

Optimization and Construction of Cross-border E-commerce Product Descriptions Based on Error Correction of Machine Translation

Ying Du*

School of Foreign Languages, Shanghai Technical Institute of Electronics and Information, Shanghai 201411, China

**Author to whom correspondence should be addressed.*

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Due to the diverse product descriptions, dense technical terms, and significant contextual differences in cross-border e-commerce, machine translation is prone to issues such as word ambiguity, grammatical incoherence, and pragmatic deviations, which affect the transmission of product information and the reusability of the data. This paper focuses on error identification, corpus correction, and resource construction, establishing a systematic vocabulary and terminology calibration system, improving the corpus processing procedures, correcting syntactic and semantic transformation rules, addressing target language environmental factors, and then conducting the collection, cleaning, comparison, and annotation of product description information. On this basis, a dataset consisting of the original text, machine translation, error types, corrected text, and quality grades is constructed, forming a high-level cross-border e-commerce product introduction corpus that can be retrieved, tracked, and dynamically maintained. The aim is to improve the accuracy and readability of the translations and to support the generation of product content, post-editing, and the use of multiple languages.

Keywords: Machine translation; Error correction; Cross-border e-commerce; Product description corpus

Online publication: June 26, 2026

1. Introduction

The product description of cross-border e-commerce plays a role in transmitting product attributes, facilitating platform search, and guiding consumer decision-making. The quality of its translation directly affects the presentation method of the product and the international market communication effect. When enterprises initially provide multilingual services, machine translation often has difficulty in correctly identifying professional terms, standardized numerals, grammatical structures, and cultural contexts, resulting in information distortion, awkward expressions, and keyword mismatches. The *Opinions on Expanding Cross-border E-commerce Export and Promoting Overseas Warehouse Construction* issued by nine

departments including the Ministry of Commerce require strengthening the international business capabilities of cross-border e-commerce enterprises and standardizing the description of information. Currently, some enterprises mainly use manual modification, but they lack unified standards for error classification, correction rules, corpus annotation, and storage methods. Therefore, this paper focuses on error identification and correction in machine translation and studies the optimization, alignment annotation, and construction methods of product description corpora, providing support for multilingual content management and corpus reuse.

2. The significance of optimizing and constructing product description corpora for cross-border e-commerce

Product descriptions for cross-border e-commerce need to provide product-related information and enable consumers to search for related product information through search engines. Their accuracy is related to whether the product attributes, features, and other relevant information can be accurately presented to consumers. By correcting machine translation errors and optimizing the corpus, errors in professional terms, sentence structure, semantic omission, and contextual imbalance can be identified, leading to unified terminology expressions, sentence conversions, quality control standards, and reducing the cost of secondary human proofreading^[1]. Constructing a product description corpus with clear sources, accurate corresponding original texts, and traceable correction processes and versions is conducive to accumulating high-frequency product language resources. Standardized corpora help with content migration between platforms, types, and markets across different countries, enhancing the uniformity, naturalness, and cultural adaptability of product descriptions, providing data support for cross-border e-commerce enterprises to conduct multilingual operations, brand marketing, and intelligent content production.

3. Problems existing in machine translation error correction for cross-border e-commerce product descriptions

3.1. Inaccurate translation of professional terms and terminology

Machine translation often makes errors in processing product materials, specifications, functions, and manufacturing information. It may have deviations in word meaning judgment and terminology selection, with some words having multiple meanings, which vary greatly in different fields such as clothing, electronic products, and household goods^[2]. There is a lack of stable mapping for brand names, abbreviations, measurement units, and combined parameters, resulting in omissions, mistranslations, and inconsistencies in context. The usage norms of title keywords, attribute words, and marketing words on different platforms vary, and the general terminology database cannot match specific scenarios, causing the translation to deviate from the common expressions used in the target market. Inaccurate terminology can lead to weakened product selling points, ambiguous parameters, and incorrect keyword combinations. Manual correction is greatly influenced by subjective factors and lacks error annotation information and version records. Similar errors repeatedly occur, so the correction results cannot form reusable corpus resources, affecting the quality and stability of product description corpora.

3.2. Deviation in syntactic and semantic conversion

Multilingual product information typically consists of a title, attribute section, key selling points, and

detailed descriptions. During machine translation, issues such as improper syntactic restructuring and semantic coherence gaps often occur. Elisions frequently occur in clauses, compound sentences, and parameter descriptions regarding conditional relationships, functional connections, and the degree of applicability, leading to unclear expression of critical information. Consistently, titles, selling points, and product details may exhibit inconsistent wording, semantic redundancy, or fragmented information. Such discrepancies compromise translation readability and credibility, potentially causing misinterpretations, inappropriate application assessments, and post-sale disputes. Current correction efforts primarily rely on manual sentence-by-sentence verification without rigorous syntactic or semantic annotation, failing to ensure stable error classification or provide high-quality, long-term reference materials—thereby undermining the standardization of corrected translations and the quality of subsequent database entries.

