IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 20, 2020, accepted October 1, 2020, date of publication October 19, 2020, date of current version October 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3032173

A Topic Learning Pipeline for Curating Brain Cognitive Researches

YING SHENG[1], JIANHUI CHEN[2], XIAOBO HE[1], ZHE XU[1], JIANGFAN GAO[1], AND SHAOFU LIN[3]

[1]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2]Beijing Key Laboratory of MRI and Brain Informatics, Beijing University of Technology, Beijing 100124, China
[3]Beijing Institute of Smart City, Beijing University of Technology, Beijing 100124, China

Corresponding author: Shaofu Lin (linshaofu@bjut.edu.cn)

This work was supported in part by the Science and Technology Project of Beijing Municipal Commission of Education under Grant KM201710005026, in part by the National Basic Research Program of China under Grant 2014CB744600, and in part by the Beijing Natural Science Foundation under Grant 4182005.

ABSTRACT Cognition is the most basic but complex process of human beings. Benefit from noninvasive neuroimaging technologies, a series of important brain projects have been carried out to model cognition from different aspects and levels. Because modeling such a complex phenomenon requires characterizations of numerous entities and cannot only depend on the efforts of one or more laboratories within a project cycle, a lot of neuroimaging text mining researches have focused on curating neuroimaging-based brain cognitive raw data, derived data and result data, to collect multi-aspect information about brain cognitive researches for comprehensively and objectively characterizing key entities of brain cognition. However, the data-centric perspective leads to the shortcomings of poor topic semantics and topic independent results. This paper proposes a brand-new perspective of big data sharing in neuroimaging, that is, curating brain cognitive researches. A new task definition of neuroimaging text mining and a topic learning pipeline integrating the heterogeneous deep neural networks and density clustering of topic relations are designed to realize this new perspective. The experimental results on actual data sets show that the proposed method can obtain more accurate and complete research topics for effectively characterizing brain cognition and its researches.

INDEX TERMS Neuroimaging text mining, topic learning, data curation, brain cognition.

I. INTRODUCTION

Cognition is the most basic but complex process of human beings and also an important research topic in psychology, neuroscience and artificial intelligence [1]. Various neuroimaging technologies, especially functional magnetic resonance imaging (fMRI) [2], can non-invasively detect the connection between cognitive states and patterns of brain activity, and have received widespread attentions. A series of important brain projects, such as Human Connectome Project [3], Adolescent Brain and Cognitive Development (ABCD) [4], and UK Biobank [5], have been launched to harness brain big data for modeling cognitive functions of the brain. From fMRIDC [6], LONI-IDA [7], OpenfMRI [8] to OpenNeuro [9], various open neuroimaging data centers or platforms have also been developed to support such a data-driven approach. However, creating a detailed map of mechanism of brain cognition needs to decode various cognitive states from complex patterns of brain activity, and is involved with characterization of numerous entities, such as brain regions, brain networks. Although the structural and functional aspects of these entities can partly be evaluated in the laboratory, the complexity and the cross-scale nature of the modelled phenomena prohibit a comprehensive evaluation of all aspects at play [10]. Hence, brain cognitive data must be complemented with metadata, which record the whole scientific work flow, to support the systematic evaluation and integration of research results cross laboratories or even disciplines for obtaining comprehensive and objective characterizations of key entities of brain cognition.

A lot of efforts have been made to generate metadata for curating neuroimaging-based brain cognitive raw data, derived data (e.g, unthresholded statistical map), and result data (e.g, peak coordinates of brain activation), as well as brain cognitive researches behind the data. fMRIDC [6] provides manual descriptions of neuroimaging raw data.

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

BrainMap [11] proposes a manual approach to describe peak coordinates of brain activation and the corresponding key research information, such as experimental tasks, subjects. NeuroVault [12] supports the manual inputting of neuroimaging research information for sharing unthresholded statistical maps of the human brain. Brainspell [13] provides a distributed collaboration platform to collect neuroimaging research information by a crowdsourcing approach. Provenances [14] are the core research issues of neuroimaging metadata. The representative work is Neuroimaging Data Model (NIDM), which provides metadata schemas, including NIDM-Experiment, NIDM-Workflow [15], NIDM-Results [16], to describe neuroimaging-based brain cognitive experiments, data analytical processes, and data analytical results respectively. At present, NIDM has been regarded as an effective approach to realize the findable, accessible, interoperable and reusable (FAIR) neuroimaging data sharing [17]. The commonly used neuroimaging data analytical tools, such as SPM and FSL, and some important neuroimaging data platforms, such as NeuroVault, have provided the support for NIDM.

The above researches have brought enormous influences. For example, the statistical data on the Neurosynth platform shows that 14371 Neurosynth-based literatures had been published by July 2018. Related researches are involved with cognitive functions decoding of the whole brain [18], [19] or certain brain regions [20].

However, existing researches on curating neuroimaging-based brain cognitive data mainly adopt manually generation or tool exporting to collect information related to brain cognitive researches, and lack a quick and steady acquisition approach. The open access literature provides a huge and steadily growing information sources for brain cognitive researches. However, existing neuroimaging text mining technologies cannot meet the requirements:

- Though various neuroimaging text mining technologies [10], [21]–[25] have been developed to realize automatic collection of research knowledge from neuroimaging literature, they mainly focused on curating derived data and result data, and led to the problem of poor topic semantics. Related researches mainly recognized cognitive functions, disorders and activated brain regions by using frequency-based or probabilistic model-based methods. The obtained topics often lacked enough semantics. For example, 15 entitity categories recognized by Abacha *et al.* [24] focused on brain regions (e.g., Gross brain anatomy,Functional neuroanatomy), function (e.g., Brain function, Body function), Medical problem, Stimuli/responses (e.g., Gustation, Visual) and cannot effectively cover the whole process of cognitive researches. Furthermore, too many domain independent general words (e.g. using, repeat, asked) were included in topics [24] and the relations among topic words were also ambiguous. Such topics with poor semantics cannot effectively characterize the whole research process for supporting systematic evaluation and integration of research results cross laboratories or even disciplines.
- Some recent researches on neuroimaging text mining [10], [24], [25] transformed the open-domain topic learning task with unknown types of topics to the close-domain named entity recognition task with predefined types of topics for enriching extracted results and removing general words. These technologies partly solved the shortcoming of poor topic semantics, but the shortcoming of topic independent results became apparent. Those recognized entities in literature may not describe the authors' study. They are generalized domain knowledge but cannot characterize the authors' study for supporting systematic evaluation and integration of research results, as well as other complex applications, such as brain mechanism meta-analysis [26], research path recommendation [27].

