Journal of Literature and Art Studies, June 2025, Vol. 15, No. 6, 483-489
doi: 10.17265/2159-5836/2025.06.006

A Review of Automatic Pre-editing Approaches in Machine Translation

WANG Jun-song, MENG Ya-qi
Northwestern Polytechnical University, Xi'an, China

WANG Ai-qing
University of Liverpool, Liverpool, United Kingdom

With the development of machine translation technology, automatic pre-editing has attracted increasing research attention for its important role in improving translation quality and efficiency. This study utilizes UAM Corpus Tool 3.0 to annotate and categorize 99 key publications between 1992 and 2024, tracing the research paths and technological evolution of automatic pre-translation editing. The study finds that current approaches can be classified into four categories: controlled language-based approaches, text simplification approaches, interlingua-based approaches, and large language model-driven approaches. By critically examining their technical features and applicability in various contexts, this review aims to provide valuable insights to guide the future optimization and expansion of pre-translation editing systems.

Keywords: automatic pre-editing, machine translation, controlled language, text simplification, large language models

Introduction

With the rapid advancement of machine translation (MT) technologies, automatic pre-editing has attracted increasing attention from both academia and industry as a critical step for improving translation quality and reducing post-editing effort. Pre-editing denotes the purposeful adjustment of source texts before machine translation, aiming to optimize the quality of the generated output. Automatic pre-editing, in particular, entails the utilization of automated tools to perform such adjustments prior to the production of MT output. Typically, this process involves normalization, simplification, and structural optimization, aiming to enhance the source text's readability and machinability. In essence, it intended to improve what is often termed the "source text quality" for translation. Compared with post-editing focusing on output correction, pre-editing emphasizes control over the input, which can significantly improve translation consistency and predictability while reducing human intervention.

Despite its growing importance, research on automatic pre-editing remains limited. This review presents a thorough review of the current state of research, analyzing the main approaches and their application scenarios,

Acknowledgments: This research was supported by Chunhui Collaborative Research Project funded by the Ministry of Education of China [Grant No. 202200490] and Humanities and Social Sciences Research Project funded by the Ministry of Education of China [Grant No. 23YJAZH139].

WANG Jun-song, Ph.D., Associate Professor, School of Foreign Languages, Northwestern Polytechnical University.
MENG Ya-qi, Postgraduate Student, School of Foreign Languages, Northwestern Polytechnical University.
WANG Ai-qing, Senior Lecturer, Department of Languages, Cultures and Film, University of Liverpool.

and evaluating their strengths and limitations. It also examines the impact of pre-editing on machine translation performance, considering both its benefits and challenges. Additionally, the review explores strategies for optimizing pre-editing techniques. The aim is to provide insights that can inform the continued development of pre-editing in this emerging area.

Research Design

Data Sources

This review primarily draws on literature indexed in Google Scholar and Web of Science, covering the period from 1992 to 2024. The search utilized keywords such as “pre-editing,” “pre-editing systems,” “controlled language,” “text simplification,” and “machine translation.” Following an initial screening, the snowballing method was applied to examine the references of relevant publications, expanding the sample further. After excluding non-academic materials, such as announcements, newsletters, calls for papers, and book reviews, 99 of the 124 publications met the inclusion criteria and were selected for analysis.

Research Methods

To facilitate systematic categorization, this review utilizes the UAM Corpus Tool 3.0 to annotate and classify the selected literature based on key technical features of automatic pre-editing. This tool supports multi-level text encoding and visual statistical analysis, making it ideal for examining complex textual features in translation studies. A preliminary annotation scheme was developed, focusing on the core technical characteristics and methodologies of automatic pre-editing. The initial annotation was conducted independently by two researchers, followed by cross-checking and adjustments to ensure coding reliability. Based on this annotated corpus, comparative analysis, inductive reasoning, and clustering techniques were employed to identify patterns among different technological approaches, contributing to the development of a potential classification framework.

