Construction, Performance, and Application of the New Era People's Daily Word Segmentation Corpus (I): Corpus Construction and Evaluation Postprint

Authors: Huang Shuiqing, Wang Dongbo

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] The construction of a People's Daily word segmentation corpus adapted to the new era provides up-to-date, finely-annotated data for Chinese information processing and a novel linguistic resource for the diachronic analysis of modern Chinese. [Method/Process] Through analyzing existing Chinese word segmentation corpora, this paper describes the data sources, annotation standards, and procedural workflow of the constructed New Era People's Daily corpus, evaluates its performance via an automatic word segmentation annotation model, and conducts comparative analyses with existing corpora. [Results/Conclusion] The New Era People's Daily corpus adheres to fundamental processing specifications for modern Chinese corpora, characterized by large scale and extensive temporal coverage. Utilizing the January 2018 subset, a Conditional Random Fields (CRF)-based segmentation model is constructed to perform performance evaluation and comparison with the January 1998 People's Daily corpus. The comprehensive evaluation metrics obtained indicate that the New Era People's Daily corpus demonstrates superior overall performance, that the 1998 corpus is not substitutable, and that the current construction of this corpus is imperative.

Full Text

Preamble

Construction, Performance, and Application of the New Era People's Daily Segmented Corpus (I)
—*Corpus Construction and Evaluation*

Huang Shuiqing[1],[2], **Wang Dongbo**[1],[2]
[1]College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095
[2]Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing 210095

Abstract

[**Purpose/Significance**] This study constructs a segmented corpus of *People's Daily* adapted to the new era, providing the latest precisely annotated data for Chinese information processing and offering new linguistic resources for analyzing modern Chinese from a diachronic perspective. [**Method/Process**] Based on an analysis of existing Chinese word segmentation corpora, we describe the data sources, annotation specifications, and workflow of the constructed New Era People's Daily Corpus. We evaluate its performance by building automatic word segmentation models and compare it with existing corpora. [**Result/Conclusion**] The New Era People's Daily Corpus (NEPD) follows basic modern Chinese corpus processing standards, featuring large scale and long time span. Selecting the January 2018 portion, we construct a segmentation model based on Conditional Random Fields (CRF) and conduct performance evaluation and comparison with the January 1998 People's Daily corpus. The specific evaluation metrics demonstrate that the New Era People's Daily corpus exhibits outstanding overall performance, the 1998 corpus cannot serve as a substitute, and constructing this new corpus is highly necessary.

Keywords: new era; People's Daily; automatic word segmentation; conditional random field model; corpus; NEPD

1. Introduction

A corpus is a dataset composed of authentic linguistic materials annotated manually or automatically. Corpora serve as effective tools and means for natural language-related research, enabling the study of both universal linguistic patterns and specific texts. Chinese natural language processing requires an additional automatic word segmentation step compared to other languages. Chinese automatic word segmentation is the foundation of all Chinese information processing tasks, and its quality directly determines the performance of subsequent tasks such as part-of-speech tagging, entity extraction, automatic syntactic parsing, and machine translation. Currently, the mainstream technology for Chinese word segmentation is machine learning, which automatically learns the distributional characteristics and knowledge of vocabulary from precisely processed corpora to identify words in Chinese strings. Word segmentation corpora represent one of the most important types of Chinese corpora. While different machine learning models built on the same corpus can produce different segmentation models, the performance gap is generally not substantial. Instead, the annotation accuracy of the training corpus has a greater impact on segmentation

results. In Chinese information processing research, training corpora typically consist of general-domain and domain-specific corpora. Among Chinese general word segmentation corpora, the 1998 People's Daily corpus constructed by Peking University's Institute of Computational Linguistics is the most representative and influential. However, with the passage of time, the 1998 corpus can no longer reflect the richness and coverage of contemporary vocabulary. The New Era People's Daily Corpus (NEPD) includes articles published after 2012, when socialism with Chinese characteristics entered a new era. To distinguish it from Peking University's 1998 People's Daily corpus, we name this corpus the New Era People's Daily Corpus (NEPD). Currently, NEPD covers nine months of *People's Daily* articles (January–June 2015, and January 2016, 2017, and 2018). To promote open access and sharing of corpus resources, NEPD will be made available to the academic community for research purposes, with continued supplementation of the latest materials. NEPD possesses both dynamic diachronic span and static semantic richness. This series of articles will discuss NEPD's basic characteristics, construction process, segmentation performance, optimal segmentation models, and explore contemporary Chinese syntactic and stylistic features from a diachronic perspective. This first article introduces the construction process, specifications, and principles of the New Era People's Daily Corpus, evaluates its performance through a CRF-based segmentation model, and compares it with the January 1998 People's Daily corpus. The results demonstrate that NEPD significantly outperforms the 1998 corpus when processing recent *People's Daily* articles, proving the necessity of constructing NEPD.