A. CONTRIBUTION

Based on the above observations, this paper proposes a topic learning pipeline to extract key information related to neuroimaging-based brain cognitive researches from full texts of open access literature. The main contributions may be summarized as follows:

- Firstly, this paper adopts a brand-new perspective to share neuroimaging big data, i.e., curating brain cognitive researches. The existing neuroimaging data centers or platforms were developed to assist the data-driven modeling of brain cognition. Therefore, all neuroimaging metadata or provenances were constructed to curate neuroimaging-based brain cognitive data, including raw data, derived data and result data. However, although the neuroimaging research has entered the era of big data [28], it is still an arduous work to aggregate cross-laboratory experimental data because of multiple reasons, including economics, data ownership, privacy protection etc. In order to solve this problem, this paper adopts a brand-new perspective of data sharing, which curates brain cognitive researches rather than only data by extracting key information about research processes and results from literature. A large number of brain cognitive research instances can be aggregated quickly for systematically evaluating and modeling of brain cognition.
- Secondly, this paper designs a new task definition of neuroimaging text mining following the new perspective of curating brain cognitive researches. Full texts of open access neuroimaging literature are selected as the information source to obtain the complete descriptions of researches. The rapidly growing open access neuroimaging literature ensures the feasibility of the selection. Based on this information source, a brain informatics (BI) provenance model is defined as the objective of neuroimaging text mining giving consideration to both importance and availability of information. This model includes nine kinds of domain entities

and forty-five kinds of relations for characterizing the whole process of neuroimaging-based brain cognitive researches.

- Thirdly, this paper proposes a heterogeneous topic learning pipeline, called CNN-BiLSTM-CNN-DC, to extract key research information according to the new task definition of neuroimaging text mining. Initially, learning from the topic learning incorporated with domain knowledge, this pipeline adopts a learning framework driven by both knowledge and data. During the data-driven model learning process, the BI provenance model and its domain term dictionaries are used as prior knowledge for the objective definition of model learning and the construction of train datasets. This not only improves the automation level of proposed methods but also effectively solves the shortcoming of poor topic semantics by transforming the open-domain topic learning task to close-domain one. Furthermore, this pipeline adopts a new topic learning model which integrates the heterogeneous deep neural network and density clustering to realize an effective combination of close-domain and open-domain topic learning. Such a combination follows the research trend of close-domain topic learning based on deep neural networks and solve the shortcoming of poor topic semantics. It also solves the shortcoming of topic independent results by using the open-domain topic clustering. Finally, the topic clustering is based on the relational density between topics, which can better represent the topics of literature than word frequency adopted by various probabilistic topic models.

B. STRUCTURE

The rest of this paper is organized as follows. Section II summarizes previous work related to neuroimaging text mining and topic learning. Section III describes the model and its learning framework of the proposed pipeline CNN-BiLSTM-CNN-DC. Experiments are presented in Section IV to validate the effectiveness of the proposed method. Finally, Section V gives concluding remarks.

II. RELATED LITERATURE

A. NEUROIMAGING TEXT MINING

Neuroimaging texts, especially scientific literature, are growing fast. Taking only fMRI as an example, 482 relevant literatures have been published in the journal PLoS One in 2018. They are valuable knowledge resources for studying human intelligence, pathological mechanism of brain and mental diseases, brain-computer interface, and so on.

Neuroimaging text mining is to extract information or knowledge from neuroimaging texts and has been paid much attention. In order to decode a wide range of cognitive states, Neurosynth [29] recognized terms based on frequency and used the Naive Bayesian Classifier to predict the occurrence of specific terms based on the whole-brain activation patterns. Poldrack *et al.* [21] adopted the Latent Dirichlet Allocation (LDA) method to identify topics of literature in the neurosynth database and mapped these topics to brain activation data for discovering mechanisms of brain functions and psychological processes. French *et al.* [22] developed a co-occurrence-based method to extract brain regions and their relations from neuroscience literature. Alhazmi *et al.* [23] extracted topic words based on frequency and constructed relations between semantic spaces of topics and brain activated regions by using correspondence analysis and hierarchical clustering.

As mentioned in the previous section, the above researches have the shortcoming of poor topic semantics. Because of only focusing on the relations between research topics and activated brain regions, they mainly adopted the frequency-based or probabilistic model-based topic learning methods, and the obtained topics lacked enough semantics. Firstly, too many general words were included in topics. Topic words in Neurosynth include many general words (e.g. using, repeat, asked) and domain general words (e.g. magnetic resonance, brains) [24]. Poldrack *et al.* had to remove general words by referring to concepts in the Cognitive Atlas [28]. In addition, the relations between topic words were also ambiguous. Alhazmi *et al.* had to adopt hierarchical clustering to form an implicit relation among topic words [23]. Such topics with poor semantics can be used to construct the mapping between topics and activated brain regions by some furthering processing operations, but are far from effectively characterizing the whole research process, i.e., curating brain cognitive researches for systematically evaluating and modeling of brain cognition.

In recent years, a few researches of neuroimaing text mining began to expend the extraction perspective from the relations between research topics and activated brain regions to the whole research process. Multi-aspect research information extraction during the research process has gained more attentions. Ben Abacha *et al.* [24] adopted the rule-based method and conditional random field (CRF) based methods to recognize 15 functional neuroimaging entity categories, including gross brain anatomy, functional neuroanatomy, medical problem, stimuli/responses etc. Shardlow *et al.* [10] recognized various entities, including brain regions, experimental values, neuron types, etc., by using active and deep learning, for curating information in computational neuroscience. Riedel *et al.* [25] completed a comprehensive evaluation on recognizing various entities related to the cognitive experiment, which were defined by the Cognitive Paradigm Ontology [30] and involved with behavioral domain, paradigm class, diagnosis, context, instruction, stimulus modality, stimulus type, etc., based on multiple corpus features (abstract-only and full-text) and classifiers (Bernoulli naïve Bayes, k-nearest neighbors, logistic regression, and support vector classifier).

However, these neuroimaing text mining researches oriented to the research process only focused on the experimental process and neglected some important entities during the data analytical process, such as analytical tools and

methods, analytical results, which are also key information for systematically evaluating and modeling of brain cognition. Their extraction tasks were also limited to entity recognition. Though the use of domain knowledge, such as the Cognitive Paradigm Ontology, in the task definition solved the shortcoming of general words, the shortcoming of poor topics semantics was still existing because the relations among entities weren't recognized. Furthermore, the shortcoming of topic independent results became apparent because those entities appeared in literature might not be used to describe authors' study. For example, a cognitive state stated in the ''Related Work'' chapter may only be a control region rather than the result of current study. It is possible to cause a deviation using these entities to characterize the authors' study for systematically evaluating and modeling of brain cognition.

B. TOPIC LEARNING

Extracting research information from neuroimaing literature can be regarded as topic learning, which learns meaningful expressions of texts from document sets [31]. It is a basic work in text semantic analysis [32] and text mining [33].

The most classic topic learning methods are various probabilistic topic models, in which the most widely used one is the Latent Dirichlet allocation (LDA) model proposed by Blei *et al.* [34]. It detects the global semantic topic structure by mining the co-occurrence pattern of words, and gives topics of each document in the form of probability distribution. However, LDA uses the bag of words to convert text information into digital information. This kind of method ignores the word order and textual structure and cannot effectively model documents [35]. In order to solve this problem, various improved LDA models have been developed. Balikas *et al.* [36] proposed the sentence LDA (senLDA) model, which introduced the information of textual structures and word dependence into topic modeling for achieving the higher topic granularity. Nguyen *et al.* [37] proposed the Latent Feature Topic Modeling (LFTM) which integrated quantitative contextual information to extend the traditional LDA and DMM models by using word embedding and achieved remarkable results in the topic consistency evaluation. Li *et al.* [38] proposed a generative topic embedding model, which combined the traditional probabilistic topic model with word embedding and mined word collocation patterns from both the global document and the local context for generating coherent topics.