3.3. Contextual adaptation imbalance in target language context

For product descriptions targeting different markets, they involve consumer habits, cultural cognition, platform norms, and expression taboos. In the process of context migration, there is an imbalance in machine translation adaptation. If the exaggerated rhetorical styles, functional promises, and emotional colors in the source language are directly transplanted to the target language, it will result in stiff translations, offensive expressions, or exceeding the acceptance range of the target market ^[3]. Some translations, although having complete grammatical structures, fail to meet the reading habits and product information organization needs of local consumers, leading to weakened product selling points, chaotic brand tone, and increased platform review risks. Due to the lack of tag information such as region, platform, and audience in the existing corpus, it is impossible to accurately record the expression differences of product description corpora in different regions. Manual correction often relies on experience to handle, and contextual errors are difficult to classify. The corrected content also lacks consistency, affecting the value of product description corpora for cross-regional use.

4. Optimization paths for cross-border e-commerce product description corpora based on machine translation error correction

4.1. Establishing a vocabulary and terminology calibration system

Vocabulary and terminology calibration should be carried out from three aspects: product classification, attribute items, and the corresponding target market. Hierarchical sorting of names, materials, specifications, functions, processes, and usage scenarios should be conducted. Enterprises can collect platform pages, brand manuals, and industry norms to form a terminology table including source language terms, target language translations, applicable product categories, context explanations, and prohibited expressions. During the calibration stage, word frequency analysis, context correlation comparison, and manual verification methods are used to confirm the correct translation, and codes are set for polysemous words, abbreviations, measurement units, and brand-defined terms. The corrected words are simultaneously entered into the machine translation dictionary and corpus annotation system to ensure the uniformity of expressions in the title, attribute column, selling points, and detail page ^[4]. Taking a cross-border home appliance enterprise's insulated cup product as an example, information was collected on Amazon, its own platform, and product instructions. The terms were organized into a table based on material, capacity, insulation time, sealing form, and usage. For terms such as “food-grade stainless steel” and “double-wall vacuum insulation,” the Chinese source terms, English translations, applicable fields, and compliance labels should be unified. The system

locks the terms in the terminology database before machine translation and conducts consistency verification based on the title, selling points, and detail page after translation. Reviewers rewrite according to market conventions, and the modified content and version information are stored in the corpus. New descriptions for similar products can directly adopt term mapping, maintaining uniformity in the terms in the title, parameters, and selling points, and the calibration records are used for model training and corpus expansion.

4.2. Improving syntactic and semantic conversion rules

For different construction methods of titles, attribute columns, selling points, and detailed explanations, different information analysis and sentence generation rules can be set for each. Titles focus on compressing core attributes, attribute columns focus on parameter matching, selling point sentences focus on linking functional relationships, and detailed explanations maintain complete connections of conditions, objects, and results. For multi-parameter and continuous features, clause splitting, clause combining, and component mapping templates can be set, and the conversion results are verified based on the coverage of information points and consistency of semantic roles. Taking the description of a cross-border clothing enterprise's windproof coat as an example, the project team decomposed all the items such as fabric, waterproof grade, hat structure, and applicable seasons into individual information units, forming a three-level matrix of "material as service attribute, function as realization of function, climate as application limit." "Using multi-layered fabric can resist a small amount of raindrops and is suitable for daily office use in spring and autumn seasons" has been transformed into three types of English expressions: fabric description, mild waterproof feature, and commuting scenario. The key information and type words in the title have been retained, and the main selling point sentence adopts the active voice to reflect the protective function. The parameters column indicates the seasons used. After post-editing and information slot checking, the main body, attribute values, and application scope are fully consistent. During the translation stage, dependency analysis was conducted to determine the headword, modifying components, and parallel levels. Based on the target language's word order, the subject-predicate-object positions were arranged and combined. Structured conversion templates were provided for similar categories such as jackets, windbreakers, and raincoats.

4.3. Context adaptation correction in target language implementation

Based on consumer perception, cultural habits, regulatory laws, and brand image, target market models can be established in different countries and sales platforms. The tone level, salutation format, measurement system, performance commitment, and sensitive expressions can be transformed into operable matching dimensions. Each market version is labeled with country, platform, audience group, scenario, compliance level, etc. The corrected content also has a clear release scope. At the same time, the expression characteristics of different regions are retained ^[5]. Taking a cross-border beauty product as an example, the product description texts for the British version and the Southeast Asian version were generated by the project team. The British version focuses on commuting scenarios, skin sensation experience, and reapplication prompts, using relatively conservative functional expression methods. The Southeast Asian version emphasizes ventilation, waterproof, sweat-proof, and external functions in hot and rainy environments. For words like "cool and non-greasy" and "high sun protection," the team optimized the combination based on local e-commerce high-frequency search terms, unified the formatting of product capacity, applicable skin types, and usage frequency, and limited the efficacy description according to advertising norms. Reviewers checked the tone, culture, and regulatory

restrictions based on the brand image to form two different correction schemes. The final draft language information is bound to regional and platform tags, and subsequent products such as sun protection, isolation lotion, and body sun protection can be called upon as needed to achieve regional expression consistency of the same brand. Platform feedback and user comments can further adjust the tone parameters, promoting continuous updates of the market version.

