Application and Exploration of AI-Based Proofreading Systems for Postprints

Authors: Yang Gengxiu, Sun Jiafei, Feng Enda, Yin Lin

Date: 2023-10-08T00:00:00+00:00

Abstract

As media convergence enters a critical phase, news content originates from increasingly diverse sources and in greater volumes, dissemination channels are progressively diversified, and timeliness demands are ever more stringent, presenting new challenges for quality control and publication security in media organizations' content production. Currently, the application of artificial intelligence technology in proofreading can already evaluate the accuracy and appropriateness of word choice and sentence construction within manuscripts, and can even detect contradictions in emotional tone or logical inconsistencies. The All-Media Command Center of Xinhua Newspaper Media Group implements full-process intelligent proofreading and cross-proofreading across different systems throughout all critical stages of content production, leveraging proofreading results for deep learning to form a continuously iterative and improving closed-loop system.

Full Text

Application and Exploration of AI-Based Proofreading Systems

Abstract: As media convergence enters a deep-water zone, news content sources have become more diverse and voluminous, distribution channels have multiplied, and timeliness requirements have intensified. Media organizations now face new challenges in quality control and publication security. Currently, applying artificial intelligence technology to proofreading can already determine the accuracy and rationality of word and sentence usage in manuscripts, and even identify contradictions in emotional tone or logical inconsistencies. The All-Media Command Center of Xinhua Daily Media Group adopts full-process intelligent proofreading and cross-proofreading by different systems, integrating these capabilities throughout all key stages of content production, and utilizes proofreading results for deep learning to form a continuously iterative and improving closed loop.

Keywords: proofreading; verification; review; artificial intelligence; NLP; content security

Authors: Yang Gengxiu, Sun Jiafei, Feng Enda, Yin Lin

1. Project Background

News manuscript proofreading is an essential and critical step in the news production and publication process, serving as the key defense line to ensure publication safety and maintain journalistic rigor. Although manuscript review processes vary across major media publishing institutions, they all share one common step before official release—proofreading.

Alongside the evolving needs of the media industry for text proofreading, proofreading systems have undergone three generations of development. The first generation primarily relied on computer storage and basic operations, accumulating large error word databases to perform character-by-character and word-by-word matching against manuscript content, achieving word-level verification. The second generation employed intelligent technology for sentence-level text checking, capable of identifying unreasonable word collocations based on the overall context of sentences. The third-generation proofreading system is human-like, building upon the second generation' s capabilities to achieve semantic analysis through deep learning, conducting comprehensive analysis and understanding of manuscript content. Based on the overall viewpoint and tone of the full text, it determines whether each sentence and word is reasonable, and whether there are contradictions in emotional tone or logical inconsistencies.

As media convergence enters a deep-water zone, news dissemination channels have become increasingly diversified, and timeliness requirements continue to rise. The market demands greater speed, breadth, depth, and volume in content production, presenting new challenges for media organizations in quality control and publication safety. The All-Media Command Center project leverages the latest developments in semantic analysis and deep learning to explore the introduction of AI proofreading into the content production workflow and conducts statistical evaluations of proofreading effectiveness.

2.1 Challenges of Intelligent Proofreading

The main challenge of intelligent proofreading lies in emotion and semantic analysis. Since proofreading work primarily deals with text, NLP enables computers to read and understand content, provide error prompts, and automate proofreading tasks.

2.2 Application of Natural Language Processing

Natural Language Processing (NLP) is one of the most important technologies in the information age and a crucial component of artificial intelligence. Applications derived from NLP technology have been widely used in various fields,

including spell checking, machine translation, speech recognition, and chatbots. Deep learning provides a flexible, general, and learnable framework that has achieved breakthrough progress in speech recognition and computer vision.

2.3 Construction of Intelligent Proofreading System

For current mainstream proofreading systems, analyzing numerous typical erroneous entries reveals that the most common errors in Chinese proofreading systems include character-level errors, grammar-level errors, and semantic-level errors. Character-level errors mainly result from wrong characters, alternate characters, missing characters, extra characters, or transposed characters. By performing character-by-character and word-by-word matching against manuscript content, words matching entries in the error database are identified as character errors and prompted to users, such as “倡仪” (should be “倡议”), “国家” (should be “国家”), or “总埋” (should be “总理”). Grammar-level errors primarily refer to incorrect word collocations or missing characters. By extensively learning correct corpora, computer systems autonomously analyze and summarize language usage patterns and conventions, enabling machines to develop certain understanding and judgment capabilities at the sentence level, thereby identifying anomalies and unreasonable content within sentences to achieve proofreading purposes.

The intelligent proofreading system combines active proofreading and automatic proofreading, adopting a SAAS layout model that allows the intelligent proofreading system to be either embedded in the manuscript editing system or used as an independent auxiliary review module. Intelligent proofreading work is distributed across all key stages of content production, allowing editors to initiate AI proofreading for their current manuscripts at any time. This distributes error detection and correction throughout the manuscript circulation process, minimizing pressure on the final proofreading stage and reducing proofreading omissions caused by tight deadlines and large volumes.