As stated in the previous section, traditional probabilistic topic models, such as LDA, often produce poorly semantic and incomprehensible topics. In order to overcome this shortcoming, integrating domain knowledge into topic modeling has become an important optimization direction of topic modeling [39]. By using domain knowledge to guide the modeling process, knowledge-based topic modeling can obtain the more coherent and meaningful topics. Yao *et al.* [40] combined LDA with the large-scale probabilistic knowledge base to improve semantic consistency and accuracy of topics. Amplayo *et al.* [41] proposed the MicroASM model which introduced the external seed dictionary into topic modeling and obtained rich semantic topics based on the seed-topic word pairs.

In recent years, topic learning based on deep neural network has become another important optimization direction of topic modeling. By using the deep neural network to model the context, a topic model with deep semantic representation can be constructed to overcome various shortcomings of traditional probabilistic topic models, such as poor model scalability, poor topic semantic coherence, insufficient feature expression ability, which are caused by the shallow feature structure and the probabilistic generation mode [42]. Dieng *et al.* [43] proposed the TopicRNN model based on Recurrent Neural Network (RNN), which can capture remote semantic dependency between potential topics to improve the ability of generating reasonable topics. Zhang *et al.* [44] proposed the topic-enhanced LSTM model TE-LSTM+SC, which captured contextual features of textual sequences by using LSTM and obtained potential semantic topics as diverse as possible based on the topic modeling layer and the similarity constraint (SC) strategy. Yang *et al.* [45] proposed a deep learning-based topic learning method which used the candidate topics obtained by LDA to construct the feature inputs of deep neural network for obtaining the more accurate topics.

As stated above, the probabilistic topic models are still the most important topic learning methods. Various knowledge-based or deep-learning-based improved methods are still based on word frequency and only optimize topics by the domain knowledge base [40], term dictionary [40] or deep neural network [45]. However, aiming at full texts of literature, the word frequency often cannot represent the research topics. For example, ''bold'' and ''signal'' are the most frequently occurring words in the literature ''Decoding Vigilance with NIRS'' [46]. Obviously, they are just two general domain terms rather than topics of this literature. Therefore, the shortcoming of poor topic semantics cannot be effectively solved. This has also been proved by previous researches of neuroimaging text mining [21], [29]. In fact, we find that the relational density among topics can represent the topics of literature better because researchers often describe the topics in detail from different aspects.

Some recent topic learning researches [43], [44] transformed topic learning from an open-domain task with unknown types of topics to a close-domain task with predefined types of topics. Topic extraction mainly depends on word features and contextual features obtained by word embedding and RNN. These methods still lead to the shortcoming of topic independent results.

In summary, mining full texts of neuroimaging literature, the existing topic learning methods still cannot effectively solve the two shortcomings of neuroimaging text mining, i.e., poor topic semantics and topic independent results. It is necessary to develop a new topic learning method for curating brain cognitive researches.

TABLE 1. Entity types of brain cognitive.

Entity Type	Definition	Example
Gross Brain Anatomy	The gross brain anatomy is an anatomical region of the cerebral cortex and used to mark the occurrence location of brain response in the brain cognitive research.	superior frontal gyrus
Cognitive Function	The cognitive function is an ability of human brain to process information and used to denote the brain function implied by brain response in the brain cognitive research.	memory retrieval
Subject	The subject is a participant in the brain cognitive research and recorded for behavioral or brain physiological data.	children, adult
Medical Problem	The medical problem is an abnormal process of life activity and used to denote the subject's disease or abnormal symptom in the brain cognitive research.	tinnitus, diabetes
Sensory Stimuli or Response	The sensory stimuli or response is used to denote the sensory channel of stimulus presentation in the brain cognitive research.	olfactory, visual
Experimental Task	The experimental task is a task (e.g., questions, games, etc.) which is performed by subjects in the brain cognitive research.	color-word stroop task
Experimental Measurement	The experimental measurement is a kind of brain testing equipment used in the brain cognitive research.	functional magnetic resonance imaging
Analytical Tool and Method	The analytical tool and method is a mining algorithm or tool which is used to analyze experimental data in the brain cognitive research.	independent component analysis
Brain Network	The brain network is a kind of brain response which is mined from experimental data in the brain cognitive research.	default mode network

III. METHODOLOGY

In this section, we will introduce the proposed method, including the BI provenance model and the topic learning pipeline CNN-BiLSTM-CNN-DC.

A. BRAIN INFORMATION PROVENANCE MODEL FOR CURATING BRAIN COGNITIVE RESEARCHES

Brain cognitive researchers are interested in brain responses in specific experimental tasks, including locations of responses and brain functions implied by these responses [47]. Therefore, experimental tasks, brain regions and brain functions are three types of important entities for curating brain cognitive researches. Furthermore, as stated in [48], task-based data sets are generally more unique than structural or resting-state fMRI data sets because each study typically employs a different task manipulation to examine a specific psychological process. Hence, the entities related to the cognitive experiment, including sensory stimuli or responses, subject and experimental measurement, are also the important entities for curating brain cognitive researches. Besides those experimental entities, analytical entities, including analytical tools and methods, and analytical results, are very important for evaluating the reliability of study results.

Following the above fundamental observations, 9 types of domain entities are defined in Table 1. Ignoring the direction of the relations, 45 types of relations between entities are described in Table 2. These entities and relations form a BI provenance model [49], which can effectively characterize the whole process of brain cognitive researches, including the experimental process and the analytical process, for systematically evaluating and modeling of brain cognition. Fig. 1 shows the BI provenance model.

B. TOPIC LEARNING PIPELINE: CNN-BILSTM-CNN-DC

1) OVERVIEW

In order to mine the full texts of neuroimaging literature according to the BI provenance model, this paper proposes a topic learning pipeline, called CNN-BiLSTM-CNN-DC. It includes a multi-layer heterogeneous topic learning model and a knowledge-data fusion driven, asynchronous learning framework. Fig. 2 shows the overall architecture of topic learning model, which consists of three layers: candidate topic recognition layer, topic relation extraction layer and topic filter layer. In the following sections, we will describe each layer in detail.

2) CANDIDATE TOPIC RECOGNITION LAYER

The candidate topic recognition layer is used to identify 9 types of domain entities in the BI provenance model as candidate topics from the inputted literature texts. As shown in Fig. 2, it includes three sub layers: text vectorization, topic feature modeling and candidate topic annotation.

a: TEXT VECTORIZATION LAYER

The text vectorization layer encodes sentences of literature and constructs textual vectors as the input of the next layer.

First, three types of vectors are constructed based on lexical units, upper and lowercase features and domain terminology dictionaries in the sentence.