2. Analysis of Chinese Word Segmentation Corpora and Models

2.1 Existing Chinese Word Segmentation Corpora

Among general Chinese word segmentation corpora, the most representative and influential is the People's Daily corpus from Peking University. The publicly released portion primarily consists of January 1998 *People's Daily* articles, completed by Professor Yu Shiwen and colleagues at Peking University's Institute of Computational Linguistics. The development process included establishing annotation specifications and research on retrieval methods [1-2]. The second is the National Language Commission's Modern Chinese Balanced Corpus, characterized by its balance and large scale, covering not only news but also materials from economics, military, sports, and other domains [3]. The third is the segmentation corpus in the Tsinghua Chinese Treebank, which implements Chinese word segmentation based on Li Jinxi's linguistic theory that "parts of speech are determined by syntactic function" [4]. The fourth is the segmentation corpus in the Penn Chinese Treebank, which follows structuralist linguistic theory [5]. Among these four corpora, the first two are relatively large-scale and employ consistent segmentation principles and specifications, but their timeliness has become increasingly problematic over time. The latter two adopt unique

linguistic theories but are relatively small in scale and also suffer from poor timeliness.

2.2 Chinese Word Segmentation Models

Common machine learning models for Chinese word segmentation include Hidden Markov Models, Maximum Entropy Models, and Conditional Random Fields (CRFs). Among these three, CRF has become the mainstream technology for Chinese word segmentation because it addresses independence assumption and label bias problems while allowing arbitrary feature knowledge to be added during training. Representative research includes C. Huang's [6] review of Chinese word segmentation progress from 1997–2007, which highlighted that statistical learning methods achieved significant breakthroughs in handling unknown words compared to rule-based methods, and emphasized the importance of publicly available evaluation datasets. Li Shuanglong et al. [7] defined feature functions as arbitrary real-valued functions rather than binary functions, reducing feature quantity and selection complexity, achieving an F-score of 95.2% in closed testing on the 1st SIGHAN test set. Shen Qinzhong et al. [8] incorporated character position probability features from the perspective of character word-formation capability, demonstrating that this feature improved F1-score by 3.5% to 94.5%. Chi Chengying et al. [9] achieved accuracies of 95.8% and 95.9% in closed tests on the SIGHAN 2006 Bakeoff Uppen and Msra corpora, respectively, while noting that CRF models performed poorly on multi-character unknown words. Song Yan et al. [10] proposed a character-word joint decoding method integrating CRF and Bigram language models, achieving an F-score of 93.9% on Bakeoff3. Liu Zewen et al. [11] proposed a 5-Tag labeling method using LCCRF for Chinese short texts, with dictionary-based correction achieving average F-scores exceeding 95% on four SIGHAN Bakeoff 2005 test sets, showing that inappropriate features not only decreased F-scores but also increased time and space complexity. Feng Xue [12] designed a statistical model incorporating dictionary information into character sequence labeling and word beam search, achieving good performance in both in-domain and cross-domain settings. Wang Ruojia et al. [13] constructed a 100,000-level medical dictionary combining authoritative domestic dictionaries, official standards, and medical supplements, achieving 82% F-score for electronic medical record segmentation.

Given CRF's strong performance in linear sequence tasks like word segmentation, this study selects the CRF model to build a segmentation model based on NEPD. Using CRF also facilitates performance comparison between models built on January 2018 and January 1998 People's Daily corpora.

3. Corpus Acquisition and Preprocessing

3.1 Raw Corpus Acquisition

NEPD's raw corpus was downloaded from the *People's Daily* full-text image retrieval system. Raw corpus refers to unannotated character sequences extracted

from texts. To ensure vocabulary coverage and diachronic span, NEPD includes all articles from nine months: January–June 2015, and January 2016, 2017, and 2018. A sample of the acquired data source is shown in Figure 1 [Figure 1: see original paper].

The acquisition process involves: (1) determining the target time periods and organizing personnel to download all articles from these periods; (2) storing all raw corpus uniformly as text files while preserving original paragraphing and formatting to facilitate manual annotation; (3) organizing all text files by month to form complete raw corpora for each period.

3.2 Data Preprocessing

The raw corpus requires preprocessing: (1) removing non-body content such as page headers like "People's Daily 2015.01.27 Page 6 Feng Hua" that were inadvertently copied; (2) unifying character encodings to UTF-8 for consistent subsequent processing. After preprocessing, the annotated *People's Daily* corpus is obtained, with sample examples shown in Table 1 .