Most article sentiment analysis primarily focuses on research into learning dictionary modeling and machine learning algorithms. By analyzing sentiment dictionaries, negation dictionaries, degree adverb dictionaries, and stop-word dictionaries, it calculates contextual sentiment orientation. It analyzes the collocation relationships between news themes and word modifiers to calculate word polarity, integrating dictionary resources to construct sentiment lexicons while employing weighted linear combination methods to determine article sentiment orientation.

Machine learning-based article sentiment analysis treats sentiment as a multi-classification problem, belonging to supervised learning methods. The machine learning approach requires processes including text preprocessing, feature selection, feature weighting, classifier training, and classification. This method’s classification performance is superior to the traditional TF-IDF (term frequency-inverse document frequency) feature weighting method.

The intelligent proofreading system applied in the All-Media Command Center not only achieves vocabulary and sentence checking but also performs certain sentiment analysis, conducting comprehensive analysis and understanding of manuscript content. Based on the full text' s viewpoint and tone, it determines whether each sentence and word is reasonable and whether there are viewpoint contradictions or logical inconsistencies. Through deep learning based on theme fusion, it employs Chinese text preprocessing methods to convert unstructured or semi-structured information into structured information that computers can understand, enabling comprehensive content analysis and understanding to automatically identify text sentiment categories and achieve intelligent verification.

Article themes and article sentiment typically share certain commonalities. Deep learning models can improve article sentiment classification model accuracy by integrating vectors. The proofreading system introduces a bidirectional LSTM sentiment algorithm to achieve contextual information fusion for words, overcoming both the gradient vanishing problem of traditional RNNs and addressing the limitation of traditional LSTM that can only effectively integrate preceding context information while lacking integration of following context information. By integrating text theme features, it constructs more accurate sentiment classification models.

In the manuscript collection and editing stage, the proofreading system participates in real-time. Editors and journalists can select proofreading, and the system will provide prompts for word misuse, semantic expression errors, etc., in the text manuscript and offer modification suggestions, helping editors write better from the first stage. Simultaneously, through proofreading intelligent assistants interacting with editors, when editors click each prompt on the right side, the focus in the editing box will locate accordingly, saving editors time searching for corresponding points in the original text. Meanwhile, when editors make decisions to correct or ignore prompted errors, the intelligent proofreading system records and learns from these decisions.

3. Building Full-Process Content Security

Traditional news manuscript proofreading is typically the final step before publication, with tight deadlines and heavy workloads; detected errors require return for revision and re-proofreading. In the environment of deep media integration, manuscript volumes have increased explosively, and real-time news on mobile terminals often pursues the fastest possible publication—delaying even one second may mean losing the optimal dissemination opportunity for a news item. Under such circumstances, the practice of placing all proofreading work in the final pre-publication stage can no longer meet current media requirements for multi-format news manuscripts, low time tolerance, and zero-error tolerance, let alone satisfy the long-term goal of building "Four All Media" in the future.

By integrating the application of intelligent proofreading systems and recon-

structing the content production workflow, the technological achievements of natural semantic analysis and deep learning are introduced into the entire content production process. After a period of operation, based on surveys of usage by editorial staff and statistical reports of manuscript errors at each stage, it shows advantages over traditional proofreading, detecting some critical errors that traditional proofreading cannot detect.

In the manuscript issuance stage, if editors fail to completely revise all issues in the news manuscript before submission, or if new errors are introduced during revision and editors submit directly to the manuscript database without noticing, when the manuscript is being issued, reviewers can utilize the intelligent proofreading system to re-proofread the manuscript once again. By implementing secondary proofreading at necessary workflow nodes, error correction work can be scheduled as early as possible to preceding nodes.

To avoid homogenization tendencies in proofreading results from the same intelligent proofreading system, the All-Media Command Center system introduces another proofreading system to conduct batch proofreading of manuscripts in the "final manuscript database," providing error risk alerts.

4. Dual-System Cross-Proofreading

Currently, single intelligent proofreading systems based on semantic analysis and deep learning still miss some detectable errors in practical applications, and learning results based on different corpora also produce differences in understanding and judgment of vocabulary, semantics, emotion, and other elements. In addition to layering and advancing proofreading work within the workflow, the intelligent proofreading system simultaneously introduces two different intelligent proofreading systems to conduct cross-proofreading of news manuscripts. The first system is responsible for proofreading individual manuscripts, while the second system conducts full-text proofreading again on manuscripts that have passed through the first system, forming an error warning table for manuscripts in the issuance database through statistical lists, and feeding these results back to the learning module of the intelligent proofreading system, enabling continuous self-improvement. This approach fully utilizes the strengths of each system, maximizing the effectiveness of intelligent proofreading in controlling manuscript quality.

Batch Cross-Proofreading Error Risk Alert List

In the future, intelligent proofreading systems will continue to improve in two aspects: learning based on private data and learning based on internet big data. Through localized learning, proofreading rules will be further refined to continuously enhance rigor; through internet big data learning, the system will follow the development of industry leaders in manuscript proofreading standards while timely understanding new internet expression patterns, fully leveraging the 叠加 effect of full-process proofreading and cross-proofreading to achieve a "1 + 1 > 2" effect.

(Author Affiliation: Xinhua Daily Media Group)

Note: Figure translations are in progress. See original paper for figures.