- Word vector

The word vector is the vectorization of words. Based on the statistical information of global co-occurrences and local contexts, word vectors are learned to contain as much semantic and grammatical information as possible [50]. This study adopts the Glove word vector model which was trained on 6 billion words of Wikipedia and web texts [51].

TABLE 2. Relation types of brain cognitive.

Relation Type	Type ID	Definition
is-part-of	BRI-BRI	"is-part-of" is a relation between two "Gross Brain Anatomy" entities and denotes the inclusion relation between them.
produce-the-activation-in	COG-BRI	"produce-the-activations-in" is the relation between the entities "Cognitive Function" and "Gross Brain Anatomy" and denotes that "Cognitive Function" produces the activation in "Gross Brain Anatomy".
has-the-medical-problem-of	SUB-MDI	"has-the-medical-problem-of" is the relation between the entities "Subject" and "Medical Problem" and denotes that "Subject" suffers from "Medical Problem".
performs	SUB-TSK	"performs" is the relation between the entities "Subject" and "Experimental Task" and denotes that "Subject" completes "Experimental Task" in the brain cognitive research.
is-involved-with	TSK-SEN	"is-involved-with" is the relation between the entities "Experimental Task" and "Sensory Stimuli or Response" and denotes that the subject receives and performs "Experimental Task" by "Sensory Stimuli or Response" in the brain cognitive research.
uses	TSK-MES	"uses" is the relation between the entities "Experimental Task" and "Experimental Measurement" and denotes that the brain cognitive research uses "Experimental Measurement" to collect brain data related to "Experimental Task".
reflects	TSK-COG	"reflects" is the relation between the entities "Experimental Task" and "Cognitive Function" and denotes that "Experimental Task" is used to assess "Cognitive Function" in the brain cognitive research.
is-included-in	BRI-RLT	"is-included-in" is the relation between the entities "Gross Brain Anatomy" and "Brain Network" and denotes that "Gross Brain Anatomy" is the part of "Brain Network".
is-mined-by	TOL-RLT	"is-mined-by" is the relation between the entities "Brain Network" and "Analytical Tool and Method" and denotes that "Brain Network" is recognized by using "Analytical Tool and Method".
……	……	……

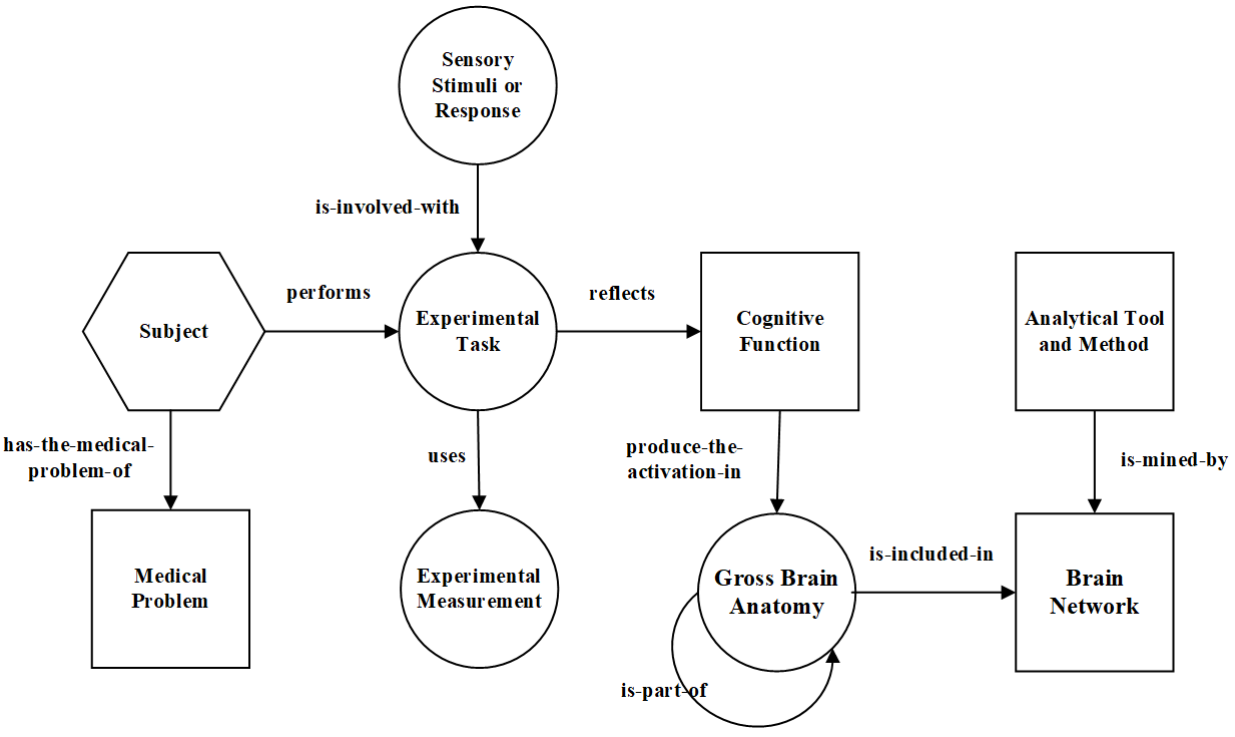

FIGURE 1. The BI provenance model with part of relations.

- Case feature vector

The 9 types of domain entities and their contexts are involved with a large number of domain terms, which often have capitalized abbreviated names, such as fMRI, ICA (Independent Component Analysis), DMN (Default Mode Network). Therefore, identifying the upper- and lower-case features is important for domain entity recognition. Because the case features of words have been eliminated in the word vector training and data set preprocessing, this study constructs a one-hot case feature vector for each word.