4. Corpus Annotation and Specifications

4.1 Annotator Training

To ensure NEPD annotation quality, annotators receive comprehensive training in knowledge, skills, and specifications: (1) Knowledge requirements: understanding modern Chinese vocabulary definitions, systems, and linguistic theories; the value of word segmentation in Chinese information processing; definitions of segmentation inconsistency; and distinctions between combinatory and intersectional ambiguities. (2) Skill requirements: ability to implement programs for word frequency statistics and Zipf's law analysis, maximum matching segmentation algorithms, and rule-based ambiguity resolution. (3) Specification requirements: thorough familiarity with the national standard "Modern Chinese Word Segmentation Specification for Information Processing" (GB/T13715-92) and ability to apply its examples to new cases.

4.2 Annotation Process

After training, annotation proceeds through three steps: (1) First-pass segmentation by Group 1, marking word boundaries with "/". For example, "坚持依法治国、依法执政、依法行政共同推进" becomes "坚持/依法/治国/、/依法/执政/、/依法/行政/共同/推进". (2) Second-pass verification by Group 2, checking compliance with specifications. For instance, the idiom "扎扎实实" must be annotated as one word, not split. (3) Third-pass verification by Group 3 to ensure final accuracy.

Following three rounds of annotation, specialized programs perform machine proofreading of punctuation, as annotators often focus on words and overlook punctuation. After these steps, the final annotated NEPD corpus is obtained, with sample results shown in Table 2 .

4.3 Special Annotation Specifications

NEPD adopts special specifications for certain cases: (1) Personal names are segmented into surname and given name separately to facilitate future part-of-speech tagging and surname statistics, and to enable comparison with the 1998 corpus. (2) Idioms are treated as complete words based on semantic compositionality and convention, while longer proverbs and colloquialisms are split into multiple words. (3) Number-measure word combinations are consistently segmented, e.g., "一名" (one) is split into "一/名".

5. NEPD Segmentation Experiments and Performance Evaluation

5.1 Experimental Design

J. Lafferty et al. proposed Conditional Random Fields in 2001 as a conditional probability model for sequence labeling and segmentation [14]. To evaluate NEPD's performance, we extract January 2018 data from NEPD for comparison with Peking University's January 1998 corpus. Experiments use a self-developed platform encapsulating CRF++ 0.58, a widely-used, highly available open-source toolkit with excellent usability, accuracy, stability, and portability for NLP tasks including segmentation and named entity recognition.

The experimental approach involves: (1) designing tag sets and feature templates based on corpus characteristics; (2) processing selected corpora into CRF++-recognizable formats; (3) constructing feature templates through feature combination; (4) training segmentation models on training data; (5) applying models to similarly processed test data; (6) evaluating performance using precision, recall, and F-score:

Precision (P) = Correctly tagged tokens / Total tagged tokens × 100%
Recall (R) = Correctly tagged tokens / Total tokens that should be tagged × 100%
F-score (F) = (2×P×R)/(P+R) × 100%

5.2 Model Construction and Performance Comparison

For fair comparison, both January 2018 and January 1998 corpora are randomly divided into 10 equal parts at a 1:9 test-to-train ratio. Only character-level features are used without additional features. Based on word length distribution in People's Daily, a 4-tag set is adopted: B (word beginning), M (word middle, unlimited repetition based on length), E (word end), and S (single-character word). Tag semantics are shown in Table 4 .

Using CRF++ on training data produces segmentation models that output tag sequences to test datasets. Tags are placed in the final column, and characters are combined into words based on tag sequences to segment both corpora. For

consistent comparison, identical tag sets and feature templates are used across experiments.

5.3 Performance Results

Ten segmentation models are built from each corpus, with overall performance shown in Tables 5 and 6 .

1998 Corpus Results: The best model achieves 97.18% F-score, with average F-score of 97.10%. For multi-character words, initial-character F-score reaches 97.79% (average 97.74%), middle-character 93.69% (average 93.47%), and final-character 97.76% (average 97.71%). Middle-character performance affects overall multi-word F-score due to lower recall (minimum 93.27%), caused by long-span words like foreign names.

2018 Corpus Results: The optimal model achieves 97.80% F-score, 0.62% higher than the 1998 best model, with average F-score of 97.74% (0.63% higher). Initial-character F-score reaches 98.59% (average 98.54%), 0.8% higher than 1998. Middle-character F-score reaches 90.86% (average 90.64%), 2.83% lower than 1998, attributed to increased vocabulary length span. Final-character F-score reaches 98.46% (average 98.41%), 0.70% higher than 1998. Single-character word recognition achieves 97.78% maximum and 97.68% average F-score, 1.01% and 1.17% higher than 1998, respectively.