Results and Discussion

Through inductive analysis of the collected literature, this review identifies four technical approaches to automatic pre-editing. The following sections will systematically examine the research trends and technical features, followed by an analysis of the strengths and limitations of different approaches.

Research Trends in Automatic Pre-editing Approaches

This study systematically reviews the current state of research on automatic pre-editing approaches by categorizing and quantifying the existing literature. Figure 1 presents the distribution of the four major types of approaches identified in the reviewed publications: Controlled Language–Based Approaches (CL-based approaches), Text Simplification Approaches, Interlingua-Based Approaches, and Large Language Model–Driven Approaches (LLM-driven approaches).

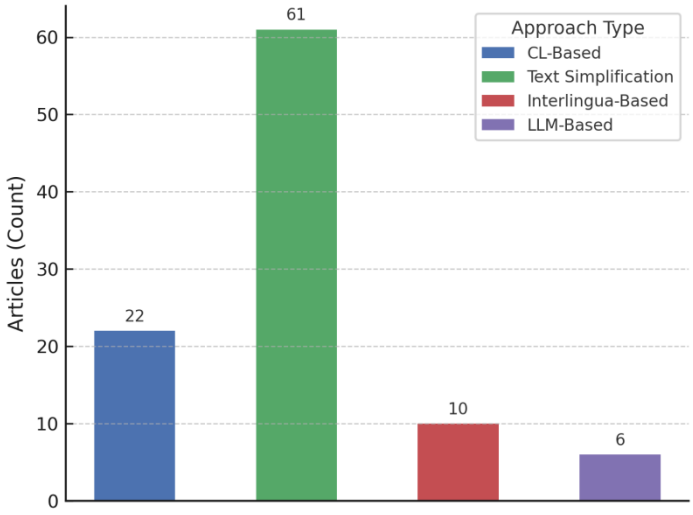

Figure 1. Distribution of Research Articles on Pre-Editing Approaches (1992-2024).

As shown in Figure 1, text simplification approaches dominate the field with 61 publications, reflecting strong research interest in enhancing the processability and adaptability of source texts. CL-based approaches follow in second place with 22 studies, and are primarily applied in fields with strict regulatory or stylistic requirements. Interlingua-based approaches are featured in 10 publications, indicating limited attention and mostly explored in multilingual or low-resource language contexts. In contrast, LLM-driven approaches are represented in only 6 studies, suggesting that this area of research is still in its early stages despite its significant potential. Overall, traditional approaches remain the primary focus, while emerging approaches like LLM-driven pre-editing are gradually gaining momentum.

Analysis of Automatic Pre-editing Approaches

After presenting the overall distribution of four translation pre-editing approaches, this section will analyze and introduce these approaches in detail.

(1) CL-Based approaches

CL-based approaches enhance the standardization of source texts by establishing linguistic norms, such as limiting sentence length, avoiding ambiguity, and standardizing terminology. These constraints help reduce ambiguity and irregular expressions, thereby simplifying the machine translation process. Initially applied in high-stakes domains like aviation, healthcare, and law, representative systems include the KANT system, which uses rule-based analysis and structural rewriting for syntactic parsing and terminology normalization. Another example is Acrolinx, which leverages a knowledge base to detect stylistic deviations and recommend standardized expressions. The ACCEPT project integrates pre-translation, in-process, and post-translation editing specifically for user-generated content. The main strengths of these approaches are their high consistency and controllability, making them particularly well-suited for technical documentation, contracts, and other specialized texts. However, several limitations are evident. Controlled language has a narrow scope of application, high system maintenance costs, and limited flexibility for open-domain content. The initial construction of rule databases is labor-intensive, and while long-term operational costs are manageable, scalability in cross-domain applications remains limited.