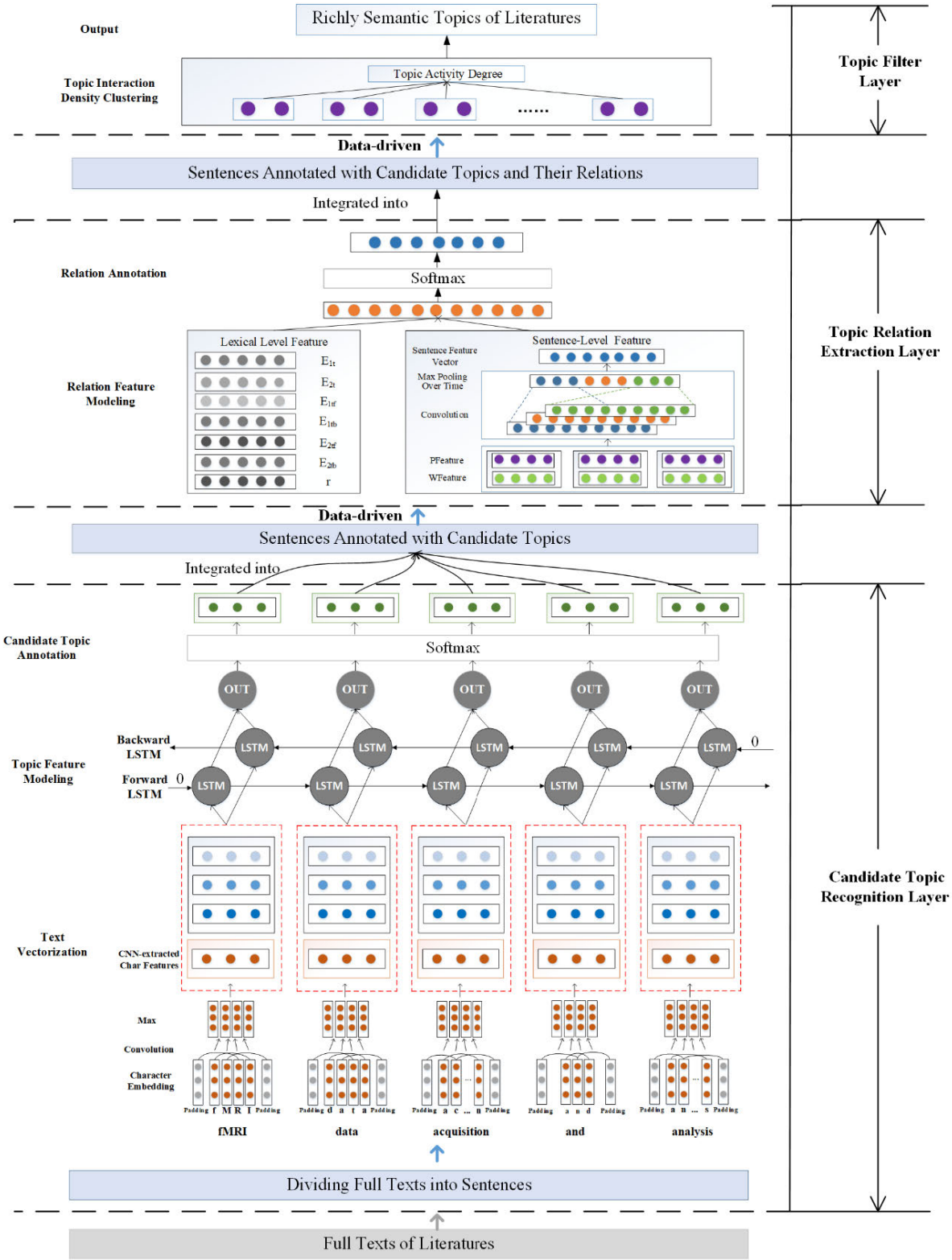

FIGURE 2. The proposed multi-layer heterogeneous topic learning model.

It includes six dimensions: ''numeric'', ''allLower'', ''allUpper'', ''initialUpper'', ''mainly_numeric'', ''contains_digit'' and ''other''.

- Terminology dictionary vector

The term dictionary encoding plays an important role in multi-class named entity recognition [52]. This study

constructs 9 term dictionaries for the 9 domain entity types defined in Table 1. Based on these term dictionaries, the one-dimensional terminology dictionary vector can be constructed. When a word matches any term in the dictionaries, a label index is set as the dimensional value of terminology dictionary vector. For example, if the word ''fMRI'' matches the term in the domain dictionary ''Experimental Measurement'', a label ''B-MES'' can be obtained. The ''B'' means that the word ''fMRI'' matches the beginning of term and the ''MES'' means that the term belongs to the domain dictionary ''Experimental Measurement''. The index of label ''B-MES'' in the label set is just the dimensional value of terminology dictionary vector of ''fMRI''.

Then, for any word, a combined word vector can be constructed by using the above three types of vectors [53]:

$$v_{word} = [\mathrm{v}_w, \mathrm{v}_c, \mathrm{v}_t] \tag{1}$$

where v_w, v_c,v_t are the corresponding word vector, case feature vector and terminology dictionary vector.

Secondly, the CNN is used to construct the character feature vector v_{char} based on the character embedding. The character embedding is a 25 dimensional vector and obtained by lookup the random table which is initialized with values drawn from an uniform distribution with range $[-0.5, 0.5]$. As shown in Fig. 2, the input character set of CNN includes all characters in the sentence and the special token PADDING. All PADDINGs are added on both sides of words according to the size of CNN window.

b: TOPIC FEATURE MODELING LAYER

The topic feature modeling layer adopts the BiLSTM to model contextual information based on the above text vectorization. For a sentence $S = [\mathrm{w}_1, \mathrm{w}_2, \ldots \mathrm{w}_n]$, the whole process can be described as follows:

$$f_j = [v_{wordj}, v_{charj}], \mathrm{j} \in [1,n] \tag{2}$$

$$\vec{h}_j = \mathrm{LS}\vec{T}M(f_j), \mathrm{j} \in [1,n] \tag{3}$$

$$\overleftarrow{h}_j = \mathrm{LS}\overleftarrow{T}M(f_j), \mathrm{j} \in [1,n] \tag{4}$$

$$h_j = [\vec{h}_j, \overleftarrow{h}_j] \tag{5}$$

where f_j is the combined vector of the w_j, $\vec{h}_j$ is the output of the forward hidden layer, $\overleftarrow{h}_j$ is the output of backward hidden layer, and h_j is the output of the hidden layer of BiLSTM.

c: CANDIDATE TOPIC ANNOTATION

The candidate topic annotation layer decodes the output of the topic feature modeling layer into log-probabilities of entity types by using the log-softmax [54], and then annotates recognized candidate topics and their types in literature.

3) TOPIC RELATION EXTRACTION LAYER

The topic relation extraction layer is used to identify the relations between candidate topics from annotated literature texts. As shown in Fig. 2, it includes two sub layers: relation feature modeling and relation annotation.

a: RELATION FEATURE MODELING LAYER

The relation feature modeling layer encodes the lexical level and sentence-level features of topic relations to construct the relation feature vector as the input of the next layer.

- Lexical Level Feature.

Word embedding, annotated topics, contexts of topics and relation types are integrated to construct the lexical level feature vector. For example, there is a sentence ''Intrinsic functional connectivity alterations of the [primary visual cortex] in primary [glaucoma] patients'', which contains two candidate topics ''primary visual cortex'' and ''glaucoma'' and a relation type ''BRI-MDI''. Its lexical level feature vector can be constructed as follows

$$V_{lf} = [\mathrm{E}_{1t}, \mathrm{E}_{2t}, \mathrm{E}_{1tf}, \mathrm{E}_{1tb}, \mathrm{E}_{2tf}, \mathrm{E}_{2tb}, \mathrm{r}] \tag{6}$$

where E_{1t} is the word vector of the first topic ''primary visual cortex'', E_{2t} is the word vector of the second topic ''glaucoma'', E_{1tf} is the word vector of the previous word ''the'' of the first topic, E_{1tb} is the word vector of the latter word ''in'' of the first topic, E_{2tf} is the word vector of the previous word ''primary'' of the second topic, E_{2tb} is the word vector of the latter word ''patients'' of the second topic, and r is the index of the relation type ''BRI-MDI'' in the topic relation table.

