Post-print: An Exploration of a Media Corpus-Based Verification System

Authors: Gao Wen

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

Abstract

In news gathering and editing, aside from certain typographical errors, many errors are latent semantic-level errors. Semantic errors necessitate examining whether the semantics and pragmatics expressed in statements violate certain standards, and conventional text proofreading methods struggle to identify these semantic errors. For instance, incorrect formulations regarding the Taiwan issue of China that appear in newspapers and online articles are considerably difficult to correct using automatic proofreading tools. Nevertheless, the impact of such errors on news organizations must not be underestimated; some may constitute political errors that influence public opinion guidance, representing the foremost priority in editorial department proofreading. Consequently, leveraging media corpora for news gathering and editing proofreading constitutes a crucial development direction for automatic news text proofreading.

Full Text

Exploring a Proofreading System Based on Media Corpora

Xinhua News Agency Communications Technology Bureau, Beijing 100803

Abstract: In news gathering and editing, beyond typographical errors, many errors are latent semantic-level mistakes. Semantic errors require checking whether the meaning and pragmatics expressed in a statement violate certain standards, which are difficult to detect using conventional text proofreading methods. For instance, incorrect formulations regarding Taiwan-related issues that appear in newspapers and online articles pose considerable challenges for automated correction tools. Yet such errors cannot be underestimated in their impact on news organizations—some may constitute political errors affecting public opinion orientation, representing the top priority for editorial proofreading. Therefore, fully leveraging media corpora for news gathering and editing

proofreading constitutes an important direction for automated news text proofreading.

Keywords: news gathering and editing; proofreading; corpus; machine learning; knowledge graph

CLC Number: H131.2
Document Code: A
Article ID: 1671-0134(2021)11-142-03
DOI: 10.19483/j.cnki.11-4653/n.2021.11.044

Citation Format: Gao Wen. Exploring a Proofreading System Based on Media Corpora[J]. China Media Technology, 2021(11): 142-144.

In the Chinese proofreading domain, numerous proofreading software and services currently exist. Analysis of research findings and practical usage experience regarding corpus coverage, lexicon capacity, proofreading algorithms, and system compatibility reveals that due to inherent software limitations, capabilities in semantic-level proofreading remain constrained. The inability of some lexicons to update automatically online represents another major drawback, still requiring proofreading personnel to conduct further manual verification. However, manual proofreading entails high labor intensity and costs, and issues such as proofreader responsibility or visual fatigue still result in missed errors. Sometimes certain errors, particularly those with semantic implications, can be amplified and even exploited by malicious actors or media outlets, causing adverse effects. This further necessitates research emphasis on proofreading technologies incorporating semantic analysis and even knowledge base priori strategies to compensate for deficiencies in current proofreading software.

1. Constructing Effective Metadata Using Xinhua' s Existing Media Corpus

A corpus is a large database that applies computer technology to conduct statistical analysis of massive natural language materials. As a national news agency, Xinhua possesses a vast and authoritative news corpus, providing absolute advantages for constructing dedicated media corpus metadata for Xinhua news reports. A typical corpus system should include: document extraction and metadata creation; automatic part-of-speech/grammar annotation; and functional modules for indexing, retrieval, and statistical analysis. Among these, the most critical components are part-of-speech, syntactic, and error annotation, as the annotation methods and accuracy provided by the system directly determine the scale of corpus construction and the quality of research outcomes [1]. Therefore, while fully utilizing the rich news materials in Xinhua' s multimedia database to construct specialized metadata, experienced editors and proofreaders must cooperate in screening and providing positive and negative case samples for training proofreading models. The availability of these sam-

ple sets' priori data will directly affect the error detection rate of proofreading models.

Machine learning and deep learning algorithms can automatically mine statistical patterns from data to learn computational models [2]. Data quality and quantity are crucial for model effectiveness, with big data support being a key factor in deep learning' s recent successes. For semantic proofreading in news gathering and editing, establishing a large-scale dedicated corpus is essential.

1.1 Published News Content Data

Xinhua has accumulated massive and authoritative news corpora through long-term news gathering and editing work, providing absolute advantages for constructing dedicated media corpora for Xinhua news reports. These corpora have undergone rigorous software and manual proofreading, possessing high accuracy and standardization. Meanwhile, their linguistic characteristics and statistical properties of vocabulary and phrase collocations can provide rich positive samples for model training. The scale of this corpus will far exceed that used in previous academic research, providing powerful data support for training large-scale deep learning models.

