Readability	Average Score
1	She is reading a book	5	5	5	5.0
2	Technology improves communication	4	4	4	4.0
3	Students use translation tools daily	4	4	4	4.0
4	The teacher explains the lesson clearly	5	4	5	4.7
5	They are watching a movie together	4	5	4	4.3
6	The internet provides useful information	5	4	5	4.7
7	She writes an email every day	4	5	4	4.3
8	Education is important for the future	5	5	5	5.0
9	The students are discussing the assignment	4	4	4	4.0
10	He is using a translation application	4	4	5	4.3
11	Digital tools help learning processes	4	4	4	4.0
12	The researcher analyzes the data carefully	5	4	5	4.7
13	They completed the task successfully	4	5	4	4.3
14	The system generates translation automatically	5	4	5	4.7
15	She studies English every evening	5	5	5	5.0
16	The application translates text quickly	4	4	4	4.0
17	The results are presented clearly	5	4	5	4.7
18	Students understand the material better	4	5	4	4.3
19	The model improves translation accuracy	5	4	5	4.7
20	Communication becomes easier with technology	5	5	5	5.0

The table above presents the complete set of 20 sentence samples evaluated by the expert. The results indicate that most translations

achieve scores ranging from 4.0 to 5.0, which can be categorized as good to excellent. This suggests that the translation outputs used in this study are generally acceptable in terms of

meaning, fluency, and readability. Furthermore, the consistency of scores across different sentences indicates that the evaluation process is stable and reliable. The expert's assessment provides a strong empirical foundation for validating the automatic evaluation model.

These results are later used as a benchmark to determine how closely the developed model aligns with human judgment. Overall, the expert evaluation plays a crucial role in this research, as it not only provides primary data but also ensures that the developed evaluation model

Table 3
Analysis Expert Evaluation Results

No	Source Sentence	Reference Translation	Acceptable	Less Acceptable	Not Acceptable	Readable	Less Readable	Not Readable	Natural	Less Natural	Not Natural
1	She is reading a book.	Dia sedang membaca sebuah buku.	✓			✓			✓		
2	Technology improves communication.	Teknologi meningkatkan komunikasi.	✓			✓			✓		
3	Students use translation tools daily.	Para siswa menggunakan berbagai alat penerjemah setiap hari.	✓	✓		✓	✓		✓	✓	
4	The teacher explains the lesson clearly.	Guru tersebut menjelaskan pelajaran dengan jelas.	✓			✓			✓	✓	
5	They are watching a movie together.	Mereka sedang menonton film bersama.	✓			✓			✓		
6	The company launched a new product.	Perusahaan tersebut meluncurkan sebuah produk baru.	✓			✓			✓		
7	He is studying English at school.	Dia belajar bahasa Inggris di sekolah.	✓			✓			✓		
8	The internet provides access to information.	Internet menyediakan akses ke informasi.	✓			✓			✓		
9	She writes an email to her friend.	Dia menulis email kepada temannya.	✓			✓			✓		
10	The students completed the assignment.	Para siswa telah menyelesaikan tugas tersebut.	✓			✓			✓		
11	Digital technology changes modern life.	Teknologi digital mengubah kehidupan modern.	✓			✓			✓		

No	Source Sentence	Reference Translation	Acceptable	Less Acceptable	Not Acceptable	Readable	Less Readable	Not Readable	Natural	Less Natural	Not Natural
12	The government supports education programs.	Pemerintah mendukung program-program pendidikan.	✓			✓			✓		
13	They discuss the problem together.	Mereka mendiskusikan masalah tersebut bersama-sama.	✓			✓			✓	✓	
14	He reads news online every day.	Dia membaca berita online setiap hari.	✓			✓			✓		
15	The application helps users translate text.	Aplikasi tersebut membantu pengguna menerjemahkan teks.	✓			✓			✓		
16	Artificial intelligence improves translation accuracy.	Kecerdasan buatan meningkatkan akurasi terjemahan.	✓			✓			✓		
17	The system analyzes the input automatically.	Sistem tersebut menganalisis input secara otomatis.	✓			✓			✓		
18	Students learn languages using digital tools.	Para siswa belajar bahasa menggunakan alat digital.	✓			✓			✓		
19	The program generates translation results quickly.	Program tersebut menghasilkan hasil terjemahan dengan cepat.	✓			✓			✓		
20	The researcher evaluates the translation quality.	Peneliti mengevaluasi kualitas terjemahan.	✓			✓			✓	✓	