- Sentence-Level Feature

Firstly, the word representation is constructed for each word in the sentence. It is used to describe the local features of topic relations and can be defined as follows:

$$Y = [\mathrm{WFeature}, \mathrm{PFeature}] \tag{7}$$

where WFeature is the word vector of the current word and PFeature is the distance vector between the current word and two topics. For example, aiming at the ''connectivity'' in the sentence ''Intrinsic functional connectivity alterations of the [primary visual cortex] in primary [glaucoma] patients'' stated above, WFeature is its word vector and PFeature= [-4,-9] is the distances between it and the two topics.

Secondly, the sentence-level global features are obtained by CNN [55]. The word representation only reflects the local features of relations. Based on the word representation, CNN is used to fuse all local features for predicting topic relations from the global perspective. After convolution, the most useful features of each convolution kernel are extracted by max pooling [56]. The process can be described as follows:

$$m = \max(W_1 Y) \tag{8}$$

where W_1 is the linear transformation matrix of hidden layer, Y is the word representation, and m is the optimal feature corresponding to each convolution kernel. Then, the sentence-level feature can be calculated as follow:

$$V_{sf} = \tanh(W_2 \mathrm{m}) \tag{9}$$

where W_2 is the linear transformation matrix of hidden layer, and tanh is the activation function.

After extracting the lexical level feature and sentence-level feature, they are directly linked into a relation feature vector $v = [V_{lf}, V_{sf}]$.

b: RELATION ANNOTATION

The relation annotation layer adopts the softmax classifier to decode the output of the relation feature modeling layer into the probability of topic relations, and then annotates the recognized topic relations.

4) TOPIC FILTER LAYER

The topic filter layer uses density clustering to identify research topics of literature and their relations from candidate topics. The process is described in algorithm 1.

Firstly, a relation density clustering of topics is performed by using DBSCAN [57]. It classifies topics into different clusters based on the weighted density of relations among topics, as shown in Step 16. The weighted density means that the density of relations is calculated based on not only the number of relations but also the proportion in the same type of relations for classifying really relevant topics into a clustering. Therefore, two important parameters of DBSCAN, the relation radius *rRadius* and the relation density threshold *rdThreshold*, are calculated as follows:

- In Step 1-13, the number of relations between any two topics t_{i} and t_{j} is normalized based on the corresponding relation type to obtain the relational distance $rdis_{i,j}$, and then *rRadius* is set as the average value of all relational distances in Step 14.
- *rdThreshold* is calculated in Step 15. For each topic t_{i}, its $tNum_i$ is set as the number of relations whose relational distances are bigger than rRadius, and then *rdThreshold* is set as the average value of all $tNum_i$.

Secondly, research topics are recognized from the obtained topic clusters, as shown in Step 17-23. Considering both richness and relevance of topics, a focus score of topic clusters can be calculated in Step 20. num^i_{typ} is the number of topic types in the cluster $\mathrm{clu_i}$ and num^i_{rel} is the number of relations in the cluster $\mathrm{clu_i}$. Select the cluster with the highest score as the research focus area and all topics in this cluster are the research topics of the literature l.

C. MODEL LEARNING PROCESS GUIDED BY DOMAIN KNOWLEDGE

Learning from the topic learning integrating domain knowledge [40], [41], this study introduces BI provenances and related domain ontologies or term dictionaries into the data-driven model learning process. As shown in Fig. 3, it includes two stages, model learning of candidate topic recognition and model learning of topic relation extraction.

1) MODEL LEARNING OF CANDIDATE TOPIC RECOGNITION

During model learning of candidate topic recognition, BI provenances are regarded as domain knowledge for the task definition and the training data set construction:

Algorithm 1 Topic Identification Based on Relational Density Clustering

Input:

the sentence set $S = \{s_1, s_2, \ldots, s_n\}$ of literature l in which each s_i is a sentence annotated with candidate topics and their relations;

the entity type set of brain cognitive $ET = \{E_1, E_2, \ldots, E_9\}$ in which each E_i is a entity type in Table 1.

Initialize:

a topic set $T = \{t_1, t_2, \ldots, t_n\}$ which includes all candidate topics in l;

a relational distance matrix of topics RDIS;

a distance matrix of relation types RT;

the relation radius rRadius = 0;

the relation density threshold rdThreshold = 0;

Output:

the research topic resTopics of literature l

1. **for** each element $rdis_{i,j}$ in RDIS
2. **if**(i ≤ j)
3. $rdis_{i,j}$ = the number of relations between the candidate topics t_i and t_j in l;
4. **else** $rdis_{i,j} = 0$;
5. **end**
6. **for** each element $rt_{i,j}$ in RT
7. **if**(i ≤ j)
8. $rt_{i,j}$ = the number of relations belonging to the relation type $r_{i,j}$ which is the relation type between E_i and E_j;
9. **else** $r_{i,j} = 0$;
10. **end**
11. **for** each element $rdis_{i,j}$ in RDIS
12. $rdis_{i,j} = rdis_{i,j}/rt_{m,n}$, where the topic t_i belongs to the Entity type E_m and the topic t_j belongs to the Entity type E_n;
13. **end**
14. $\mathrm{rRadius} = \frac{\sum_{\mathrm{RDIS}} rdis_{i,j}}{|rdis_{i,j} \neq 0|}$;
15. $\mathrm{rdThreshold} = \frac{\sum_{\mathrm{T}} tNum_i}{|rdis_{i,j} \neq 0|}$, where $tNum_i$ is the number of $rdis_{i,j} or rdis_{j,i} > rRadius$, j = 1 . . . n;
16. the cluster set $\mathrm{CLUS} = \{\mathrm{clu_1}, \mathrm{clu_2}, \ldots, \mathrm{clu_m}\}$ = DBSCAN(rRadius, rdThreshold), where $\mathrm{clu_i}$ is a cluster of relations;
17. **for** each element $\mathrm{clu_i}$ in CLUS
18. $\mathrm{num^i_{typ}}$ = the number of entity types of topics in $\mathrm{clu_i}$;
19. $\mathrm{num^i_{rel}}$ = the number of relations in $\mathrm{clu_i}$;
20. $\mathrm{score_i} = \frac{\mathrm{num^i_{typ}}}{\sum_{i=1}^{m} \mathrm{num^i_{typ}}} + \frac{\mathrm{num^i_{rel}}}{\sum_{i=1}^{m} \mathrm{num^i_{rel}}}$
21. **end**
22. **if** ($\mathrm{score_i}$ ==max(score))
23. resTopics = all topics in $\mathrm{clu_i}$
24. **return** resTopics;

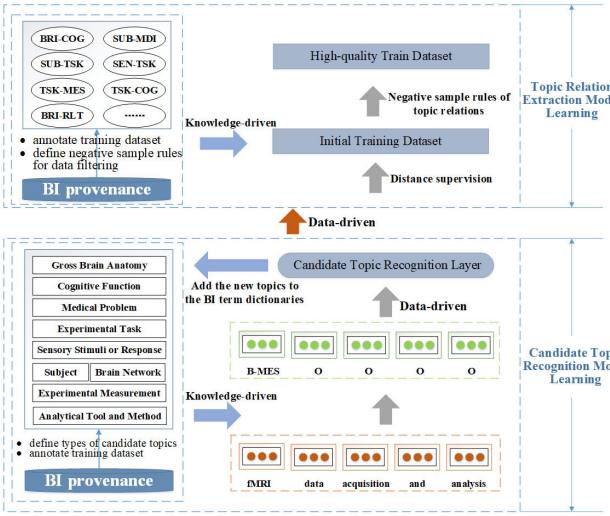

FIGURE 3. The model learning process guided by BI provenances.