1.2 Expert-Corrected News Text Data and Pre-Correction Data

Negative samples are essential for machine learning models. Learning algorithms need to mine various patterns of semantic errors from negative samples and generalize them. Semantic errors in negative samples should possess characteristics as similar as possible to actual errors. Xinhua' s proofreaders have examined and corrected various types of semantic errors in their long-term work, ensuring news report accuracy. These manually detected errors and the pre-correction data represent highly valuable negative sample data, providing excellent references and inspiration for automatic negative sample generation algorithms. For instance, some semantic errors often involve inputting incorrect homophones or near-homophones—words that are not erroneous in themselves but create semantically obvious differences when collocated with surrounding words. Based on this pattern, negative samples can be generated by randomly replacing words with homophones or near-homophones in positive samples. Through both collection and generation methods, massive and high-quality negative sample data can be obtained.

1.3 Large-Scale Web News Text Data

Beyond authoritative data provided by Xinhua, substantial news data from major print or online media searchable on the internet can also serve as training samples. Data crawling techniques are typically employed to capture this data, forming a more comprehensive and powerful news database. Current web crawler technology is already mature—for example, Toutiao and Baidu News

both crawl and aggregate media data. In crawler technology, main text extraction technology can be added as assistance, as web structures vary across different data sources, thus requiring more targeted crawler design to achieve massive crawling of news data from different websites.

1.4 Domain Knowledge Data

News gathering and editing may encounter factual errors, such as incorrect correspondences or collocations among people, events, locations, and times. Such errors involve no grammatical or semantic colocation issues; they cannot be successfully detected relying solely on linguistic characteristics and must utilize domain-specific knowledge. A knowledge graph is a tool for structurally organizing and representing information. It represents various things as entities, relationships between things as links between entities, and organizes knowledge from the human world in this manner, thereby providing support for various knowledge-related applications. Knowledge graphs have seen very successful applications in search engines, information retrieval, and automatic question answering, remaining a research hotspot for major technology companies and academic institutions for many years with relatively mature related technologies.

In the proofreading system construction, this component includes knowledge graph construction and application phases. The first phase involves building a large-scale knowledge graph for news reporting, realized through combining multiple algorithms. It begins with knowledge extraction from existing large-scale knowledge bases such as Baidu Baike and Wikipedia, which contain substantial structured knowledge that can be efficiently transformed into knowledge graphs through simple operations. Additionally, internet web pages contain massive knowledge, often in unstructured forms that are more chaotic than knowledge bases but can also be extracted through natural language processing and other automated technologies and added to knowledge graphs. Finally, knowledge from multiple sources is fused, entity relationships within knowledge graphs are inferred, and erroneous knowledge is eliminated to further improve knowledge graph quality. The second phase involves designing corpus knowledge query and verification algorithms using knowledge graphs to detect factual errors, including entity extraction, entity linking, and comparison with corresponding content in knowledge graphs to return results.

2. Machine Learning-Assisted Automatic News Text Error Correction Models

Semantic analysis can employ Conditional Random Fields (CRF), deep learning convolutional neural networks, knowledge graphs, and other machine learning models to construct automatic news text error correction models. Comprehensive application of these three advanced algorithms will enable more optimized news text proofreading algorithms. Semantic error detection relies on contextual information in text, requiring models to learn contextual dependencies within

a certain range—from phrases to single sentences or even multiple sentences.

2.1 Probabilistic Model Error Prediction

Probabilistic graphs combine graph structures with probability statistics, suitable for inference and prediction tasks on data with specific structures, and have broad applications in processing sequential data like text. Conditional Random Fields are a commonly used probabilistic graph model for processing contextual information. CRF extracts features from multiple consecutive words as input and calculates the joint probability distribution of output label sequences, where output labels can be defined according to application objectives [3]. Output labels can correspond to word appropriateness levels; when appropriateness is too low, words are classified as errors. CRF has successful applications in multiple natural language processing tasks, including part-of-speech tagging, syntactic analysis, and named entity recognition. Input features can utilize shallow features such as part-of-speech features and word vector features, which have been widely applied in industry with good results. CRF can learn statistical characteristics of word collocation combinations and predict semantic errors accordingly.

2.2 Deep Learning Modeling

Conditional Random Fields are a traditional shallow model with relatively low complexity, only able to learn relatively small contextual dependencies. In recent years, deep learning models have achieved excellent results in various artificial intelligence applications. Through multi-layer neural networks, they learn highly abstract features and statistical patterns from data. Due to high modeling capacity and computational parallelism, deep learning is highly suitable for applications combined with big data, capable of fully leveraging massive corpus advantages to learn longer-range contextual dependencies in word collocations.

Using Convolutional Neural Networks and Recurrent Neural Networks for semantic proofreading, statements first undergo word segmentation preprocessing, with each word represented by an initialized word vector. The long sequence of word vectors is then input into deep neural networks, ultimately outputting the probability of errors for each word [4]. Except for the word segmentation preprocessing step, the entire process is completely automatic learning requiring no manually designed features.

