A Paradigm for Large-scale Refined General-purpose Corpus Construction: Postprint of a Book Review of "Large-scale Modern Chinese Word-segmented Corpus Construction and Application"

Authors: Qu Weiguang

Date: 2024-03-28T00:00:00+00:00

Abstract

[Purpose/Significance] This paper identifies the principal value and contributions of the book *Construction and Application of Large-Scale Modern Chinese Word Segmentation Corpus*, aiming to provide reference for Chinese corpus construction and to facilitate the rapid development of Chinese natural language processing in the era of large language models. [Method/Process] From both macro and micro perspectives, this study conducts a corpus-based quantitative linguistic analysis of the construction of the People' s Daily word segmentation corpus in the new era and related research reviews, and provides a critical evaluation of knowledge organization of the People' s Daily under deep learning. [Results/Conclusion] The book *Construction and Application of Large-Scale Modern Chinese Word Segmentation Corpus* conducts research based on the construction and application of the People' s Daily corpus in the new era, which not only inherits the system and philosophy of the Peking University People' s Daily corpus, but also provides corresponding solutions for domain-specific natural language processing tasks to a certain extent.

Full Text

Abstract

[Objective/Significance] This paper identifies the principal value and contributions of the book *Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus*, aiming to provide reference for Chinese corpus construction and to accelerate the advancement of Chinese natural language processing in the era of large language models. [Method/Process] From both macro and

micro perspectives, this study conducts a corpus-based quantitative linguistic analysis of the construction of the New Era People' s Daily Segmented Corpus and reviews related corpus research, while also evaluating the knowledge organization of People' s Daily content under deep learning frameworks. [Result/Conclusion] The book, centered on the construction and application of the New Era People' s Daily Segmented Corpus, not only inherits the system and philosophy of the Peking University People' s Daily corpus but also offers viable solutions for domain-specific natural language processing tasks.

Keywords: Corpus; People' s Daily; Deep Learning

Since deep learning ushered in a new wave of natural language processing, the importance of corpora and data has grown exponentially. Empowered by increasingly higher-quality and larger-scale corpora, deep neural network models continuously push the performance boundaries of computational language processing. With the rise of generative large language models, the quality of training corpora has become the decisive factor determining the applicability of artificial intelligence. The current technological landscape reveals a stark imbalance: over 90% of the high-quality corpora used in leading models such as GPT-4.0 are English, while Chinese corpora account for less than 1%. Against this backdrop, Chinese AI research, which takes deep learning as its benchmark, faces the urgent challenge of acquiring and constructing high-quality, ultra-large-scale corpora.

In response to this challenge, Professor Huang Shuiqing and Professor Wang Dongbo from Nanjing Agricultural University present a comprehensive solution in their book *Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus* [**?**]. Drawing upon more than a decade of dedicated research and accumulated expertise in corpus construction, the authors introduce the New Era People' s Daily Segmented Corpus (NEPD) in detail. They systematically examine the current state and existing problems of domestic corpus development, and explore applications of NEPD including performance evaluation, stylistic computation, deep learning model construction, and news-domain tasks such as keyword extraction, automatic summarization, text classification, and lexical-level retrieval. The book provides a comprehensive demonstration of NEPD' s advantages and characteristics, successfully constructing the largest refined word segmentation corpus resource currently available for Chinese natural language processing.

1. Constructing the Largest Refined General Word Segmentation Corpus

The NEPD corpus introduced in *Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus* offers numerous advantages. First, the raw corpus material is of exceptionally high quality. The corpus draws from *People' s Daily* publications following the entry of socialism with Chinese characteristics into a new era, specifically selecting all articles from January-June

2015, January 2016, January 2017, January 2018, and January 2022—totaling ten months of publications. This raw material features standardized, normative language with distinct contemporary characteristics. Second, the corpus achieves unprecedented scale, exceeding 30 million Chinese characters and far surpassing the 1-million-character Peking University January 1998 People' s Daily corpus, the 1-million-word Tsinghua Treebank, and the 2-million-word Penn Chinese Treebank. Third, the corpus has undergone precise manual segmentation. Unlike most corpora that combine machine annotation with human proofreading [**?**], NEPD was constructed entirely through meticulous manual processing, ensuring its accuracy and usability. In summary, NEPD represents the world' s largest refined modern Chinese general word segmentation corpus, holding significant application value across multiple domains. It facilitates lexical analysis and stylistic computation from a linguistic perspective, supports research on word segmentation ambiguities and dictionary compilation, and advances information organization and service studies, thereby contributing to linguistics, information science, artificial intelligence, natural language processing, and data science.

