IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 October 2023, accepted 13 November 2023, date of publication 16 November 2023,
date of current version 27 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3333338

METHODS

Text Segmentation via Hierarchical Document Attention Model

YANHUA WANG AND CHUNFANG MIN
Faculty of Arts, Lanzhou University, Lanzhou 730000, China

Corresponding authors: Yanhua Wang (dpengzhanchi@126.com) and Chunfang Min (feifei0001feifei@163.com)

This work was supported in part by the National Social Science Fund Major Bidding Project on Research on the Language of Hui Dialects in Northwest Ethnic Areas, Hui Jing Tang Language, and Children's Classic Language under Grant 17ZDA311; in part by the Key Project of Central Universities at Lanzhou University under Grant 17LZUJBWZD004 and Grant 2018skzy023; in part by the Special Fund for Basic Scientific Research Business Expenses of Central Universities; and in part by the Fundamental Research Funds for the Central Universities under Grant 2019jbkyzx007 and Grant 2019jbkyzx016.

ABSTRACT With the rapid development of natural language processing technology, text segmentation has become an important task in text processing. However, existing text segmentation methods often perform poorly when faced with long texts and complex structures, requiring a more efficient and accurate approach. In this paper, we propose a new text segmentation method based on the Hierarchical Document Attention (HDA), which automatically identifies and segments different paragraphs in the text by analyzing and weighting the hierarchical structure of the text sequence data. Compared with existing methods, the model has higher accuracy and efficiency, and better supports tasks such as text analysis and information extraction. The main contribution of this paper is the proposal of a text segmentation method based on the HDA, which effectively models text sequences through multi-level attention mechanisms. Experimental verification on public datasets shows that this model exhibits good performance in text segmentation tasks.

INDEX TERMS Natural language processing, text segmentation, hierarchical document attention, attention mechanism.

I. INTRODUCTION

As natural language processing technology advances rapidly, text segmentation [1], [2], [3] has emerged as a crucial task in text processing. However, existing text segmentation methods face challenges such as long texts, complex structures, and limited efficiency and accuracy. To address these limitations, we propose a novel text segmentation method based on the Hierarchical Document Attention (HDA) approach.

Compared to traditional text segmentation methods that rely on rules, statistics, and machine learning techniques [1], [2], [3], [4], [5], our HDA-based method offers several advantages. Firstly, rule-based methods often struggle with complex text structures [1], [2], requiring predefined rules. In contrast, HDA automatically identifies and segments paragraphs by analyzing the hierarchical structure of the text sequence data, eliminating the need for predefined rules and effectively handling complexity.

Secondly, statistical methods may accumulate errors when processing long texts [3], [4], resulting in decreased segmentation accuracy. In contrast, HDA's multi-level attention mechanisms consider the hierarchical nature of the text, enabling more accurate modeling of the text sequence and improving segmentation performance.

Thirdly, machine learning methods typically demand a large amount of training data and extensive feature engineering [2], [5], limiting their flexibility. In contrast, HDA's multi-level attention mechanisms allow it to adapt to various text structures without extensive feature engineering, providing a more flexible and efficient approach to text segmentation.

Experimental results confirm the superiority of our HDA-based method over existing approaches. It achieves higher levels of accuracy and efficiency in text segmentation tasks, while also better supporting tasks such as text analysis

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Loconsole.

and information extraction. By effectively modeling text sequences through multi-level attention mechanisms, our method demonstrates its potential and importance in the field of natural language processing.

Additionally, the HDA approach proposed in this paper distinguishes itself from the Hierarchical Attention Network (HAN) by incorporating non-parametric unsupervised Bayesian methods. This utilization of Bayesian techniques enhances the flexibility and accuracy of HDA in text segmentation tasks. By leveraging these non-parametric methods, HDA can adapt to various text structures without relying on predefined rules or extensive feature engineering, allowing for a more robust and accurate segmentation process. The incorporation of non-parametric unsupervised learning enables HDA to effectively capture the underlying patterns and hierarchical dependencies within the text, resulting in improved segmentation accuracy and overall performance.

The paper is organized as follows: Section II introduces preliminaries. Section III describes the HDA. In Section IV, we describe the inference process. The experiment and analysis are described in Section V. We summarize this research in Section VI and point out our future work.

II. RELATED WORK

Currently, research on text segmentation primarily focuses on three categories: rule-based methodologies, statistical approaches, and machine learning techniques.

Rule-based methods: Utilizing predefined language rules and patterns to determine the segmentation points of the text. For instance, ''Text segmentation of health examination item based on character statistics and information measurement'' [1] proposes a character-based approach for segmenting lengthy health examination texts, achieving a precision of 78.6%. This study validates the valuable clues for text comprehension in historical data without requiring domain knowledge. ''Touching text line segmentation combined local baseline and connected component for Uchen Tibetan historical documents'' [2] introduces a method for segmenting text lines in Tibetan historical documents that are distorted, broken, or connected. By detecting pseudo text lines, handling connected regions, and segmenting connected components, this approach achieves accurate segmentation with high robustness and precision. These methods significantly contribute to the development of text segmentation research.

Statistical methods: Segmenting text through analyzing statistical features and probability models. For example, ''HMM-BiMM: Hidden Markov Model-based word segmentation via improved Bi-directional Maximal Matching algorithm'' [6] introduces a novel word segmentation algorithm that combines a hidden Markov model with an enhanced bi-directional maximal matching algorithm (HMM-BiMM). This algorithm achieves the highest efficiency and accuracy in symptom text segmentation, with a 3% precision improvement and approximately 70% reduction in runtime compared to the BiMM algorithm. ''DEFT: A corpus for definition extraction in free- and semi-structured text'' [7] presents a robust English corpus and annotation pattern for exploring non-intuitive term-definition structures in free and semi-structured text. It overcomes the challenges posed by complex and unstructured natural language data with practical solutions and real-world data. These studies offer new perspectives and methodologies for text segmentation development.

Machine learning methods: Leveraging machine learning algorithms and training data to learn text segmentation patterns and rules, enabling automated segmentation. For instance, ''Mask TextSpotter v3: Segmentation Proposal Network for Robust Scene Text Spotting'' [8] proposes an end-to-end trainable method for scene text localization, replacing region proposal networks with segmentation proposal networks (SPN). It achieves outstanding results on text instances with extreme aspect ratios or irregular shapes. ''Scene Text Segmentation via Multi-Task Cascade Transformer With Paired Data Synthesis'' [9] introduces a method that explicitly learns text attributes by generating paired data and utilizing a multi-task cascade Transformer network. It outperforms existing methods in scene text segmentation tasks. ''An Empirical Study of TextRank for Keyword Extraction'' [10] conducts an empirical study optimizing TextRank parameters and identifies the optimal settings for keyword extraction, demonstrating the best TextRank performance regardless of text length. These novel algorithms have made significant contributions to the field of text processing, offering new avenues for text segmentation research.

