文 献 与 数 据 学报
第 6 卷　第 1 期　　2024 年 3 月

大规模精加工通用语料库建设的范例

——《大规模现代汉语分词语料库构建及应用》书评

曲维光[1,2]
（1. 南京师范大学中北学院，镇江　212399；
2. 南京师范大学计算机与电子信息学院 / 人工智能学院，南京　210023）

摘　要：[**目的 / 意义**] 指出《大规模现代汉语分词语料库构建及应用》一书的主要价值与贡献，旨在为中文语料库的构建提供借鉴，以促进大语言模型下中文自然语言处理的快速发展。[**方法 / 过程**] 从宏观和微观的视角，对新时代人民日报分词语料库的构建和语料库相关研究综述进行了基于语料库计量语言学的分析，并对深度学习下的人民日报知识组织等内容进行了评介。[**结果 / 结论**]《大规模现代汉语分词语料库构建及应用》一书基于新时代人民日报分词语料库构建及应用进行研究，不仅传承了北京大学人民日报语料库的体系和理念，而且在一定程度上为解决领域化自然语言处理的任务提供了相应的方案。

关键词：语料库　人民日报　深度学习
分类号：G202
DOI：10.31193/SSAP.J.ISSN.2096-6695.2024.01.10

自深度学习引领自然语言处理的新浪潮以来，语料和数据的重要性与日俱增。在更高质量更大规模语料的助力下，深度神经网络模型不断刷新计算机处理自然语言的性能上限。随着生成式大语言模型的兴起，语料的优劣更是直接决定了人工智能的应用程度。当前的技术现状是，以 GPT4.0 为代表的大语言模型使用的优质语料 90% 以上都是英文语料，而中文语料的占比甚至还不到 1%。当前国内人工智能的研究以深度学习为标杆，获取和构建高质量的超大规模语料库是亟待解决的技术难题。

在这一背景下，南京农业大学黄水清教授和王东波教授合著的《大规模现代汉语分词语料库构建及应用》[1] 给出了应对上述挑战的解决方案。该书源于他们在语料库构建领域十余年的深耕和积累。该书在详细介绍了他们推出的新版人民日报分词语料库（New Era People's Daily Segmented Corpus，简称 NEPD）①的基础上，细致考察了当前国内语料库建设的发展现状和存在问题，并围绕 NEPD 语料库进行了性能评测、语言风格计算、深度学习模型构建以及新闻领域的

[作者简介] 曲维光，男，教授，博士生导师，研究方向为自然语言处理、计算语言学、语言工程，Email：wgqu@njnu.edu.cn。

关键词抽取、自动摘要、自动分类和词汇级检索等应用。该书全面展示了 NEPD 语料库的优势和特点，完成了当前汉语自然语言处理领域最大规模的分词语料资源的构建。

1　构建最大规模的精加工通用分词语料库

《大规模现代汉语分词语料库构建及应用》推出的 NEPD 语料库具有诸多优势。首先，原始语料质量高。该语料库以中国特色社会主义进入新时代以后的《人民日报》为原始语料素材，选取了 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》所刊发的全部文章，原始语料的语言规范、标准，时代特色鲜明。其次，语料库规模大。该语料库整体规模超过 3000 万汉字，远超 100 万字的北京大学 1998 年 1 月人民日报语料库、100 万字的清华树库和 200 万字的宾州汉语树库。而且，语料库完成了精准的人工分词。不同于绝大多数结合机器标注和人工校对的语料库[2]，该语料库全部采用人工分词精加工方式构建，其准确性和可用性具有充分的保证。总的来说，NEPD 语料库是目前世界上规模最大的精加工现代汉语通用分词语料库。该语料库在诸多领域具有重要的应用价值，有助于语言学视角的词汇分析[3]、风格计算研究，有助于词语切分歧义研究及词典编纂工作，有助于信息组织与服务研究，对于语言学、情报学、人工智能、自然语言处理、数据科学等研究都具有促进作用。

2　梳理语料库研究及语料库建设现状

按照《大规模现代汉语分词语料库构建及应用》中的定义，“语料，即语言材料，指的是为一定目的收集的真实语言环境中出现过的语音、句子、词汇、语法等素材”[1]。通过对语料库早期研究的追溯和理论概念的探讨，《大规模现代汉语分词语料库构建及应用》还将语料库定义为“对真实语料进行人工或机器加工、标注后形成的数据集”[1]，且存储类型包含了数据库方式和文本文件等非数据库方式[4]。在此基础上，该书通过全面的文献调研，从定量数据分析、定性内容考察、应用现状阐释以及代表性语料库梳理等方面对语料库的建设和发展特点进行了全面的总结。