2.3 Advanced Semantic Error Correction Based on Knowledge Graphs

Deep learning can learn larger-range contextual dependencies but struggles to detect and correct factual errors. Such proofreading tasks require knowledge graphs. The main tasks involve designing algorithms for entity extraction, entity linking, and entity relationship extraction. Entity extraction identifies entities representing specific people, places, objects, etc., from statements. One method involves using part-of-speech tagging to identify nouns in statements,

which often correspond to entities. Proprietary entity lexicons can also be established for text matching queries. Entity linking connects extracted entities with entities stored in knowledge graphs, eliminating entity ambiguity and potential polysemy [5]. This task also requires utilizing contextual information, allowing reapplication of the aforementioned CRF and deep learning models. Entity relationship extraction identifies semantic relationships between two entities from statements. Many methods can be used for relationship extraction, including feature-based supervised learning methods, bootstrapping-based semi-supervised learning methods, and clustering-based unsupervised extraction methods. After completing these steps, knowledge graphs can be queried and compared to return knowledge verification results.

Implementation Considerations

Leverage Xinhua' s existing corpus to standardize news gathering and editing language. Xinhua issues over 2,000 stories daily, with nearly 800 Chinese-language pieces alone. Additionally, Xinhua has published the "Xinhua News Agency Prohibited Terms in News Reporting" to standardize terminology in several reporting domains. These rich and professional corpora undoubtedly represent valuable assets for Xinhua. Meanwhile, deep learning prerequisites require large amounts of effective data. Therefore, Xinhua' s news corpus is of great significance for training proofreading models. Using existing correct data from the corpus as prior knowledge, combined with generated or added massive negative samples, the proofreading model can be iteratively improved to achieve supervisory verification of news reporting language standards.

Optimize proofreading workflow and improve work efficiency. Currently, proofreading software used by major editorial departments can resolve some typographical errors. Adding semantic-level error recognition functionality undoubtedly represents further optimization and improvement of the proofreading workflow, enhancing news work efficiency while better ensuring news product quality. It should be noted that automatic proofreading tools can only serve as auxiliary means in news work and cannot completely avoid all textual and grammatical errors. Therefore, while optimizing workflow and improving efficiency, automatic proofreading tool software saves time and energy for journalists but does not eliminate the need for manual review. Instead, it allows journalists to devote more energy to optimizing news writing.

Train proprietary proofreading models to better serve different content domains. In different news content domains such as education, healthcare, energy, and politics, commonly used words and specialized vocabulary rarely intersect. Training domain-specific proofreading models using deep neural networks in specialized fields will yield more accurate semantic-level recognition. Currently, in Xinhua' s editing system, each manuscript contains a dedicated field for article classification. Using this field to select proofreading models and calling corresponding proofreading algorithms for analysis of current texts yields more accurate proofreading results.

Intelligent proofreading based on media corpora is not a simple technical task; it requires support from experienced news workers. For deep learning based on media corpora, corpus selection is undoubtedly one of the decisive conditions. In Xinhua' s multimedia database, massive news manuscripts serve as positive examples for deep learning, while negative example construction requires accumulation from experienced news workers collecting typical and common error types as important components of negative examples. The advantage of deep neural networks' multi-layer structure is representing complex functions with fewer parameters. Simultaneously, this requires training sample data to cover future samples as much as possible, enabling learned multi-layer weights to predict new samples well. Therefore, sample construction requires support from news workers with rich writing and proofreading experience to screen and construct positive and negative samples and determine error rate weights for different typographical and semantic errors. Only with sufficient effective samples can recognition rates for future test samples be improved. Close integration of journalists' rich experience in manuscript writing and proofreading with artificial intelligence technology enables tool software to provide better services for news work.

References

[1] Wei Naixing. Corpus-based and Corpus-driven Collocation Research[J]. Contemporary Linguistics, 2002(2): 101-114, 157.
[2] Wang Wentong, Wang Lichun. A Review of Deep Learning Research[J]. Journal of Beijing University of Technology, 2015(1): 48-59.
[3] Hong Mingcai, Zhang Kuo, Tang Jie, Li Juanzi. Chinese Part-of-Speech Tagging Method Based on Conditional Random Fields (CRFs)[J]. Computer Science, 2006(10): 148-151, 155.
[4] Zhou Feiyan, Jin Linpeng, Dong Jun. A Review of Convolutional Neural Network Research[J]. Chinese Journal of Computers, 2017(6): 1229-1251.
[5] Liu Qiao, Li Yang, Duan Hong, Liu Yao, Qin Zhiguang. A Survey of Knowledge Graph Construction Techniques[J]. Journal of Computer Research and Development, 2016(3): 582-600.

Author Bio: Gao Wen (1985-), female, from Dongying, Shandong, Engineer. Research direction: Computer application technology.

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