III. PRELIMINARIES

A. ATTENTION MECHANISM

Attention mechanism [11], [12], [13], [14], [15], [16], [17] is a widely used technique in the fields of machine learning and natural language processing. It enables models to focus more accurately on relevant parts of input sequences while processing them. In natural language processing, attention mechanism compares each element of the input sequence with the current state and calculates a weight that represents the importance of that element to the current state. These weights can be used to weight the elements in the input sequence, resulting in a weighted vector that serves as the model's output. The basic formula for attention mechanism is as follows:
address memory (score function):

$$e_{ij} = F(\bar{h_i}, h_j) \tag{1}$$

normalize (alignment function):

$$y_{ij} = softmax(e_{ij}) = \frac{exp(e_{ij})}{\Sigma_{k=1}^{K} exp(e_{ik})} \tag{2}$$

read content (generate context vector function):

$$c_i = \sum_{j=1}^{J} \gamma_{ij} j[j] \tag{3}$$

h_j denotes the hidden state of the decoder, $\bar{h_j}$ denotes the hidden state of the encoder, F denotes the scoring function, e_{ij} denotes the attention score, γ_{ij} denotes the attention weight obtained by normalizing the evaluation results using the softmax function, and c_i represents the output generated based on the attention weight. By calculating e_{ij} using the scoring function F, the model can assess the importance of each part of the input sequence, and then normalize the evaluation results using the softmax function to obtain the corresponding attention weight γ_{ij}, which is used to generate the output c_i.

The attention mechanisms can effectively process long sequences and reduce attention to irrelevant information, thereby significantly improving the efficiency and accuracy of task processing. Additionally, attention mechanisms can enhance the robustness and generalizability of a model, making it more suitable for diverse application scenarios.

B. HIERARCHICAL DIRICHLET PROCESS

The Dirichlet Process (DP) [18], [19], [20] is a sophisticated stochastic distribution in probability theory that serves as an infinite mixture model capable of generating infinite categories. However, when dealing with large datasets, it may result in an excessive number of categories that are difficult to interpret. To address this issue, the Hierarchical Dirichlet Process (HDP) [19], [21], [22], [23], [24] extends the DP by introducing an additional layer of random processes to limit the number of generated categories. It first selects a global DP as the high-level process, and then chooses several sub-DPs as the low-level processes to ensure that each dataset has enough categories to accurately describe its features while also ensuring that the number of categories is reasonable. The HDP can more accurately interpret data, effectively reduce the number of categories, make the model more compact, better describe data features, and discover correlations between data to better utilize them for inference and prediction.

In the HDP, a base distribution G_h is constructed with a concentration parameter α_h to create $G_0 \sim DP(\alpha_h, G_h)$. Then, $G_j \sim DP(\alpha_0, G_0)$ is built for each group of data using G_0 as the prior distribution. This way, each group of data has an independent DP, and the topics of each document conform to the G_h distribution. Finally, the Dirichlet Process mixture model is constructed using G_j as the prior distribution, as shown in the following formula:

$$\begin{aligned} &W_{ji} \sim F(\theta_{ji}) \\ &\theta_{ji}|G_j \sim G_j \\ &G_j|G_0 \sim DP(\alpha_0, G_0)\, for\, j = 1, 2, 3, \ldots, J \\ &G_0 \sim DP(\alpha_h, G_h) \end{aligned} \tag{4}$$

To represent its probability distribution using stick-breaking, it is sampled in a uniform manner. The formula is as follows:

$$\begin{aligned} &\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \sigma_{\vartheta_k} \\ &G = \sum_{j=1} \pi_k \\ &\beta_k \sim Beta(1, \alpha) \\ &\vartheta_k \sim G_0 \end{aligned} \tag{5}$$

Adopting HDP can better comprehend the hierarchical structure of data, further improving the accuracy and efficiency of data modeling and analysis. Moreover, HDP has broad prospects for application, bringing important development opportunities to multiple fields and promoting the continuous advancement of data science and artificial intelligence technology.

IV. PROPOSED MODEL

In the realm of text segmentation tasks, Hidden Markov Model (HMM) [25], [26], [27], [28], [29] is widely adopted as a generative model for modeling sequence data. HMM models the discrete state sequence as a double stochastic process of Markov chains, where each hidden HMM state represents a theme. For each state, there exists an emission probability distribution function (PDF) denoted as B, which can be inferred from the n-gram topic-related word distribution. A transition matrix A of size $N \times N$ is utilized to model the transition probability between states.

The HDA, a text segmentation model based on the attention mechanism, is a probabilistic graphical model that incorporates both HMM and attention mechanism to better handle text segmentation tasks. In the HDA, each word is assigned a probability indicating its likelihood of belonging to each theme. The model dynamically adjusts the weight of each word through learning attention weights, thus better capturing the semantic information within the text. The graphical model of the proposed model is shown in *Fig*.1.

The HDA employs hierarchical analysis and attention weighting to automatically recognize and segment text sequence data. Specifically, the model divides the text sequence into multiple levels and assigns weights to each level using a multi-level attention mechanism, resulting in an effective representation of the text sequence. In the HDA, words are conditionally independent observations given the state, denoted as $w_t|z_t, m_t, \quad m_t \sim b(\lambda_{z_t}, m_t)$, where λ_{z_t} and m_t represent the distribution parameters of state z_t and mixture component m_t. The transition probability from state z_{t-1} to z_t only depends on the previous state and is represented as $z_t|z_{t-1} \sim \pi_{z_t, z_{t-1}}$.

In the HDA, an attention mechanism can be utilized to compute the similarity between words and states, resulting in a better text sequence representation. For each word t and each HMM state s, the attention score $e_{t,s}$ can be calculated

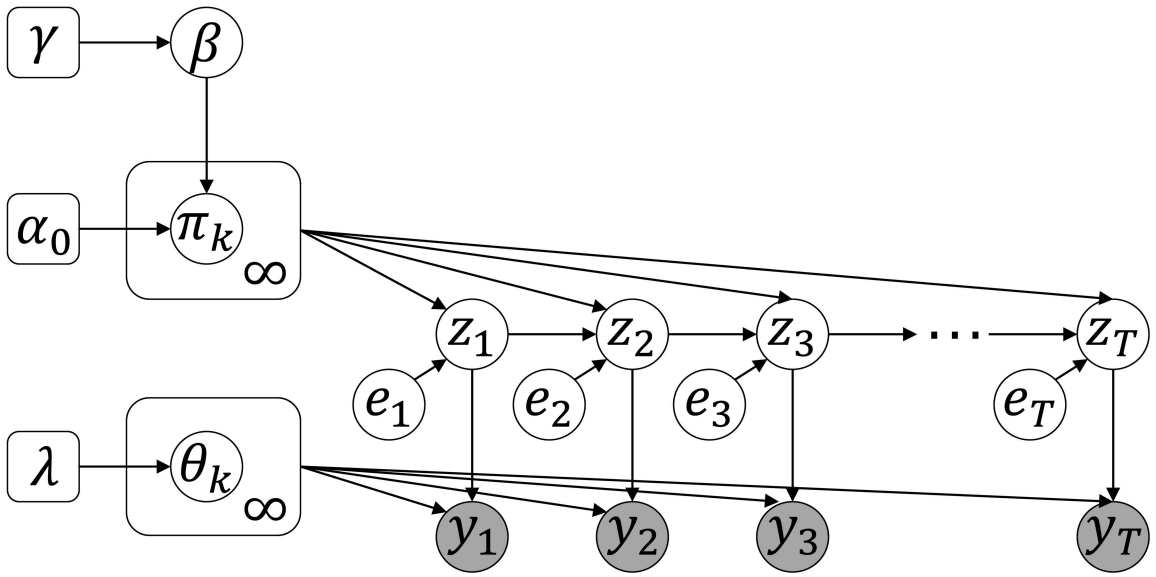

FIGURE 1. The probabilistic graphical model of the HDA: Capturing hierarchical document attention for text segmentation.