首先，《大规模现代汉语分词语料库构建及应用》一书完成了国内语料库研究的定量分析，通过发文分析、合作分析、主题演变等角度，总结出研究对象逐渐多元化、技术内容愈发细致深化两个重要趋势。其次是语料库的研究内容考察，该书将语料库研究归纳为语料库构建和语料库应用两大类型，前者梳理出语料库的规范构建流程、语料库构建过程中的数据标注粒度和标注策略等问题，后者总结了语料库在语言教学、领域词表和词典编撰、信息检索和信息抽取、语言对比和翻译研究、自然语言处理等诸多领域的应用实例。最后是对国内代表性的语料库的整理，作者从通用单语语料库、汉英双语平行语料库、其他汉外平行语料库、其他特色语料库等四个方面，介绍了国家语委现代汉语通用平衡语料库、中国科学院汉英平行语料库、北京大学计算语言研究所双语平行语料库、汉语中介语语料库等 17 种国内最具代表性的各类语料库。该部分的研究内容从定量和定性两个角度进行了深入分析，对于全面了解国内语料库研究近 30 年来的发展现状具有重要的参考价值。

3 介绍大规模通用分词语料库构建方法并测评 NEPD 的性能

《大规模现代汉语分词语料库构建及应用》一书梳理了基于规则、基于统计和基于序列标注的汉语自动分词方法，为后续的分词模型构建和分词性能测评提供了理论和实践依据。书中详细介绍了《人民日报》全文本原始语料的获取方法及其流程，包括数据预处理、加工策略、编码方式等技术细节，完整呈现了 NEPD 的标注规范、过程及结果，包括标注人员培训策略、多步骤标注流程、特例规范说明等内容。

本书还展示了 NEPD 语料库分词性能测评。书中以序列化标注模型为依据，对比了 1998 年 1 月人民日报语料和 2018 年 1 月人民日报语料的分词性能，发现了两者之间的明显差异。在 1998 年语料基础上训练出的分词模型已经无法适应于当今时代的现代汉语文本切分需求，而 2018 年语料则能够充分满足这一需求，这充分证明了构建 NEPD 语料库的必要性[5]。NEPD 显著地弥补了北京大学人民日报语料没有持续更新的不足，可以认为是对已有人民日报语料库在当前时代的延续和扩充，是对北京大学人民日报语料库的创建者俞士汶先生学术事业最好的传承。同时，NEPD 也可为命名实体识别、语义检索和浅层句法分析等任务提供有力的语料资源支撑[1]。

4 考察当代汉语文本的语言风格和分词歧义

《大规模现代汉语分词语料库构建及应用》对现代汉语句长分析的研究进行了回顾和整理，并在此基础上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[3]。该书还从字词两个维度分别统计了各类型句子长度的分布情况，有助于全面了解 NEPD 语料库。此外，得益于该语料的规模，词分布上的齐普夫定律得到了充分验证，相关分析结果对于计量语言学研究具有重要参考价值。《大规模现代汉语分词语料库构建及应用》还对分词歧义进行了系统分析，根据 NEPD10 个月的语料分别统计了不同词长的词频，发现了能充分体现 NEPD 时代特征的部分重要词语。在变异词及异例词的词频统计分析中，该书全面考察了 17 种词性下变异词的从合度和句法特征，为现代汉语词性和句法研究提供了鲜活的数据支撑。

5 助力新闻语料的多类型信息组织和服务

《大规模现代汉语分词语料库构建及应用》一书在后半部分重点对 NEPD 在深度学习下的应用场景进行了实践探索[6]。第一，基于 NEPD 语料库构建了深度学习分词模型[5]。该模型主要基于 Bi-LSTM 和 Bi-LSTM-CRF 两类深度学习框架，通过序列化标注形式，F1 值分别达到了 97.16% 和 97.67%，为当前现代汉语自动分词研究提供了可靠的参考指标。第二，面向 NEPD 的新闻特点，开展了新闻关键词的自动抽取研究[7]，综合采用了 TF-IDF、Yake、TextRank、Rake、LDA 和 LSI 等 6 种不同的算法，结合人工审核的方法获取了每个月语料中的前 500 个关键词。从结果来看，关键词反映出《人民日报》文章的内容主题特点，反映出社会发展各阶段的大事件和