(2) Text simplification approaches

Text simplification-based pre-editing seeks to improve the readability and translatability of source texts by reducing linguistic complexity through techniques like lexical substitution, syntactic splitting, and structural rewriting. It is commonly used in educational materials, technical manuals, and user support content to enhance user experience and machine translation performance. These approaches generally follow three main approaches: (1) Rule-based approaches, such as the SPIDER system, which rely on manually crafted rewriting rules—offering transparency but limited coverage; (2) Classifier-based approaches, which use trained models to identify complex structures and apply template-based transformations; (3) Data-driven approaches, which leverage large parallel corpora to train neural models, with systems like SIMPLIFICA and LexiSiS serving as key examples. These approaches are highly adaptable and easy to implement, particularly for content intended for a general audience. However, they face challenges such as semantic loss, weakened inter-sentential logic, inconsistent terminology, and heavy dependence on large-scale corpora.

(3) Interlingua-based approaches

Interlingua-based pre-editing approaches involve the use of a manually constructed intermediary structure, or "third language," that transforms the original text into a neutral expression before it is processed by the machine translation system. These approaches aim to address issues of semantic loss and syntactic mismatch in direct translation between low-resource languages, often applied in multilingual systems and specialized domain translation tasks. Typical strategies include syntactic reordering approach, which restructures Chinese sentences according to English syntax; the image-anchor-based visual intermediary approach; and the pivot-language bridging approach, where a third language, such as English, facilitates translation between under-resourced language pairs, like Chinese and Vietnamese. However, these approaches are technically complex, requiring substantial resources for modeling and preprocessing. Their ability to handle complex sentence structures is limited, and they often lead to semantic weakening and residual ambiguity.

(4) LLM-Driven approaches

In recent years, large language models have shown significant potential in pre-editing tasks. These approaches leverage the contextual understanding and generation capabilities of pre-trained models to perform text reconstruction, disambiguation, style adjustment, and terminology standardization. Compared to traditional approaches, LLM-driven approaches stand out for their wide applicability and intelligent behavior. Through prompt engineering, users can customize linguistic style, simplification level, and target domain, enabling personalized pre-editing. LLMs are particularly suited for high-demand fields such as law, medicine, and technology, where terminology consistency and cultural adaptability are critical. However, these approaches face challenges, including high computational costs, limited controllability during text generation, and the inherent "black-box" nature of LLMs. In high-risk or high-consistency domains, human intervention and rule-based strategies remain essential to ensure translation quality. Additionally, the deployment of such systems encounters technical barriers related to real-time performance and stability.

Multidimensional Analysis of Automatic Pre-editing Approaches

To provide comparison of the performance of these four automatic pre-editing approaches across multiple dimensions, this review establishes a performance framework based on five core indicators: accuracy,

applicability, flexibility, technical complexity, and cost. A radar chart is used to visualize the comparative analysis.

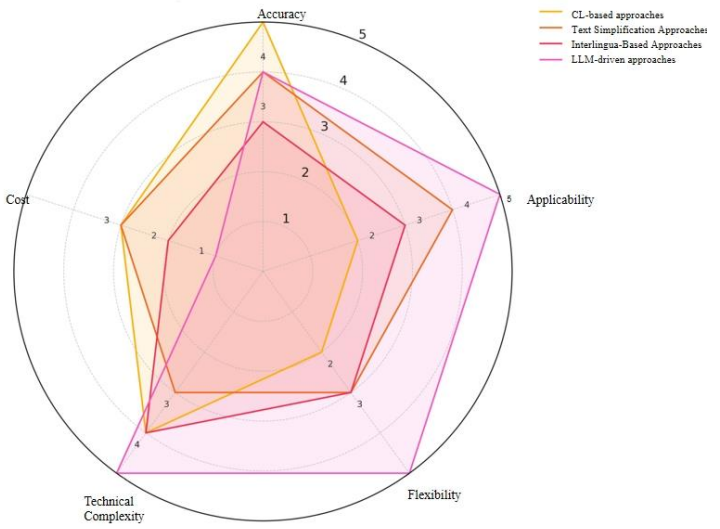

Figure 2. Comparative Analysis of Four Types of Automatic Pre-Editing Technologies.