using the following formula:

$$e_{t,s} = \alpha(s) \cdot \beta(s, w_t) \tag{6}$$

$\alpha(s)$ denotes the prior probability of state s, which can be modeled using a Dirichlet process, and $\beta(s, w_t)$ denotes the similarity between word t and state s, which can be computed using a neural network [30], [31], [32]. Specifically, a Bidirectional Long Short-Term Memory network (BiLSTM) [33], [34] can be employed to compute the context vector, followed by a feedforward neural network [35] to calculate the similarity:

$$\beta(s, w_t) = \text{softmax}(\text{MLP}([\text{BiLSTM}(w_{t-k}, \ldots, w_{t+k}); \theta_s])) \tag{7}$$

θ_s denotes the attention parameter of state s, and MLP refers to a multi-layer perceptron used to compute the similarity. Finally, the attention weight $\alpha_{t,s}$ between word t and state s is computed based on the attention score $e_{t,s}$:

$$\alpha_{t,s} = \frac{\exp(e_{t,s})}{\sum_{s'} \exp(e_{t,s'})} \tag{8}$$

In this way, the attention weight $\alpha_{t,s}$ can be utilized to update the state transition probability in the HMM-HDP model, resulting in a better representation of the text sequence.

V. INFERENCE

With a given word sequence $W = [w_1, w_2, \ldots, w_T]$ and a trained HMM, we can employ a topic label sequence $Z = [z_1, z_2, \ldots, z_T]$ by solving the following optimization problem:

$$\hat{z} = \max_{z} \sum_{t=1}^{T} \log p(w_t|z_t) + \log p(z_t|z_{t-1}) \tag{9}$$

The probability of word w_t given topic z_t, denoted as $p(w_t|z_t)$, can be computed using a topic-dependent language model. Similarly, the transition probability from state z_{t-1} to z_t, denoted as $\log p(z_t|z_{t-1})$, can be calculated based on the state transition probability in the HMM model.

By the application of the Bayesian rule, the aforementioned optimization problem is equivalent to the following formula:

$$\hat{z} = \max_{z} \sum_{t=1}^{T} \log p(w_t|z_t) + \log p(z_t|z_{t-1}) + \log p(z_1) \tag{10}$$

$\log p(z_1)$ denotes the prior probability of the state sequence, which can be modeled using DP. To perform inference on the formula using variational inference, it is necessary to first decompose the joint probability distribution:

$$\begin{aligned}
&\log p(W, Z) \\
&= \sum_{t=1}^{T} \log p(w_t|z_t) + \sum_{t=2}^{T} \log p(z_t|z_{t-1}) + \log p(z_1) \\
&= \sum_{t=1}^{T} \log p(w_t|z_t) + \sum_{t=2}^{T} \log p(z_t|z_{t-1}) + \log p(z_1|z_0) \\
&= \sum_{t=1}^{T} \log p(w_t|z_t) + \sum_{t=1}^{T} \log p(z_t|z_{t-1}) - \sum_{t=2}^{T} \log p(z_{t-1}|z_t) \\
&\quad + \log p(z_1|z_0) - \log p(z_1) + \log p(z_T|z_{T-1}) \\
&\quad - \log p(z_T|z_{T-2})
\end{aligned} \tag{11}$$

z_0 denotes the initial state. Variational inference can be used to estimate the posterior distribution $q(Z)$, which maximizes the lower bound $L(q)$.

$$L(q) = \mathbb{E}_q(Z)[\log p(W, Z)] = \mathbb{E}_q(Z)[\log q(Z)] \tag{12}$$

$\mathbb{E}_q(Z)[\log p(W, Z)]$ represents the expectation of the joint distribution $p(W, Z)$ with respect to the posterior distribution

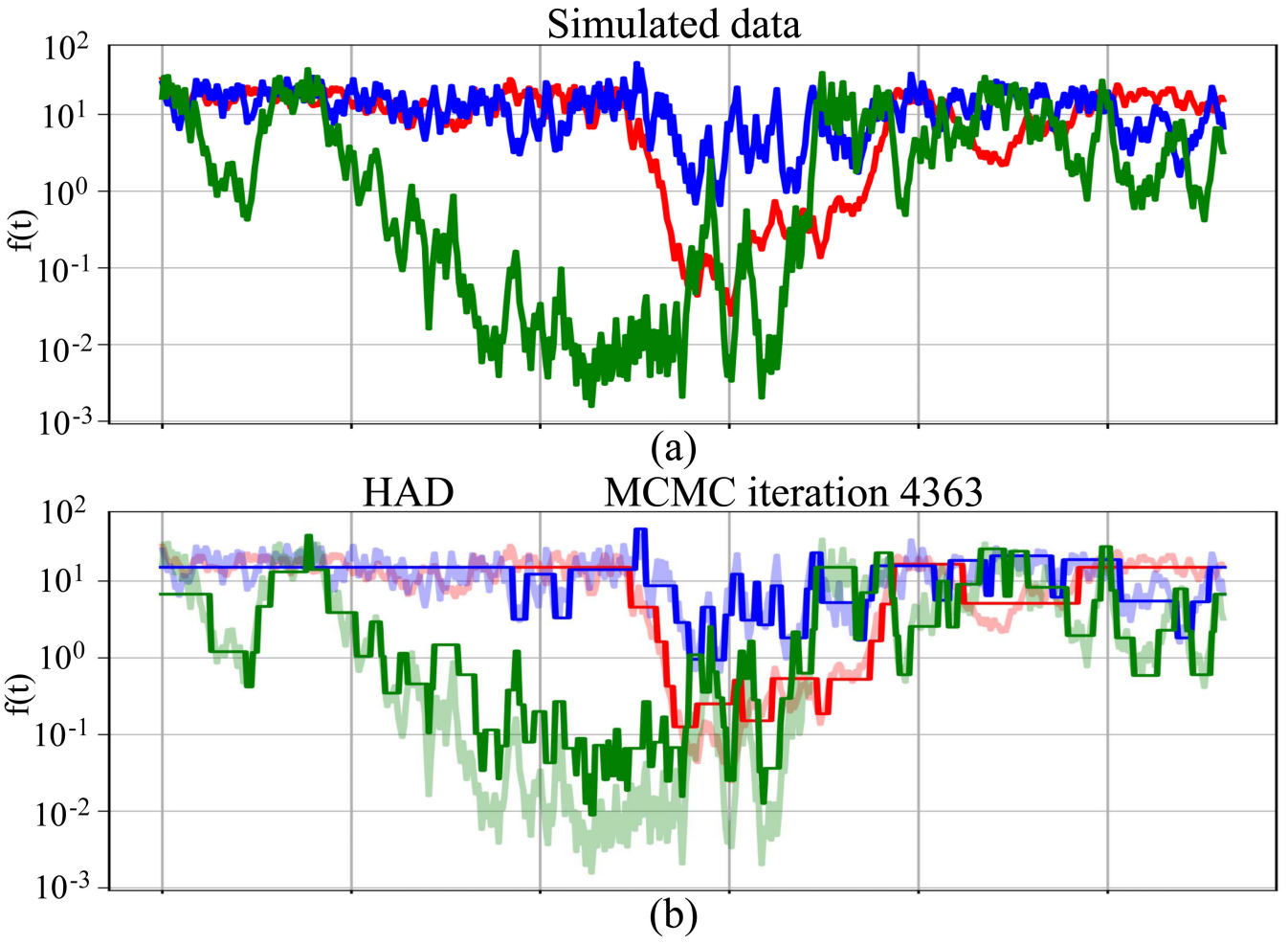

FIGURE 2. The result of HDA with synthetic data.