5. Cross-border e-commerce product description corpus construction method

5.1. Collection and layered alignment annotation of multi-source corpora

The corpus collection scope includes platform product pages, enterprise independent websites, brand manuals, product catalogs, customer service records, and manual revisions. Metadata is created based on the source, category, language, market and region, and date. After collection, all files are converted to the same character set, measurement units, punctuation marks, and field names, and duplicates, incomplete, or outdated parts are removed. The structural relationship between the title, attributes, selling points, detailed information, and specifications is retained. After cleaning, the original text, machine translation, and manual correction texts are used as basic units to achieve alignment at the chapter, sentence segment, attribute field, and vocabulary term levels. At the chapter level, the same product and version are ensured; at the sentence level, the original information sequence is maintained; at the field level, the materials, size, function, and environment are matched; and at the vocabulary level, the translation of terms and their modifications are indicated. The annotation system sets error types, correction methods, quality levels, and review status, and records the annotator, time, and version number. For polysemous words, parameter combinations and omitted expressions, context explanations, and target market labels can be added. After double review and consistency verification, qualified corpora are divided into training set, validation set, and application set, and batch numbers and quality records are established for easy search of source, annotation level, and review results.

5.2. Correction of corpus classification storage and dynamic update

After the correction of the corpus is completed, a hierarchical system can be established based on product category, target language, sales market, text location, error type, and quality level. The basic data unit for storage is “original text - machine translation - corrected text - correction basis.” The database should simultaneously save product number, platform source, version time, reviewer, and scope of application, so that each piece of corpus can be traced back to its source, compared with the content, and re-reviewed and called upon. Fields for titles, attribute bars, selling points, and detail pages can be set separately. Frequent standard expressions are placed in the main corpus layer, contextualized expressions are placed in the expanded corpus layer, and texts awaiting review are placed in the temporary storage layer, forming a hierarchical call structure. The incremental update of the system should combine new product launches, platform rule adjustments, user feedback, and post-editing records. Regularly identify new terms, invalid translations, and contextual changes based on usage frequency, correction times, and review results; form an update list; and complete replacement, supplementation, or downgrade processing on the basis of manual review. After the update is completed, the system calculates the weight of the corpus based on version differences, adjusts the affiliation relationships in the core layer, expansion layer, and temporary layer, and records the effective time and scope of application, ensuring that the effective corpus for different markets,

platforms, and categories is consistent with the current business environment. Conclusion

Based on the correction of machine translation errors to improve and form the corpus of cross-border e-commerce product descriptions, the calibration of vocabulary terms, the correction of syntax and semantics, and the adaptation to the target language context can be incorporated into the same work chain, enhancing the accuracy, coherence, and market applicability of product information in various languages. This paper starts from error identification and correction rules and links together the collection of materials, hierarchical alignment, classification storage, and dynamic update, forming a network of the original text, machine translation, corrected text, and correction basis. This model is conducive to accumulating reusable product language materials, reducing the cost of repeated manual modification, and providing data support for platform content release, brand expression, and optimization of machine translation models. In the future, samples need to be continuously expanded according to different categories, languages, and sales regions; the quality evaluation and version control mechanisms need to be improved, and the ability of the corpus to cope with product iterations and market changes needs to be enhanced, promoting the construction of the cross-border e-commerce product description corpus towards standardization and refinement.

About the author

Ying Du, female, born in 1982 in E'Zhou in Hubei Province, Han, is an associate professor working in the School of Foreign Languages at the Shanghai Technical Institute of Electronics and Information. Her research interests include machine translation, corpus-based translation studies, and teaching evaluation.

Disclosure statement

The author declares no conflict of interest.

References

- [1] Deng JL, 2023, Construction of Chinese-English Parallel Corpus for Humanities and Social Sciences Literature. *Corpora Linguistics*, 10(1): 115–126.
- [2] Gu WH, Leng BB, 2024, Four Types of Terminological Errors in ChatGPT's Application in Science and Technology Translation — Taking Mechanical Engineering Terminology as an Example. *Chinese Science and Technology Translation*, 37(1): 24–27.
- [3] Geng F, Hu J, 2023, New Directions of AI-Assisted Post-Translation Editing — A Study on Translation Examples Based on ChatGPT. *Chinese Foreign Language*, 20(3): 41–47.
- [4] Wang JQ, Niu YY, 2023, Translation Quality Assessment in Computer-Assisted Translation Evaluation System. *Shanghai Translation*, (6): 52–57.
- [5] Xia L, Xiang XL, Li YM, et al., 2024, Current Situation Analysis of Machine Translation Abstracts of Medical Papers and the Research on the Collaborative Path between Authors and Editors. *Chinese Journal of Science and Technology Research*, 35(12): 1757–1766.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.