Figure 2 illustrates the performances of four automatic pre-editing approaches across five core dimensions: accuracy, applicability, flexibility, technical complexity, and cost. The scoring range is from 1 to 5, with higher values indicating better performance. As shown, CL-based approaches excel in accuracy, making them well-suited for technical documentation tasks. In contrast, LLM-driven approaches offer clear advantages in flexibility and scope of application, though they come with higher computational resource demands and operating costs.

In terms of accuracy and stability, CL-based approaches excel in terminology consistency and syntactic regularity, making them the most reliable. In contrast, while LLMs are highly intelligent, their output controllability remains uncertain, particularly in high-precision tasks, where additional mechanisms are required. When it comes to scope and task compatibility, LLM-driven and text simplification approaches are versatile, capable of handling diverse styles and domains. However, CL-based and interlingua-based approaches are better suited to specific contexts, such as technical documents and low-resource language environments.

Concerning flexibility, LLM-driven approaches demonstrate superior performance compared to the other approaches, offering dynamic adaptation and contextual coherence through prompt engineering. In contrast, rule-driven approaches tend to be rigid when handling complex linguistic phenomena due to their limited language coverage. From a technical perspective, both interlingua-based and LLM-driven approaches are more complex, requiring semantic reconstruction and substantial computational resources. Specifically, interlingua-based approaches require the creation of independent corpora for each language pair, which can result in linguistic loss and increased system maintenance.

When it comes to cost, CL-based approaches require substantial upfront investment but offer stable long-term operation, making them suitable for prolonged deployment. Text simplification in rule-based settings

involves moderate investment, while data-driven models depend heavily on large annotated corpora, leading to higher costs. LLM-driven approaches incur significant deployment and API invocation expenses, which pose major barriers to wider adoption.

Overall, CL-based approaches offer high stability but are limited in applicability; text simplification approaches provide strong generalizability at a moderate cost; interlingua-based approaches are well-suited for low-resource languages but come with high technical barriers; and LLM-driven approaches offer the highest level of intelligence and potential, though they require significant computational resources. Each approach has distinct advantages and limitations, and selection should be made flexibly based on specific translation needs and application contexts. It is recommended to integrate multiple strategies and optimize systems in practice to improve both pre-editing efficiency and translation quality. Future pre-editing systems should leverage the complementary strengths of different approaches, dynamically selecting the most suitable methods based on task types and resource conditions to achieve an optimal balance between translation efficiency and quality.

Conclusion

With the advancement of neural machine translation, pre-translation processing has become increasingly recognized as a key factor in improving translation quality and ensuring system robustness. This review focuses on automatic pre-editing approaches, systematically outlining their core methods and implementation strategies from a classificatory perspective. It provides a comparison of the conditions and performance characteristics of four approaches, and, based on this analysis, addresses current technological challenges and explores potential future developments. Looking forward, the development of automatic pre-editing is set to advance towards greater intelligence, modularity, and easier accessibility. The integration of LLMs, in particular, offers a significant breakthrough. By leveraging their robust semantic understanding and contextual modeling abilities, LLMs can enhance style control, adapt to context, and resolve ambiguities through prompt engineering, thus boosting the flexibility and intelligence of automated preprocessing. At the same time, human-machine collaboration mechanisms are likely to evolve. Future systems should prioritize interactive design and user involvement by incorporating features like visual parameter adjustment and personalized feedback loops. These elements would make the pre-editing process more controllable and participatory, thereby enhancing the role of translators in automated workflows.

References

Awasthi, A., Gupta, N., Samanta, B., Dave, S., & Sunita, S. (2022). Bootstrapping multilingual semantic parsers using large language models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 1*.

Barreiro, A. (2011). SPIDER: A system for paraphrasing in document editing and revision—Applicability in machine translation pre-editing. *International Conference on Intelligent Text Processing and Computational Linguistics.* Berlin, Heidelberg: Springer.

Bott, S., & Saggion, H. (2012). Automatic simplification of Spanish text for e-accessibility. *Computers Helping People with Special Needs. ICCHP 2012*. Berlin, Heidelberg: Springer.