$q(Z)$, which can be computed using sampling or approximation methods. On the other hand, $\mathbb{E}_q(Z)[\log q(Z)]$ denotes the entropy of the posterior distribution $q(Z)$ with respect to itself, which can be directly computed. We can further expand $L(q)$ as follows:

$$\begin{aligned} L(q) = &\sum_{t=1}^{T} \mathbb{E}q(Z_t)[\log p(w_t|Z_t)] \\ &+ \sum_{t} \mathbb{E}q(Z_t, Z_{t-1})[\log p(Z_t|Z_{t-1})] \\ &- \sum_{t=2}^{T} \mathbb{E}q(Z_t, Zt-1)[\log p(Z_{t-1})] \\ &+ \log p(z_1|z_0) - \log p(z_1) + \log p(z_T|z_{T-1}) \\ &- \log p(z_T|z_{T-2}) - \mathbb{E}[\log q(Z)] \end{aligned} \tag{13}$$

We can represent each expectation term in the above equation as a variational factor and a corresponding conditional distribution:

$$\begin{aligned} &\mathbb{E}q(Z_t)[\log p(w_t|Z_t)] \\ &\quad = \mathbb{E}q(Z_t, Z_{t-1})[\log p(Z_t|Z_{t-1})] \\ &\quad = \sum_{k=1}^{K}\sum_{j=1}^{K} q(z_{t-1}=j, z_t=k) \log p(z_k|z_{t-1}=j) \\ &\qquad + \sum_{k=1}^{K}\sum_{j=1}^{K} q(z_t=k, z_{t-1}=j) \log p(z_{t-1}=j|z_t=k) \end{aligned} \tag{14}$$

K denotes the number of topics. The following formula is used to update $L(q)$:

$$\begin{aligned} q(z_{t=k}) \propto \exp(&\sum_{j\neq t}\sum_{l=1}^{K} \mathbb{E}q(z_j)[\log p(z_j=l|z_{j-1})] \\ &+ \mathbb{E}q(z_{t+1})[\log p(z_{t+1}|z_t=k)] + \log p(w_t|z_t=k)) \end{aligned} \tag{15}$$

$\mathbb{E}q(z_j)[\log p(z_j = l|z_{j-1})]$ denotes the marginal probability of topic l, computed under the conditional distribution $q(z_j)$ that fixes the other variational factors at the current iteration. This allows us to obtain the update formula for the variational factor $q(z_t)$. Similarly, we can use the following formula to

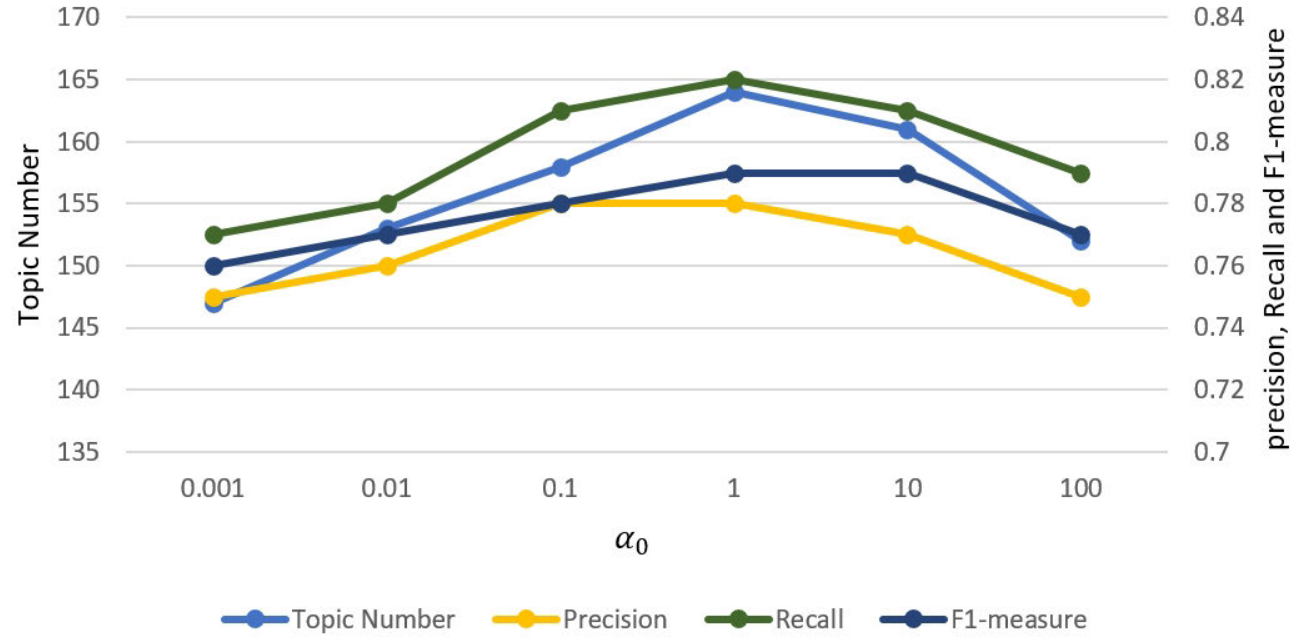

FIGURE 3. Precision, Recall and F1-measure with different value of α_0.