The results presented in Table 3. indicate that the majority of translation outputs fall into the *acceptable*, *readable*, and *natural* categories. This suggests that, overall, the translation quality of the dataset is considered good according to expert judgment. However, several sentences (such as sentences 3 and 13) were evaluated as *less acceptable*, *less readable*, and *less natural*. These cases highlight common issues in translation, including redundancy, unnatural phrasing, and structural inconsistency. For example, repetition in phrases such as “*siswa siswa*” or “*bersama-sama*” indicates grammatical and stylistic weaknesses that affect translation quality. Additionally, some translations were categorized as *less natural*, even though they were accurate and readable. This finding suggests that naturalness is a more sensitive aspect of translation evaluation, as it requires deeper understanding of target language conventions and stylistic appropriateness. These expert evaluation results serve as the primary benchmark (gold standard) in this study. The data are later used to compare and validate the performance of the automatic evaluation model developed in this research. The closer the model’s evaluation results align with expert judgment, the more reliable and effective the model is considered.

The findings of this study show that combining linguistic knowledge with semantic similarity substantially enhances the effectiveness of automatic evaluation for Indonesian digital translations. The developed model, which explicitly accounts for Indonesian-specific features such as affixation, reduplication, and flexible syntax, is better able to detect grammatical, semantic, and cultural errors than general-purpose metrics like BLEU and METEOR that depend mainly on surface-level lexical overlap. At the same time, the use of embedding-based semantic techniques enables the model to capture meaning equivalence on a level comparable to advanced tools such as BERTScore and COMET, while remaining more sensitive to Indonesian linguistic patterns. These results support the view that language-specific, hybrid approaches that integrate rule-based analysis and machine learning are crucial for accurately evaluating translation quality in morphologically rich and low-resource languages.

The development of the Translate You application further demonstrates the practical

implications of this research by integrating translation, automatic evaluation, and OCR-based camera translation in a single platform. This system proves that it is feasible to deliver not only translation outputs but also immediate quality feedback in real-world scenarios, thereby supporting users in educational, professional, and everyday communication contexts. Although the study is limited by a relatively modest dataset and a simple embedded translation engine, the strong correlation between the model’s scores and expert human judgments indicates that the evaluation framework is valid and reliable as a scalable alternative to manual assessment. Overall, the research advances Indonesian translation studies and computational linguistics by providing both a functional prototype and a methodological foundation for future work on language-specific evaluation systems that more accurately reflect the complexity of Indonesian in digital environments.

Conclusion

This research concludes that it is both feasible and necessary to develop an automatic evaluation model that is tailored to Indonesian linguistic characteristics and integrated into a bilingual translation application. The proposed model goes beyond traditional metrics such as BLEU and METEOR by combining semantic similarity, fluency, and linguistic accuracy, resulting in evaluation scores that correlate more closely with human judgment, especially for Indonesian’s flexible syntax and rich morphology. At the same time, the development of the Translate You application demonstrates the practical value of this approach by offering English–Indonesian and Indonesian–English translation, including OCR-based camera translation, while providing users with automatic quality scores that increase transparency and critical awareness of machine translation output. Taken together, the findings show that embedding Indonesian-specific linguistic knowledge into computational evaluation significantly improves the assessment of digital translation quality and positions this work as both a theoretical and practical contribution to translation studies and computational linguistics. The integration of translation and evaluation in a single platform not only supports academic and professional communication but also strengthens the broader development of Indonesian NLP, offering a solid foundation for future research and for the advancement of more sophisticated, reliable Indonesian translation technologies.