- Firstly, the 9 entity types of brain cognitive in the BI provenance model are used to define the types of candidate topics. Based on this definition, the open-domain task in which topic types are unknown beforehand is transformed to a close-domain task with predefined topic types, for overcoming the shortcomings of poor semantics or meaningless topics in traditional probabilistic topic models.
- Secondly, initial term dictionaries in BI provenances are used to annotate domain corpora for automatically constructing the training data set of candidate topic recognition.

Brain cognitive researches are involved with multiple frontier fields and most of entity types, such as ''Cognitive Function'', ''Medical Problem'', ''Experimental Task'', are constantly updated. Therefore, those initial term dictionaries are incomplete. This study adopts the CNN+BiLSTM deep neural network to recognize different types of candidate topics based on multiple lexical, contextual and domain features. New entities, which never appear in the training data set, can also be recognized. (see the experiments in Section IV for details). These entities can be added into BI term dictionaries for enriching BI provenances and expanding the training data set.

2) MODEL LEARNING OF TOPIC RELATION EXTRACTION

During model learning of topic relation extraction, BI provenances are also regarded as domain knowledge for the task definition and the training data set construction. The relation types of brain cognitive in the BI provenance model are used to define the relation types of topics. Based on these definitions, the training data set of topic relation extraction can be automatically constructed by integrating the distance supervision and rule-based negative sample filtering. The whole process includes the following two steps:

TABLE 3. Negative sample rules of topic relations.

Rule Type	ID	Rule Definition
Distance Rule	1	The distance between two topic words is greater or equal to 30 words.
	2	Two topic words are in different brackets or only one topic word is in the bracket and distance from another topic word is greater or equal to 20 words.
Syntactic Rule	3	There are negative words, such as "rather than", ", whereas", ", while", ", but", "does not", between two topic words.
	4	There is one or more semicolons between two topic words.
	5	In the case of more than three equals in a sentence, there are both equals and commas between two topic words; In the case of more than three colons in a sentence, there are both colons and commas between two topic words.

- Firstly, a large number of initial training data are automatically constructed by aligning BI provenances with the input of the candidate topic recognition layer [58].
- Secondly, a group of negative sample rules are defined based on syntactic and distance features. By using these rules, the initial training data can be filtered to obtain a high-quality train data set.

The analysis of corpora reveals that it is difficult to recognize positive samples, i.e., the sentences including one or more topic relations, by using rules. However, it is easy to recognize negative samples, which don't include any topic relation, based on rules. Hence, this study adopts negative sample rules to filter the initial training data of topic relation extraction. All of rules are given in Table 3.

D. BASELINE METHODS

LDA, MicroASM and LFTM were selected as baseline methods to validate the effectiveness of the proposed method. In order to demonstrate the value of the proposed method on brain cognitive researches more directly, experimental results were also compared with data on Neurosynth, which is one of the most influential data and knowledge sharing platform in the field of functional neuroimaging research.

1) BASELINE METHODS FOR TOPIC LEARNING

a: THE LDA-BASED METHODS

LDA is one of the most popular topic learning models. Hench, this study chooses it as a baseline method. LDA was performed on both full-texts (LDA-FullText) and abstracts (LDA-Abstract) to learn literature topics respectively.

b: THE MICROASM-BASED METHOD

The MicroASM introduces external domain knowledge into topic modeling by using a seed term dictionary, for avoiding meaningless topics. In this study, the domain dictionary ''Cognitive Function'' was used to construct the seed-topic

word pairs, such as ''perception-tinnitus''. Based on these seed-topic word pairs, MicroASM was performed on full texts of literature to obtain research topics.

c: THE LFTM-BASED METHODS

LFTM extends traditional probabilistic topic models by integrating quantitative contextual information. This study adopted both LF-LDA and LF-DMN as baseline methods. They were performed on full texts of literature to obtain research topics.

d: THE NEUROSYNTH-BASED METHOD

Neurosynth is one of the most widely used platforms for neuroimaging data and knowledge sharing. In recent years, many researchers have used the Neurosynth platform to perform the meta-analysis [59] for the mapping between brain functions and brain response patterns. In order to intuitively prove the validity of the proposed method on brain cognitive researches, this study designed a Neurosynth-based baseline method:

- Firstly, aiming at a specific brain function, two groups of literatures were chosen based on the topic words of the Neurosynth platform and the proposed method, respectively.
- Secondly, two meta-analyses of brain mechanisms were performed by using the peak coordinates of activated regions from these two groups of literatures.
- Finally, the results of meta-analysis were compared to evaluate the validity of topics between the proposed method and the Neurosynth platform.

2) BASELINE METHODS FOR CANDIDATE TOPIC RECOGNITION

Candidate topic recognition is the first step of proposed method. Its accuracy is the basis of accurate topic recognition.

Therefore, this study adopted the latest CRFs-based functional neuroimaging named entity recognition methods [24] as baseline methods of candidate topic recognition. They include the CRFs method (NER-FuncNeuro-CRFs) and the semantic features-based CRFs method (NER-FuncNeuro-CRFs+Semantic Features). The semantic features include the following two types:

- part-of-speech features, i.e., part of speech (POS) information of topic words.
- semantic features, i.e., identifying information whether the word belongs to one of three types of ''Medical problems'', ''Human Body Anatomy'' and ''Gross Brain Anatomy''.

IV. EXPERIMENTS AND EVALUATION

A. EXPERIMENTAL SETTINGS

1) EXPERIMENTAL DATA AND TERMINOLOGY DICTIONARY

a: EXPERIMENTAL DATA

The experimental data set is composed of 677 full-text literatures from the journal PLoS One. They were published in the past five years and contain any one of ''fMRI'', ''functional magnetic resonance imaging'' or ''functional MRI'' in abstracts.

b: TERMINOLOGY DICTIONARY

BI provenances contain 9 types of term dictionaries corresponding to the 9 entity types in Table 1:

- The ''Gross Brain Anatomy'' dictionary was obtained from the Whole Brain Atlas[1] and consists of 519 terms about Brodmann area, AAL (Anatomical Automatic Labeling) area and brain anatomical structures.
- The ''Cognitive Function'' dictionary consists of 839 terms obtained from Cognitive Atlas.[2]
- The ''Subject'' dictionary consists of 100 terms involved with age-group terms (adolescents, middle age, etc.), gender terms (men, women), and occupation terms (college student, teacher, etc.).
- The ''Medical Problem'' dictionary consists of 602 disease terms obtained from Wikipedia (list of diseases[3] and cancer types[4]) and Health on the Net (rare diseases[5])
- The ''Sensory Stimuli or Response'' dictionary consists of 53 sensory perception terms, such as ''Gustation'', ''Visual'', ''Emotional'', ''Olfactory'', ''Auditory'', and ''Somatosensory''.
- The ''Experimental Task'' dictionary consists of all 763 ''Tasks'' terms on Cognitive Atlas.[6]
- The ''Experimental Measurement'' dictionary consists of 21 neuroimaging detection device terms, such as MRI (magnetic resonance imaging), PET (Positron Emission Computed Tomography).
- The ''Analytical Tool and Method'' dictionary consists of 140 terms about data mining algorithms and tools obtained from Baidu Encyclopedia.[7]
- The ''Brain Network'' dictionary consists of 31 terms of brain networks.[8]

Fig. 4 gives the distribution of above 3068 terms.