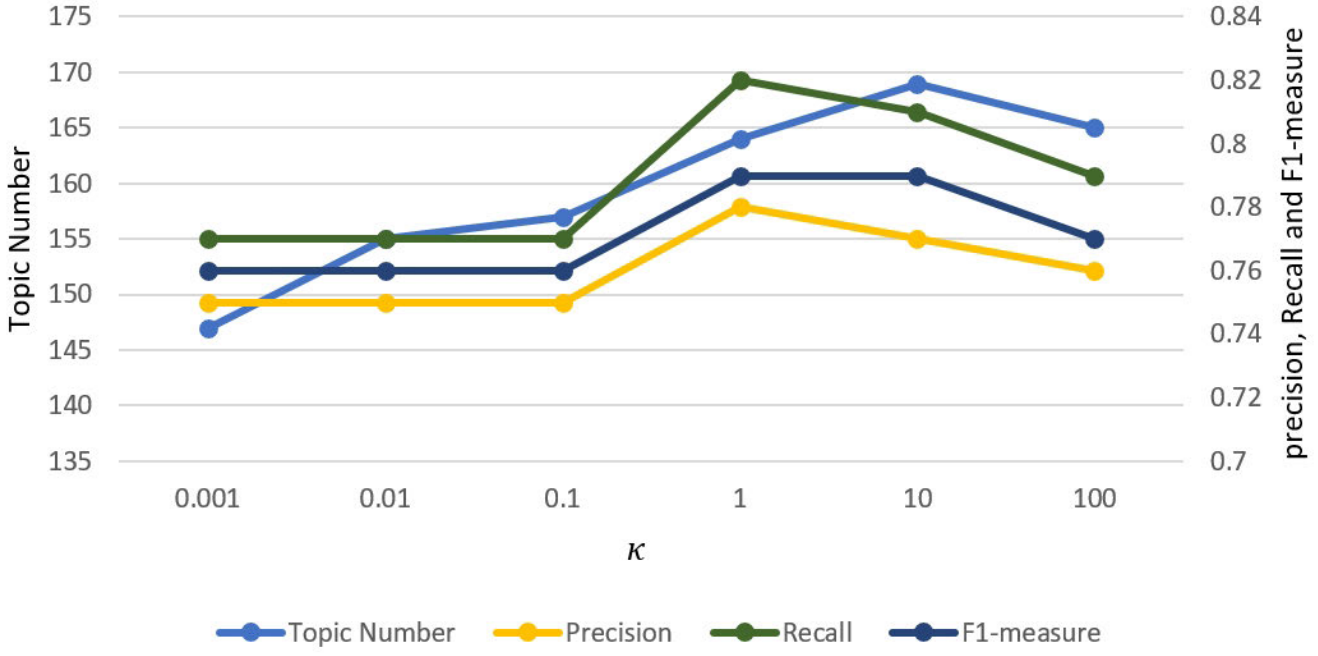

FIGURE 4. Precision, Recall and F1-measure with different value of K.

update $q(z_{t-1}, z_t)$:

$$q(z_{t-1} = j, z_t = k) \propto \exp\Big(\sum_{l \neq t-1} \sum_{i=1}^{K} \mathbb{E}q(z_i)[\log p(z_i|zi-1)] + \log p(z_t = k|z_{t-1} = j) + \log p(w_t|z_t = k)\Big) \quad (16)$$

$\mathrm{Eq}(z_i)[\log p(z_i|z_{i-1})]$ denotes the marginal probability of topic i, computed under the conditional distribution $q(z_i)$ that fixes the other variational factors at the current iteration. This allows us to obtain the update formula for the variational factor $q(z_{t-1}, z_t)$.

After updating the conditional distributions of all variational factors, we can compute the lower bound $L(q)$ and check for convergence. Upon convergence, we can label the topic categories.

VI. EXPERIMENTS

To establish the effectiveness of HDA, this section conducts a thorough investigation of its functional characteristics and verifies them through the following experimental design.

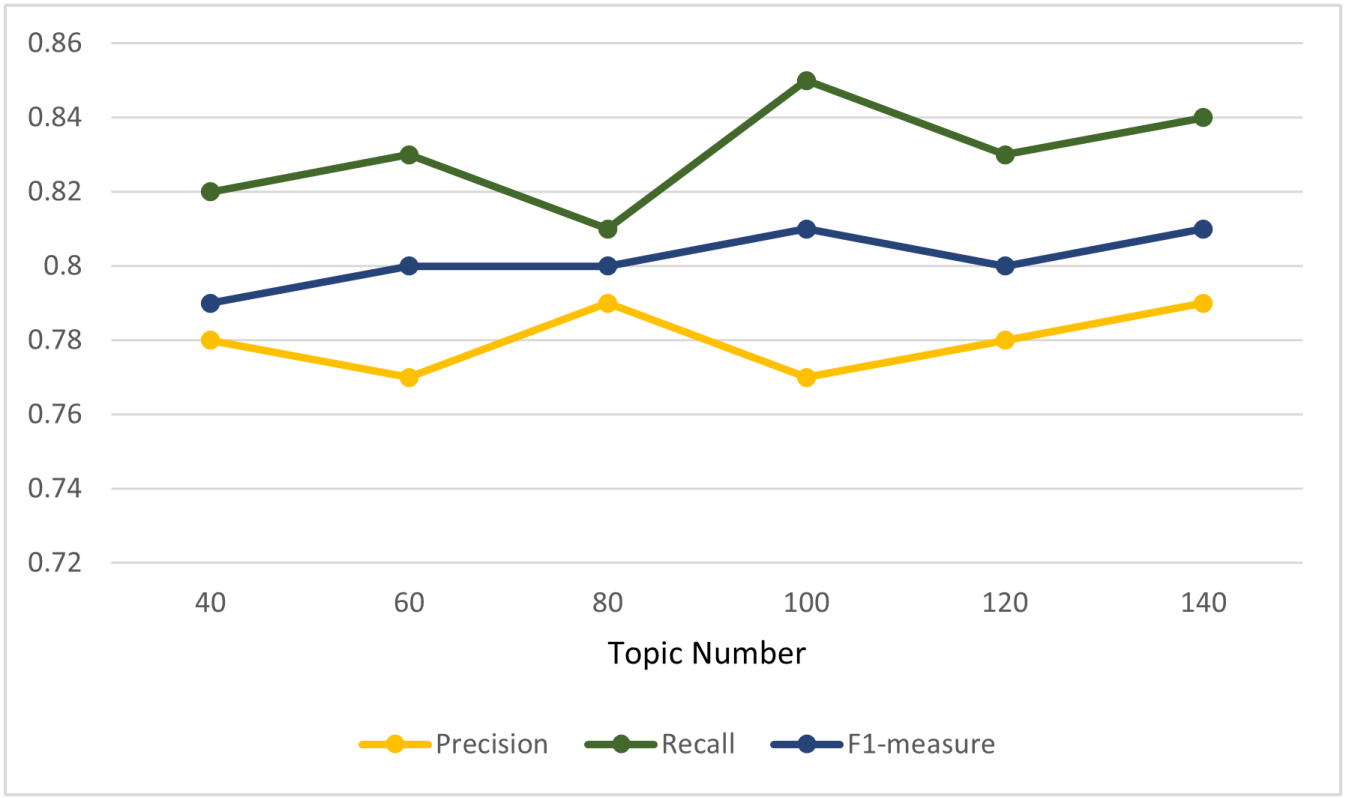

FIGURE 5. Precision, Recall and F1-measure with different number of topics.

A. EXPERIMENTAL DATASET

In the experiment, a synthetic dataset of 562 samples with 3 dimensions is utilized to explore the performance of HDA. Each sample is assumed to follow a Gaussian distribution. We use HDA to infer the probability distribution of the data and compare it with the actual distribution. The hyperparameters $\alpha = 1$, $K = 1$, and $\gamma = 1$ are set for the experiment. *Fig*.2 shows the results of 4363 sampling iterations of HDA. *Fig*.2(a) visualizes the value of the synthetic data in three dimensions. The x-axis represents time, while the y-axis represents the value of the data in the three dimensions, where each value is denoted by red, blue, and green. The corresponding probability distribution of the data values in the three dimensions is depicted on the right-hand side. *Fig*.2(b) shows the inferred distribution by HDA and compares it with the actual distribution of the synthetic data. It can be observed from (a) and (b) that the shapes of the corresponding curves are very similar, indicating that the inferred probability distribution by HDA is consistent with the actual probability distribution of the training data. This preliminary result demonstrates the effectiveness of HDA.