2. Reviewing the Current State of Corpus Research and Construction

According to the definition established in *Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus*, "corpus, or language material, refers to collections of phonetic, sentential, lexical, and grammatical samples gathered from real linguistic environments for specific purposes" [**?**]. Through tracing early corpus research and exploring theoretical concepts, the book further defines a corpus as "a dataset formed through manual or machine processing and annotation of authentic language materials"[**?**], with storage formats encompassing both database and non-database approaches such as text files [**?**]. Building upon this foundation, the authors conduct comprehensive literature surveys and summarize corpus construction and development characteristics through quantitative data analysis, qualitative content examination, application status interpretation, and inventorying of representative corpora.

The book presents a quantitative analysis of domestic corpus research, revealing two major trends through publication analysis, collaboration analysis, and thematic evolution studies: increasing diversification of research objects and progressively refined and deepened technical content. In terms of research content, the book categorizes corpus studies into corpus construction and corpus application. The former addresses standard construction procedures, annotation granularity, and annotation strategies, while the latter summarizes applications in language teaching, domain glossary and dictionary compilation, information retrieval and extraction, language comparison and translation studies, and natural language processing. Finally, the book inventories representative domestic corpora across four categories: general monolingual corpora, Chinese-English parallel corpora, other Chinese-foreign parallel corpora, and specialized corpora,

introducing 17 of the most representative corpora including the National Language Commission Modern Chinese Balanced Corpus, the Chinese Academy of Sciences Chinese-English Parallel Corpus, the Peking University Institute of Computational Linguistics Bilingual Corpus, and the Chinese Interlanguage Corpus. This section provides crucial reference value for comprehensively understanding the development of domestic corpus research over the past three decades.

3. Introducing Construction Methods and Evaluating NEPD Performance

Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus systematically reviews rule-based, statistics-based, and sequence labeling-based Chinese automatic word segmentation methods, providing theoretical and practical foundations for subsequent model construction and performance evaluation. The book details the acquisition methods and procedures for the full-text raw *People' s Daily* materials, including technical specifics on data preprocessing, processing strategies, and encoding schemes. It presents a complete picture of NEPD' s annotation specifications, processes, and outcomes, covering annotator training strategies, multi-step annotation workflows, and special case handling guidelines.

The book also demonstrates NEPD' s segmentation performance evaluation. Using sequence labeling models, it compares the segmentation performance of the January 1998 and January 2018 *People' s Daily* corpora, revealing significant differences between them. Models trained on the 1998 corpus can no longer meet the segmentation requirements of contemporary Chinese texts, whereas the 2018 corpus fully satisfies these needs, thereby validating the necessity of constructing NEPD [**?**]. NEPD effectively addresses the lack of continuous updates to the Peking University People' s Daily corpus, representing both a continuation and expansion of the existing People' s Daily corpus resources in the current era and a worthy inheritance of the academic legacy of Professor Yu Shiwen, the creator of the original Peking University People' s Daily corpus. Additionally, NEPD provides robust corpus resource support for tasks such as named entity recognition, semantic retrieval, and shallow syntactic parsing [**?**].

4. Examining Contemporary Chinese Text Style and Segmentation Ambiguity

The book reviews and consolidates research on modern Chinese sentence length analysis, subsequently 统计了各类型汉语句子在 2015 年 1 月至 6 月、2016 年 1 月、2017 年 1 月、2018 年 1 月、2022 年 1 月共 10 个月的人民日报语料在各月度的分布情况 [**?**]. It also analyzes the distribution of various sentence types from both character and word dimensions, facilitating comprehensive understanding of the NEPD corpus. Furthermore, thanks to the corpus' s scale, Zipf' s law in word distribution is fully validated, with analytical results holding important reference

value for quantitative linguistics research. The book systematically analyzes word segmentation ambiguities, 统计了不同词长的词频 based on the ten months of NEPD data, identifying key vocabulary that reflects the corpus' s contemporary characteristics. In analyzing variant and exceptional word frequencies, the book comprehensively examines the coalescence degree and syntactic features of variant words across 17 part-of-speech categories, providing fresh data support for modern Chinese lexical and syntactic research.

5. Facilitating Multi-type Information Organization and Services for News Corpora

The latter half of *Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus* focuses on practical explorations of NEPD' s applications in deep learning contexts [**?**]. First, the book constructs deep learning segmentation models based on NEPD [**?**]. Employing Bi-LSTM and Bi-LSTM-CRF frameworks through sequence labeling, these models achieve F1 scores of 97.16% and 97.67% respectively, providing reliable benchmark metrics for modern Chinese automatic segmentation research. Second, addressing NEPD's news characteristics, the authors conduct automatic keyword extraction research [**?**], integrating six algorithms—TF-IDF, Yake, TextRank, Rake, LDA, and LSI—combined with manual review to extract the top 500 keywords from each monthly corpus. The results reflect thematic characteristics of *People' s Daily* articles, capturing major social events and developmental priorities across different periods, particularly new societal focal points in the new era, thereby providing valuable references for comprehensively grasping social evolution trends.