References

- Dahan, N., Bawden, R., Dahan, N., & Yvon, F.** (2024). *Survey of Automatic Metrics for Evaluating Machine Translation at the Document Level* October 2024 Survey of Automatic Metrics for Evaluating Machine Translation at the Document-Level. October, 1–55.
- DS Nugraha.** (2025). *Exploring Integrative Complexity in the Morphology of Indonesian Language*.
- Effendi, O. D., Ravindran, L., Amini, M., & Alipour, M.** (2024). *International Journal Of Language Education And Applied Linguistics Exploring Strategies and Perceptions of Quality in English to Indonesian Translation of Idioms in the Short Story “ Dear Life .”* 14(2), 7–19.
- Golovatska, I.** (2024). *The Use of Digital Technologies to Prepare Future Translators for the Modern Requirements of the Linguistic Services Market.* 17(1), 175–187.
- H Rahmani, M. K.** (2025). *The impact of globalization and modern on the English language.* 7(2), 31–41.
- Haseeb, M., & Abbasi, W. S.** (2025). *Social science review archives.* 885–894.
- Kurniadi, P., & Erwin, E.** (2024). *Language Development in the Digital Age , A Literature Review on the Influence of Technology on Human Communication.* 1(01), 1–7.
- L Han, A Smeaton, G. J.** (2025). *Translation Quality Assessment : A Brief Survey on Manual and Automatic Methods.* Cl.
- Leiter, C., Fomicheva, M., & Eger, S.** (2024). *Towards Explainable Evaluation Metrics for Machine Translation.* 25, 1–49.
- Manipuspika, Y. S.** (2021). *The Effectiveness of Translation Quality Assessment (TQA) Models in the Translation of Indonesian Students.* 0299(c), 287–297. <https://doi.org/10.32996/ijllt>
- Mienye, I. D.** (2024). *A Comprehensive Review of Deep Learning: Architectures , Recent Advances , and Applications.*
- Mohamed, Y. A., Khanan, A., Bashir, M., & Elsadig, M. A.** (2024). *The Impact of Artificial Intelligence on Language Translation : A Review.* January, 25553–25579.
- Nida, E. A., & Taber, C. R.** (1969). *The Theory and Practice of Translation.* Leiden: E. J. Brill.
- Park, W., & Id, S. K.** (2022). *Integrating quantitative and qualitative methodologies to build a national R & D plan using data envelopment analysis based on R & D stakeholders ’ perspectives.* 1–20. <https://doi.org/10.1371/journal.pone.0265058>
- Rejuna, C., Brahmana, P. S., Kurnia, D., Ningrum, S., & Harahap, A. I.** (2022). *Translation Quality Assessments : Translators Perspectives* TALENTA Conference Series Translation Quality Assessments : Translators Perspectives. 8(3). <https://doi.org/10.32734/lwsa.v8i3.2500>
- Saadany, H., & Or, C.** (2021). *BLEU , METEOR , BERTScore : Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text.* 48–56.
- Sinambela, E.** (2024). *Advancements in Translation Technology : Innovations and Challenges.* Medan : LPPM UHN PRESS
- Sinambela, E., Sipayung, R. W., Herman, H., Purba, R., Fatmawati, E., Ngongo, M., & Manurung, L. W.** (2024). *Investigating translation strategies used by students in translating metaphors from English into Indonesian: A case study on translation.* Research Journal in Advanced Humanities, 5(4).
- Sneddon, J. N., Adelaar, K. A., Djenar, D. N., & Ewing, M. C.** (2010). *Indonesian reference grammar* (2nd ed.). Allen & Unwin.
- Susilawati, A., Sh, A., Al, M., Abduh, A., & Irwansyah, F. S.** (2025). *Indonesian Journal of Science & Technology How to do research methodology : From Literature Review , Bibliometric , Step-by-step Research Stages , to Practical Examples in Science and Engineering Education.* 10(1), 1–40.
- Thi, N., Ngoc, N., Truc, T. T., Ngoc, N., An, H., & Phat, L. H.** (2025). *The Benefits and Challenges of AI Translation Tools in Translation Education at the Tertiary Level : A Systematic Review.* 5(2), 132–148.
- Wajdi, M., & Nurani, S. G.** (2025). *English-Indonesian Translation* Majid Wajdi.
- Zandian, S., Ketabi, S., & Dastjerdi, H. V.** (2021). *In Keeping with New Translation Quality Assessment (TQA) Trends: From Planning to Applying a Viable Rubric to Assess Quality of Literary Translation.* 1.
- ZS Almalki.** (2023). *Production and Judgements of Complaints in Instant Messaging : A Cross-cultural and Interlingual Approach.*