B. BASELINE MODEL

In the experimental section, we compare our proposed HDA approach with two baseline methods: Broadcast news story segmentation using sticky hierarchical Dirichlet process (SHDP-HMM) [36] and Semantic Segmentation of Text using Deep Learning (SSOT-DL) [37].

SHDP-HMM is chosen as a baseline due to its effectiveness in segmenting broadcast news stories. It utilizes a sticky hierarchical Dirichlet process (SHDP) to model the topic structure of the news stories and employs a hidden Markov model (HMM) to capture the sequential dependencies within the text. By leveraging the SHDP-HMM model, it aims to identify coherent segments within the news stories.

On the other hand, SSOT-DL is selected as a baseline method for its capability to perform semantic segmentation of text using deep learning techniques. It employs deep neural networks to learn the representations and semantic structures of the text, enabling it to identify and segment text based on semantic boundaries and context.

By comparing HDA with these two baseline methods, we aim to evaluate the performance and effectiveness of our proposed approach in text segmentation tasks. The experimental results will provide insights into the strengths and limitations of each method, highlighting the advantages of HDA in terms of accuracy, efficiency, and adaptability in handling complex text structures.

Next, we will present the details of the experimental setup, including the datasets used, evaluation metrics, and implementation specifics, followed by a comprehensive analysis and discussion of the results obtained from comparing HDA with SHDP-HMM and SSOT-DL.

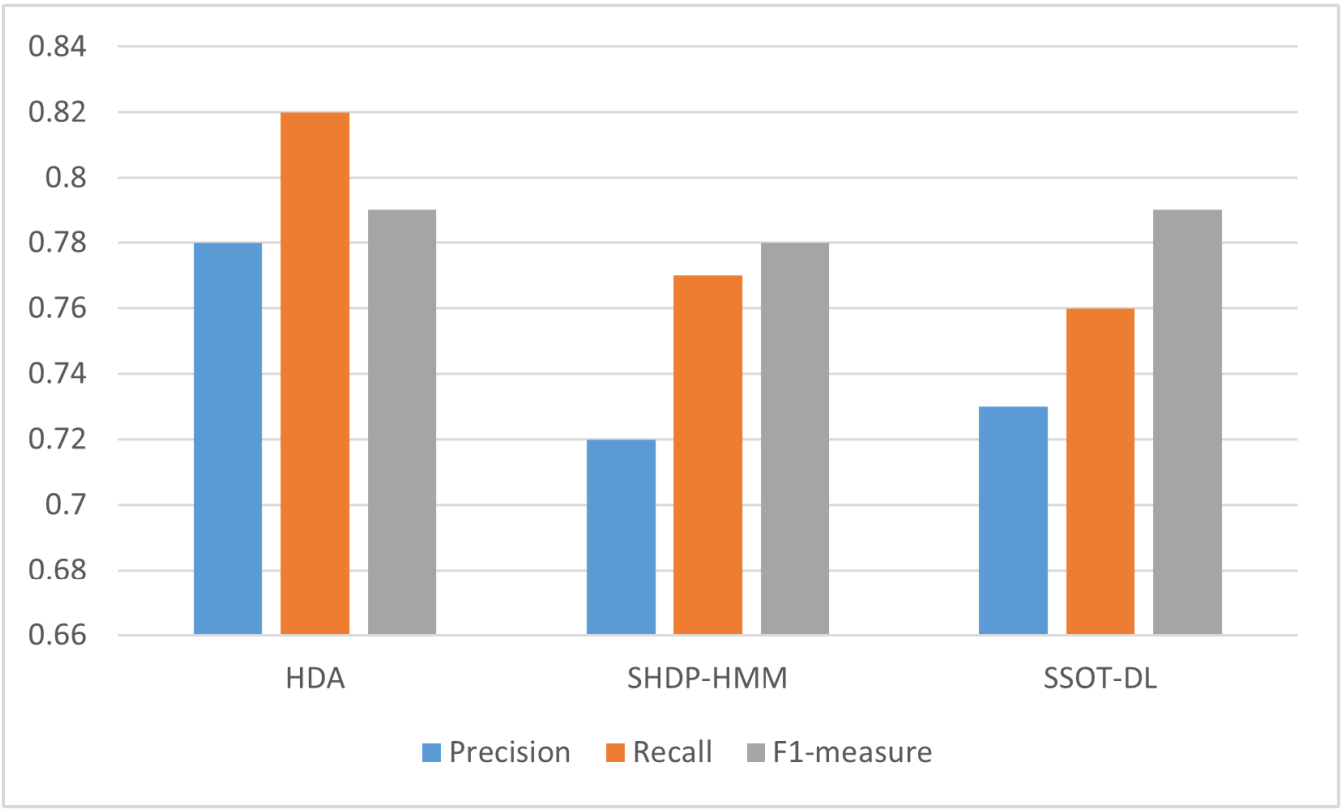

FIGURE 6. Comparison of Precision, Recall, and F1-measure between HDA and baseline.

C. EVALUATION MEASUREMENTS

To evaluate the performance of the model, this section employs widely adopted evaluation metrics, including Precision, Recall, F1-measure. Precision is calculated by

$$P = \frac{TP}{TP + FP} \tag{17}$$

Recall is calculated by

$$R = \frac{TP}{TP + FN} \tag{18}$$

F1-measure is calculated by

$$F1 = \frac{2 \times P \times R}{P + R} \tag{19}$$

TP denotes to the number of positive samples correctly predicted as positive by the model, while TN denotes to the number of negative samples correctly predicted as negative. FP denotes the number of negative samples incorrectly predicted as positive by the model, and FN denotes the number of positive samples incorrectly predicted as negative.

D. EXPERIMENTAL SETUPS

This experiment is conducted on the Topic Detection and Tracking (TDT2) corpus [38], which comprises 2,280 English broadcast news programs, containing a total of 11,406 stories. Each story has an average of 20 topics and 200 words. We divided the corpus into a training set of 2,040 programs and a test set of 240 programs. After applying the Porter Stemmer and removing meaningless stop words, we obtained a vocabulary of 49,527 words. HDA is an unsupervised method for story segmentation that assumes the number of labels and topics in the text is unknown. Therefore, we utilized sen2vec [39] to generalize sentence-level text representations without prior information. We used a blocking sampler to infer the parameters of HDA and extracted a 30-dimensional sentence vector to improve the sampling speed. Subsequently, we evaluated the model's performance using various metrics and compared the results of our model with those of the baseline in this chapter.