Third, the book explores automatic summarization for news corpora based on NEPD [**?**], implementing both extractive and abstractive approaches. Extractive summarization employs sentence weighting and traditional TextRank algorithms, while abstractive summarization utilizes the generative pre-trained T5 PEGASUS model. Results show that extractive summarization performs well on Rouge metrics, whereas abstractive summarization excels in grammatical quality, fluency, and information content. Fourth, the book investigates news text classification using NEPD, selecting 9,275 news articles across six sections (international, economic, social, sports, cultural, and political) to compare CNN, RNN, and BERT frameworks. BERT achieves the optimal performance with an F1 score exceeding 82%, reaching 98.72% accuracy specifically for sports news. Fifth, the book examines the impact of segmentation features on classification performance, demonstrating that models incorporating segmentation features achieve the highest accuracy and recall rates. Sixth, the book discusses the development of a news vocabulary-level retrieval system based on the corpus. Using the BM25 algorithm, it designs a complete retrieval system encompassing data storage and retrieval implementation, providing a detailed construction scheme with thorough explanations of data processing and algorithmic computation workflows, ultimately building a comprehensive news vocabulary-level retrieval platform.

Conclusion

Over twenty years have passed since the January 1998 *People' s Daily* annotated corpus was constructed under the leadership of Professor Yu Shiwen at Peking University [**?**, **?**]. The NEPD corpus upholds Professor Yu' s vision of building the most fundamental corpus for Chinese information processing, expanding the scale of *People' s Daily* materials, enhancing their timeliness, and demonstrating their diachronic evolution. In response to the new trends of big data and artificial intelligence development, and addressing the demand for high-quality large-scale corpora in cutting-edge information technology, *Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus* provides invaluable foundational resources and a technical application framework for modern Chinese natural language processing research through the introduction of NEPD and exploration of corpus-based quantitative analysis and information organization. The book, centered on corpus resource construction and application and integrating corpus linguistics, data science, and natural language processing techniques, offers a research pathway worthy of emulation for domain studies and disciplinary development. It is also hoped that the team of Huang Shuiqing and Wang Dongbo will continue to advance with the times, providing ever-improving corpora to drive continuous progress in Chinese information processing.

Note: Individuals or institutions can apply for access to the NEPD corpus for non-commercial research purposes at https://corpus.njau.edu.cn/.

References

[1] Huang Shuiqing, Wang Dongbo. *Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus* [M]. Nanjing: Nanjing University Press, 2023.

[2] Qu Weiguang, Tang Xuri, Yu Jingsong. Research on Ultra-large-scale Corpus Refinement Technology [J]. *Contemporary Linguistics*, 2009(2): 136-146.

[3] Huang Shuiqing, Wang Dongbo. Construction, Performance, and Application of the New Era People' s Daily Segmented Corpus (III)—Comparative Analysis of Sentence and Word Lengths [J]. *Library and Information Service*, 2019, 63(24): 5-15.

[4] Huang Shuiqing, Wang Dongbo. Review of Domestic Corpus Research [J]. *Journal of Information Resources Management*, 2021, 11(3): 4-17, 87.

[5] Huang Shuiqing, Wang Dongbo. Construction, Performance, and Application of the New Era People' s Daily Segmented Corpus (I)—Corpus Construction and Evaluation [J]. *Library and Information Service*, 2019, 63(22): 5-12.

[6] Huang Shuiqing, Wang Dongbo. Construction, Performance, and Application of the New Era People' s Daily Segmented Corpus (II)—Deep Learning

Automatic Word Segmentation Model Construction [J]. *Library and Information Service*, 2019, 63(23): 5-12.

[7] Zhou Hao, Wang Dongbo, Huang Shuiqing. Keyword Extraction and Analysis Based on the New Era People' s Daily Segmented Corpus [J]. *Journal of Literature and Data Science*, 2022, 4(1): 21-34.

[8] Liang Yuan, Wang Dongbo, Huang Shuiqing. News Automatic Summarization for People' s Daily Corpus [J]. *Knowledge Management Forum*, 2022(4): 452-464.

[9] Yu Shiwen, Duan Huiming, Zhu Xuefeng. Basic Processing Specifications for Peking University Modern Chinese Corpus [J]. *Journal of Chinese Information Processing*, 2002(5).

[10] Yu Shiwen, Zhu Xuefeng, Duan Huiming. Processing Specifications for Large-scale Modern Chinese Annotated Corpus [J]. *Journal of Chinese Information Processing*, 2000(6).

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