5.4 Cross-Corpus Validation

To demonstrate NEPD's necessity, the best 1998-based model is applied to January 2018 test data. Results in Table 7 show dramatic performance drops: maximum F-score of 83.26% and average of 83.06%, 14.54% and 14.68% lower than 2018-based models. This indicates the 1998 corpus cannot accurately annotate current texts due to insufficient vocabulary coverage and novelty. Middle-character performance is particularly poor, with maximum F-score of only 44.06% and average of 43.62%, 46.80% and 47.02% lower than 2018 models, proving the 1998 model cannot handle longer words in contemporary texts.

These results technically validate the necessity of constructing NEPD.

6. Conclusion

This paper presents NEPD's data sources, cleaning process, annotation specifications, and workflow. Using CRF models, we validate NEPD's performance from two dimensions, demonstrating its outstanding quality and necessity. NEPD addresses the limitations of the Peking University corpus for current texts, providing diachronic continuity and effective expansion of existing People's Daily corpora. It offers robust resource support for developing high-performance named entity recognition models, precise semantic retrieval systems, and shallow syntactic parsers. Future research should focus on expanding corpus scale and further improving annotation precision.

References

[1] Yu Shiwen, Duan Huiming, Zhu Xuefeng. Basic processing specifications for Peking University modern Chinese corpus[J]. Journal of Chinese Information Processing, 2002(5): 49-64.
[2] Wang Hongjun, Shi Shuicai, Yu Shiwen. Research on indexing methods for People's Daily annotated corpus[C]//Proceedings of the 8th Joint Conference on Computational Linguistics (JSCL-2005). Nanjing: Nanjing Normal University, 2005: 576-578.
[3] National Language Commission. National Language Commission Modern Chinese Corpus[EB/OL].[2019-06-02]. http://www.cncorpus.org/.
[4] Zhou Qiang. Chinese syntactic treebank annotation system[J]. Journal of Chinese Information Processing, 2004, 18(4): 2-9.
[5] ANTONY P J, WARRIER N J, SOMAN K P. Penn treebank-based syntactic parsers for South Dravidian languages using a machine learning approach[J]. International journal of computer applications, 2010, 7(8): 14-21.
[6] HUANG C, ZHAO H. Chinese word segmentation: a decade review[J]. Journal of Chinese information processing, 2007, 21(3): 8-19.
[7] Li Shuanglong, Liu Qun, Wang Chengyao. Chinese word segmentation system based on conditional random fields[J]. Microcomputer Information, 2006(10): 178-180.
[8] Shen Qinzhong, Zhou Guodong, Zhu Qiaoming, et al. Conditional random field-based Chinese word segmentation method using character position probability features[J]. Journal of Soochow University: Natural Science Edition, 2008, 24(3): 49-54.
[9] Chi Chengying, Yu Changyuan, Zhan Xuegang. Chinese word segmentation method based on conditional random fields[J]. Intelligence Magazine, 2008, 27(5): 79-81.
[10] Song Yan, Cai Dongfeng, Zhang Guiping, et al. A Chinese word segmentation method based on joint decoding of characters and words[J]. Journal of Software, 2009(9): 2366-2375.
[11] Liu Zewen, Ding Dong, Li Chunwen. Chinese short text segmentation method based on conditional random fields[J]. Journal of Tsinghua University (Science and Technology), 2015, 55(8): 906-910, 916.
[12] Feng Xue. Comparison of dictionary integration methods in Chinese word segmentation models[J]. Computer Application Research, 2019, 36(1): 14-16.
[13] Wang Ruojia, Zhao Changyu, Wang Jimin. Research on Chinese electronic medical record segmentation and entity recognition[J]. Library and Information Service, 2019, 63(2): 34-42.
[14] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceeding of international conference on machine learning. Williamstown: International Machine Learning Society, 2001: 282-289.

Author Contributions

Huang Shuiqing: Conceptualization, research design, revision and finalization of manuscript.
Wang Dongbo: Data processing and initial draft preparation.

Abstract: [Purpose/Significance] The construction of a segmented corpus of *People's Daily* adapted to the new era provides newly annotated data for Chinese information processing and offers new language resources for analyzing modern Chinese from a diachronic perspective. [Method/Process] Based on analysis of existing Chinese word segmentation corpora, we describe the data source, annotation specifications, and construction process of the New Era People's Daily Corpus, evaluate its performance through automatic segmentation models, and compare it with existing corpora. [Result/Conclusion] The New Era People's Daily Segmented Corpus (NEPD) follows modern Chinese corpus processing standards, featuring large scale and long time span. Selecting the January 2018 portion, we build a CRF-based segmentation model and compare its performance with the January 1998 People's Daily corpus. The evaluation metrics demonstrate NEPD's outstanding performance, confirm that the 1998 corpus cannot be replaced, and prove the necessity of constructing NEPD.

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