E. EXPERIMENTAL RESULTS

The parameters α_0 and K represent the dispersion of the base distribution and the self-transition probability of the states, respectively. The value of α_0 determines the probability of generating new clusters, which means that a larger α_0 tends to produce more topics. The parameter K helps to model the duration of the topics more accurately. To investigate the effect of α_0 and K on segmentation performance, we adjusted them as tuning parameters. We set $\gamma = 1$ and the degree of freedom parameter $\nu = 56$. *Fig.*3 shows the segmentation results and topic numbers for different values of α_0, with K set to 1. The x-axis represents the value of α_0, while the y-axes on the left and right sides of the figure represent the number of topics and the precision, recall, and F1-measure,

respectively. When α_0 is set from 0.001 to 100, the precision, recall, and F1-measure fluctuate, while the number of topics fluctuates within a small range of 147-164. When $\alpha_0 = 1$ and the number of topics is 164, we obtain the best segmentation result with a precision of 0.78, recall of 0.82, and F1 score of 0.79. *Fig*.4 illustrates the impact of different values of K on the segmentation results, precision, recall, and number of topics. From these two figures, we can observe that the number of topics varies slightly around the actual number of topics of 160 under different values of α_0 and K, indicating that the influence of α_0 and K on the segmentation results is minimal. When $\alpha_0 = 1$ and $K = 1$, we obtain the highest F1 score of 0.79, with a precision of 0.78 and a recall of 0.82.

Comparing HDA with the baseline models as shown in *Fig*.6, the results demonstrate that HDA surpasses the baselines in precision, recall, and F1-measure metrics. These findings highlight HDA's superior accuracy, stability, efficiency, and reliability in text segmentation tasks. The results underscore the effectiveness of HDA in handling complex text.

VII. CONCLUSION AND FUTURE WORK

This paper proposes the HDA method to investigate text segmentation, which achieves automatic recognition and segmentation of different paragraphs in text sequences through hierarchical analysis and attention weighting. Compared to existing methods, this model exhibits higher accuracy and efficiency, and better supports text analysis, information extraction, and other tasks. The experimental results demonstrate its excellent performance in text segmentation tasks.

Although the proposed text segmentation method based on the HDA has shown promising performance, there are still aspects that can be further researched and improved. For instance, experiments on larger datasets could be conducted to verify the model's generalization ability and scalability. Additionally, exploring how to apply this method to more text analysis tasks, such as text classification and sentiment analysis, is also worth investigating. In future work, we could also attempt to combine this method with other natural language processing techniques to enhance the model's performance and application scope.

REFERENCES

[1] H. An, D. Wang, Z. Pan, M. Chen, and X. Wang, "Text segmentation of health examination item based on character statistics and information measurement," *CAAI Trans. Intell. Technol.*, vol. 3, no. 1, pp. 28–32, Mar. 2018, doi: 10.1049/trit.2018.0005.

[2] P. Hu, W. Wang, Q. Li, and T. Wang, "Touching text line segmentation combined local baseline and connected component for uchen Tibetan historical documents," *Inf. Process. Manage.*, vol. 58, no. 6, Nov. 2021, Art. no. 102689, doi: 10.1016/j.ipm.2021.102689.

[3] B. Geng, "Text segmentation for patent claim simplification via bidirectional long-short term memory and conditional random field," *Comput. Intell.*, vol. 38, no. 1, pp. 205–215, Feb. 2022, doi: 10.1111/coin.12455.

[4] P. S. Patheja and N. Tripathi, "Segmentation free text recognition for overlapping characters using spectral features and bidirectional recurrent wavelet neural network," *Int. J. Intell. Eng. Informat.*, vol. 10, no. 6, pp. 464–483, 2022, doi: 10.1504/ijiei.2022.10053817.

[5] A. K. Sarkar and Z.-H. Tan, "Self-segmentation of pass-phrase utterances for deep feature learning in text-dependent speaker verification," *Comput. Speech Lang.*, vol. 70, Nov. 2021, Art. no. 101229, doi: 10.1016/j.csl.2021.101229.

[6] X. Yan, X. Xiong, X. Cheng, Y. Huang, H. Zhu, and F. Hu, "HMM-BIMM: Hidden Markov model-based word segmentation via improved bi-directional maximal matching algorithm," *Comput. Electr. Eng.*, vol. 94, Sep. 2021, Art. no. 107354.

[7] S. Spala, N. A. Miller, Y. Yang, F. Dernoncourt, and C. Dockhorn, "DEFT: A corpus for definition extraction in free- and semi-structured text," in *Proc. 13th Linguistic Annotation Workshop*, 2019, pp. 124–131.

[8] M. Liao, G. Pang, J. Huang, T. Hassner, and X. Bai, "Mask TextSpotter v3: Segmentation proposal network for robust scene text spotting," in *Computer Vision—ECCV 2020*. Glasgow, U.K., Cham, Switzerland: Springer, Aug. 2020, pp. 706–722.

[9] Q.-V. Dang and G.-S. Lee, "Scene text segmentation via multi-task cascade transformer with paired data synthesis," *IEEE Access*, vol. 11, pp. 67791–67805, 2023.

[10] M. Zhang, X. Li, S. Yue, and L. Yang, "An empirical study of TextRank for keyword extraction," *IEEE Access*, vol. 8, pp. 178849–178858, 2020.

[11] Y. Jiang, C. Zhong, and B. Zhang, "AGD-linknet: A road semantic segmentation model for high resolution remote sensing images integrating attention mechanism, gated decoding block and dilated convolution," *IEEE Access*, vol. 11, pp. 22585–22595, 2023, doi: 10.1109/ACCESS.2023.3253289.

[12] X. Lan, H. Zhai, and Y. Wang, "A novel DOA estimation of closely spaced sources using attention mechanism with conformal arrays," *IEEE Access*, vol. 11, pp. 44010–44018, 2023, doi: 10.1109/ACCESS.2023.3272617.

[13] D. Feng, J. Xie, T. Liu, L. Xu, J. Guo, S. G. Hassan, and S. Liu, "Fry counting models based on attention mechanism and YOLOv4-tiny," *IEEE Access*, vol. 10, pp. 132363–132375, 2022, doi: 10.1109/ACCESS.2022.3230909.

[14] G. Chang, S. Hu, and H. Huang, "Two-channel hierarchical attention mechanism model for short text classification," *J. Supercomput.*, vol. 79, no. 6, pp. 6991–7013, Apr. 2023, doi: 10.1007/s11227-022-04950-1.

[15] I. Alagha, "Leveraging knowledge-based features with multilevel attention mechanisms for short Arabic text classification," *IEEE Access*, vol. 10, pp. 51908–51921, 2022, doi: 10.1109/ACCESS.2022.3175306.

[16] Z. Xiao, Z. Nie, C. Song, and A. T. Chronopoulos, "An extended attention mechanism for scene text recognition," *Exp. Syst. Appl.*, vol. 203, Oct. 2022, Art. no. 117377, doi: 10.1016/j.eswa.2022.117377.

[17] Y. Feng and Y. Cheng, "Short text sentiment analysis based on multi-channel CNN with multi-head attention mechanism," *IEEE Access*, vol. 9, pp. 19854–19863, 2021, doi: 10.1109/ACCESS.2021.3054521.

[18] K. Jeong, M. Chae, and Y. Kim, "Online learning for the Dirichlet process mixture model via weakly conjugate approximation," *Comput. Statist. Data Anal.*, vol. 179, Mar. 2023, Art. no. 107626, doi: 10.1016/j.csda.2022.107626.