Bott, S., Rello, L., Drndarević, B., & Saggion, H. (2012). Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. *Proceedings of COLING*.

CHEN, S., JIN, Q., & FU, J. (2019). From words to sentences: A progressive learning approach for zero-resource machine translation with visual pivots. *arXiv preprint arXiv:1906.00872*.

Devlin, S., & Unthank, G. (2006). Helping aphasic people process online information. *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*.

FENG, Q., & GAO, L. (2017). The influence of controlled language-based pre-editing on machine translation (Jiyu shoukong yuyan de yiqian bianji dui jiqi fanyi de yingxiang 基于受控语言的译前编辑对机器翻译的影响). *Contemporary Foreign Languages Studies*, (02).

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *Proceedings of the 1st Workshop on Neural Machine Translation*.

Mitamura, T. (1999). Controlled language for multilingual machine translation. *Proceedings of MT Summit VII*.

Mitamura, T., & Nyberg, E. (2001). Automatic rewriting for controlled language translation. *Proceedings of the NLPRS2001 Workshop on Automatic Paraphrasing*.

Mitamura, T., Nyberg, E., & Nino, M. (1999). The KANT system: Fast, accurate, high-quality translation in practical domains. *Proceedings of COLING 1992, 3*.

MIN, S., LYU, X., & Sulem, E. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Pabst, H., & Siegel, M. (2009). Easier, faster communication in international teams: Lower translation costs with controlled language checking. *Proceedings of the tekom Annual Conference 2009*.

QIAN, M., & KONG, C. (2024). Enabling human-centered machine translation using concept-based large language model prompting and translation memory. *International Conference on Human-Computer Interaction. Cham: Springer Nature Switzerland*.

QIAN, M., WU, H., YANG, L., & WAN, A. (2023). Augmented machine translation enabled by GPT-4: Performance evaluation on human-machine teaming approaches. *Proceedings of the First NLP4TIA Workshop*.

Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., & Bourg, L. (2011). Text simplification in SIMPLEXT: Making texts more accessible. *Procesamiento del Lenguaje Natural*.

Saggion, H., Bott, S., & Rello, L. (2013). Comparing resources for Spanish lexical simplification. *Statistical Language and Speech Processing: Proceedings of SLSP 2013.* Berlin, Heidelberg: Springer.

Scarton, C., Oliveira, M., Candido Jr, A., Gasperin, C., & Aluísio, S. (2010). SIMPLIFICA: A tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. *Proceedings of the NAACL HLT 2010 Demonstration Session*.

Seretan, V., Roturier, J., Silva, D., & Bouillon, P. (2014). The ACCEPT Portal: An online framework for the pre-editing and post-editing of user-generated content. *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*.

Štajner, S., Calixto, I., & Saggion, H. (2015). Automatic text simplification for Spanish: Comparative evaluation of various simplification strategies. *Proceedings of RANLP*.

SUN, Y., O'Brien, S., O'Hagan, M., & Hollowood, F. (2010). A novel statistical pre-processing model for rule-based machine translation systems. *Proceedings of the 14th EAMT Conference*.

Tyagi, S., Chopra, D., Mathur, I., & Joshi, N. (2015). Classifier-based text simplification for improved machine translation. *Proceedings of the International Conference on Advances in Computer Engineering and Applications*.

Tyagi, S., Chopra, D., Mathur, I., & Joshi, N. (2015). Comparison of classifier-based approach with baseline for English–Hindi text simplification. *Proceedings of the International Conference on Computing, Communication & Automation*.

WEI, J., WANG, X., Schuurmans, D., Bosma, M., Ichter, B., XIA, F., CHI, E., LE, Q., & ZHOU, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

WU, H., & WANG, H. (2009). Revisiting pivot language approaches for machine translation. *Proceedings of ACL-IJCNLP 2009*.

XU, Y., & Seneff, S. (2008). Two-stage translation: A combined linguistic and statistical machine translation framework. *Proceedings of AMTA 2008*.