侧重点，尤其是新时代的社会新焦点，为准确掌握社会发展变化的全貌与趋势提供了有力参考。第三，基于 NEPD 开展了新闻语料的自动摘要研究[8]。该研究分别实现了抽取式自动摘要和生成式自动摘要，抽取式自动摘要采用句子权重和 TextRank 传统算法，生成式自动摘要则参考了生成式预训练模型 T5 PEGASUS 模型。从结果来看，抽取式摘要在 Rouge 指标上表现良好，生成式摘要则在语法、流畅性、信息量方面表现优异。第四，基于 NEPD 语料研究了新闻文本自动分类问题。研究选取国际、经济、社会、体育、文化、政治 6 个版面共计 9275 篇新闻报道，对比了 CNN、RNN 和 BERT 三类常见的深度学习模型框架，其中 BERT 模型的性能最优，F1 值达到了 82% 以上，尤其在体育类上表现最好，准确率高达 98.72%。第五，该书进一步考察了分词特征对分类效果的影响。结果表明，添加了分词特征的模型无论在准确率还是召回率中均能获得最高值均出的表现。第六，该书还探讨了语料库基础上的新闻词汇级检索系统的研发。基于 BM25 算法，该书设计了包含数据存储和检索实现两部分的检索系统。该书还提供了完整的检索系统构建方案，细致介绍了数据处理和算法计算流程，并构建了完整的新闻词汇级检索平台。

6 结 语

弹指韶光过，1998 年 1 月人民日报标注语料库在北京大学俞士汶先生的主持下完成构建不觉已二十多年[9-10]，NEPD 语料库秉承俞先生为中文信息处理构建最基础语料的宗旨，拓展了人民日报语料的规模，提升了人民日报语料的时效性，展示了人民日报语料的历时性。在大数据、人工智能发展的新趋势下，面向前沿信息技术对优质大规模语料的需求，《大规模现代汉语分词语料库构建及应用》一书通过推出和介绍 NEPD 语料库，并结合基于该语料库的计量分析和信息组织探索，为现代汉语自然语言处理研究提供了宝贵的基础资源和技术应用框架。该书以语料库资源构建和应用为线索，结合语料库语言学、数据科学、自然语言处理技术等，为领域研究和学科发展提供了值得借鉴的研究路径。同时也希望黄水清、王东波团队，能够与时俱进，不断提供更好的语料，推动汉语信息处理的不断进步。

【注释】

①个人或机构可以通过网址https://corpus.njau.edu.cn/ 申请NEPD语料库从事非商业行为的相应研究。

【参考文献】

［1］黄水清，王东波. 大规模现代汉语分词语料库构建及应用［M］. 南京：南京大学出版社，2023.

［2］曲维光，唐旭日，俞敬松. 超大规模语料库精加工技术研究［J］. 当代语言学，2009（2）：136–146.

［3］黄水清，王东波. 新时代人民日报分词语料库构建，性能及应用（三）——句长与词的分析比较［J］. 图书情报工作，2019，63（24）：5–15.

［4］黄水清，王东波. 国内语料库研究综述［J］. 信息资源管理学报，2021，11（3）：4–17，87.

［5］黄水清，王东波. 新时代人民日报分词语料库构建，性能及应用（一）——语料库构建及测评［J］. 图书情报工作，2019，63（22）：5–12.

［6］黄水清，王东波. 新时代人民日报分词语料库构建，性能及应用（二）——深度学习自动分词模型构建

［J］. 图书情报工作，2019，63（23），5–12.

［7］周好，王东波，黄水清. 新时代人民日报分词语料库下关键词抽取及分析研究［J］. 文献与数据学报，2022，4（1）：21–34.

［8］梁媛，王东波，黄水清. 面向人民日报语料的新闻自动摘要生成［J］. 知识管理论坛，2022（4）：452–464.

［9］俞士汶，段慧明，朱学锋. 北京大学现代汉语语料库基本加工规范［J］. 中文信息学报，2002（5）：49–64.

［10］俞士汶，朱学锋，段慧明. 大规模现代汉语标注语料库的加工规范［J］. 中文信息学报，2000（6）：58–64.

An Example of Large-scale Refinement Teneral Corpus Construction: Comments on *Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus*

Qu Weiguang[1,2]

(1. North and Middle College, Nanjing Normal University, Zhenjiang 212399, China;
2. School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, Nanjing 210023, China)

Abstract: ［**Purpose/Significance**］This paper points out the main value and contribution of the book *Construction and Application of Large-scale Modern Chinese Word Segmentation Corpus*, so as to provide reference for the construction of Chinese corpus and promote the rapid development of Chinese natural language processing under the large language model. ［**Method/Process**］From a whole and micro perspective, this paper reviews the construction of People's Daily word segmentation corpus in the new era, corpus-based quantitative linguistics analysis, and the knowledge organization of People's Daily under deep learning. ［**Result/Conclusion**］Based on the construction and application of People's Daily word segmentation corpus in the new era, the book not only inherits the system, ideas and beliefs of People's Daily corpus, but also provides corresponding solutions for solving the tasks of domanized natural language processing to a certain extent.

Keywords: Corpus; People's Daily; Deep learning

（本文责编：孔青青）