[19] J. Zhang and A. Dassios, "Truncated Poisson–Dirichlet approximation for Dirichlet process hierarchical models," *Statist. Comput.*, vol. 33, no. 1, p. 30, Feb. 2023, doi: 10.1007/s11222-022-10201-3.

[20] D. E. Troncoso Romero, M. G. Cruz Jiménez, and U. Meyer-Baese, "Hardware-efficient decimation with spectral shape approximating the nth power of a Dirichlet kernel," *Circuits, Syst., Signal Process.*, vol. 41, no. 9, pp. 4886–4905, Sep. 2022, doi: 10.1007/s00034-022-02009-3.

[21] J. Paisley, C. Wang, D. M. Blei, and M. I. Jordan, "Nested hierarchical Dirichlet processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 256–270, Feb. 2015, doi: 10.1109/TPAMI.2014.2318728.

[22] B. Can and S. Manandhar, "Tree structured Dirichlet processes for hierarchical morphological segmentation," *Comput. Linguistics*, vol. 44, no. 2, pp. 349–374, 2018, doi: 10.1162/coli_a_00318.

[23] A. M. Dai and A. J. Storkey, "The supervised hierarchical Dirichlet process," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 243–255, Feb. 2015, doi: 10.1109/TPAMI.2014.2315802.

[24] E. Karakoca, G. K. Kurt, and A. Görçin, "Hierarchical Dirichlet process based gamma mixture modeling for terahertz band wireless communication channels," *IEEE Access*, vol. 10, pp. 84635–84647, 2022, doi: 10.1109/ACCESS.2022.3197603.

[25] M. Ötting and D. Karlis, "Football tracking data: A copula-based hidden Markov model for classification of tactics in football," *Ann. Operations Res.*, vol. 325, no. 1, pp. 167–183, Jun. 2023, doi: 10.1007/s10479-022-04660-0.

[26] P. A. Chavan and S. Desai, "Effective epileptic seizure detection by classifying focal and non-focal EEG signals using human learning optimization-based hidden Markov model," *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104682, doi: 10.1016/j.bspc.2023.104682.
[27] M. Chassan and D. Concordet, "How to test the missing data mechanism in a hidden Markov model," *Comput. Statist. Data Anal.*, vol. 182, Jun. 2023, Art. no. 107723, doi: 10.1016/j.csda.2023.107723.
[28] Z. Xing, Y. Qiao, Y. Zhao, and W. Liu, "Dynamic texture classification based on bag-of-models with mixture of student's t-hidden Markov models," *Comput. Vis. Image Understand.*, vol. 230, Apr. 2023, Art. no. 103653, doi: 10.1016/j.cviu.2023.103653.
[29] F. Dama and C. Sinoquet, "Partially hidden Markov chain multivariate linear autoregressive model: Inference and forecasting—Application to machine health prognostics," *Mach. Learn.*, vol. 112, no. 1, pp. 45–97, Jan. 2023, doi: 10.1007/s10994-022-06209-5.
[30] I. Ameer, N. Bölücü, G. Sidorov, and B. Can, "Emotion classification in texts over graph neural networks: Semantic representation is better than syntactic," *IEEE Access*, vol. 11, pp. 56921–56934, 2023, doi: 10.1109/ACCESS.2023.3281544.
[31] S.-H. Kim, H. Nam, and Y.-H. Park, "Analysis-based optimization of temporal dynamic convolutional neural network for text-independent speaker verification," *IEEE Access*, vol. 11, pp. 60646–60659, 2023, doi: 10.1109/ACCESS.2023.3286034.
[32] Y. Yang, R. Tang, M. Xia, and C. Zhang, "A texture integrated deep neural network for semantic segmentation of urban meshes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4670–4684, 2023, doi: 10.1109/JSTARS.2023.3276977.
[33] Z. Li, X. Yang, L. Zhou, H. Jia, and W. Li, "Text matching in insurance question-answering community based on an integrated BiLSTM-TextCNN model fusing multi-feature," *Entropy*, vol. 25, no. 4, p. 639, Apr. 2023, doi: 10.3390/e25040639.
[34] V. S. Rathod, A. Tiwari, and O. G. Kakde, "Wading corvus optimization based text generation using deep CNN and BiLSTM classifiers," *Biomed. Signal Process. Control*, vol. 78, Sep. 2022, Art. no. 103969, doi: 10.1016/j.bspc.2022.103969.
[35] V. Ramu Reddy and K. Sreenivasa Rao, "Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks," *Neurocomputing*, vol. 171, pp. 1323–1334, Jan. 2016, doi: 10.1016/j.neucom.2015.07.053.
[36] J. Yu and H. Shao, "Broadcast news story segmentation using sticky hierarchical Dirichlet process," *Int. J. Speech Technol.*, vol. 52, no. 11, pp. 12788–12800, Sep. 2022, doi: 10.1007/s10489-021-03098-4.
[37] T. Lattisi, D. Farina, and M. Ronchetti, "Semantic segmentation of text using deep learning," *Comput. Informat.*, vol. 41, no. 1, pp. 78–97, 2022, doi: 10.31577/cai_2022_1_78.
[38] J. Fiscus, G. R. Doddington, J. S. Garofolo, and A. F. Martin, "Nist's 1998 topic detection and tracking evaluation (TDT2)," in *Proc. 6th Eur. Conf. Speech Commun. Technol. (EUROSPEECH)*, Budapest, Hungary, Sep. 1999, pp. 247–250, doi: 10.21437/Eurospeech.1999-65.
[39] T. K. Saha, S. R. Joty, and M. A. Hasan, "CON-S2V: A generic framework for incorporating extra-sentential context into Sen2Vec," in *Machine Learning and Knowledge Discovery in Databases* (Lecture Notes in Computer Science), vol. 10534, M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Dzeroski, Eds. Skopje, Macedonia, Cham, Switzerland: Springer, Sep. 2017, pp. 753–769, doi: 10.1007/978-3-319-71249-9_45.

YANHUA WANG received the Ph.D. degree in linguistics and computer intelligence applications from the School of Literature, Lanzhou University. She is possessing 15 years of experience in news media and publicity, specializing in broadcasting and journalism. She is currently the Deputy Director of the Editorial Department for "Student World." She contributed to the National Social Science Fund major bidding project on "Digital Preservation of Chinese Oral Culture." She has participated in the 2022 online workshop on "Language Contact and the Evolution of Chinese Language in Northern China" organized by the National Institute for Oriental Languages and Civilizations, France. She has presented at the 2022 International Conference on "Cutting-Edge Issues in Linguistics and Chinese Education" held at Lanzhou University. She also attended the Sixth Doctoral Forum on Linguistics at Tsinghua University, in 2022.

CHUNFANG MIN received the Ph.D. degree. She is currently a Professor with the School of Literature, Lanzhou University, and the Director of the Institute of Chinese Language and Literature. She is specializing in the study of Chinese language and literature and its intelligent applications. She received the Northwest Minzu University Young Faculty Development Award, in 2002, 2004, and 2006, the 13th Youth Faculty Development Award of Gansu Province, and the Social Sciences Award of Gansu Province, in 2006. In 2013, she was selected for the Ministry of Education's "New Century Excellent Talents Support Program."