The three types of terms for the semantic features-based CRFs method were obtained as follows:

- The ''Medical problems'' terms were obtained from Wikipedia (list of diseases, cancer types) and Health on the Net (rare diseases), as mentioned earlier;
- The ''Human Body Anatomy'' terms were obtained from human body vocabulary.[9]
- The ''Gross Brain Anatomy'' terms were obtained from the Whole Brain Atlas and Neuroanatomy.[10]

[1] http://www.med.harvard.edu/aanlib/
[2] http://www.cognitiveatlas.org/
[3] https://en.wikipedia.org/wiki/Lists of diseases/
[4] https://en.wikipedia.org/wiki/List of cancer types/
[5] http://www.hon.ch/HONselect/RareDiseases/index.html/
[6] http://www.cognitiveatlas.org/tasks/a/
[7] https://baike.baidu.com/
[8] http://www.mamicode.com/info-detail-2224948.html
[9] http://www.enchantedlearning.com/wordlist/body.shtml/
[10] https://brainaacn.org/neuroanatomy/

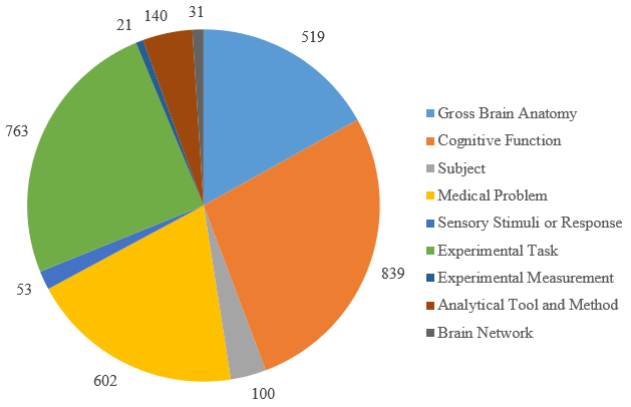

FIGURE 4. The term distribution.

TABLE 4. Examples of topic annotation.

Annotation Example 1					
visual	representation	of	the	stroop	task
B-COG	I-COG	O	O	B-TSK	I-TSK
Annotation Example 2					
DMN	hyperconnectivity		in	fatigue	has
B-RLT	O		O	B-COG	O
been	demonstrated	in	breast	cancer	survivor
O	O	O	B-MDI	I-MDI	O

TABLE 5. An example of relation annotation.

Annotation Example	
120 5 6 10 11	the contribution of the lateral prefrontal cortex to control of memory retrieval has widely been acknowledged.

2) ANNOTATION SCHEME

a: TOPIC ANNOTATION

In this study, the "BIO" tagging system was used to annotate whether a word in the sentence was part of a topic word [60]. The labeling scheme consists of "-" and the topic type abbreviation. The "Cognitive Function" type is abbreviated as "COG", the "Experimental Task" is abbreviated as "TSK", the "Brain Network" is abbreviated as "RLT" and the "Medical Problem" is abbreviated as "MDI". Two examples of topic annotation are shown in Table 4.

b: TOPIC RELATION ANNOTATION

The topic relation annotation consists of five parts: relation type, the beginning position of the first topic word, the ending position of the first topic word, the beginning position of the second topic word, and the ending position of the second topic word. An example is shown in Table 5.

In Table 5, "120" represents that the sentence is a positive relation sample between the first and second topic types. "5" and "6" respectively represent the beginning and ending positions of the first topic word "prefrontal cortex" in the sentence. "10" and "11" respectively represent the beginning and ending positions of the second topic word "memory retrieval".

3) PARAMETER SETTINGS

For the candidate topic recognition layer, the dimension size of the word vector was set at 100. The epochs was set at 50, the convolution width at 3, the CNN output size at 30, the dimensional number of LSTM hidden layer at 200, the mini-batch size at 9, and the dropout at 0.5. The Adam algorithm was used to optimize parameters.

For the topic relation extraction layer, the dimension size of the word vector was set at 50, the convolution width at 3, the dimensions of the two hidden layers at 200 and 100 respectively, and the learning rate at 0.01.

For the topic filter layer, the relation radius and the relation density threshold were set dynamically based on the distribution of topics in literature, as mentioned earlier.

4) EXPERIMENTAL EVALUATION

This study adopted the precision rate, the recall rate and F_1 values [61] to evaluate experimental results:

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{10}$$

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{11}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

The key problem is how to determine that the recognized topics are correct, i.e., the "gold standard" [62]. At present, most of studies adopted manual evaluation [63] or rule-based evaluation [64], whose results were subjective and difficult to reproduce. Therefore, this study set an objective criterion based on the literature labels given by PLoS One journal. The experimental results are considered correct when the recognized topic words are exactly the same as literature labels.

B. EXPERIMENTAL RESULTS

1) COMPARISONS OF CANDIDATE TOPIC RECOGNITION

The 10-fold cross-validation method was adopted in this study. The data set was divided into the training set and the test set according to the ratio of 9:1 [65], and the average precision rate, recall rate and F_1 value were finally obtained.

The topic type "Experimental Task" and "Medical Problem" have the lowest precision. After analyzing the experimental results, we found that 52 new terms, such as "object-location memory task", "object-scene task", "t-detection task", were regarded as false-positive recognized topics of "Experimental Task" because they were not included in the terminology dictionary. However, by querying

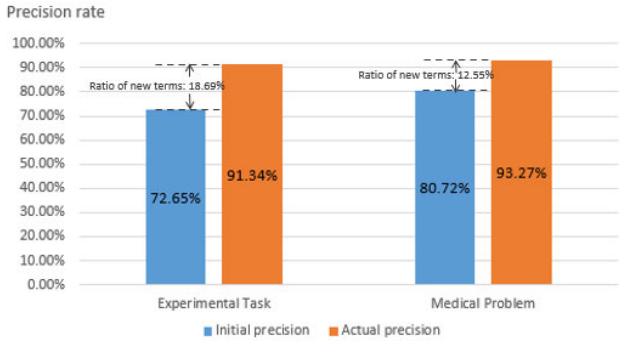

FIGURE 5. Experimental results of candidate topic recognition in "Experimental Task" and "Medical Problem".

TABLE 6. Experimental results of candidate topic recognition.

Entity Categories	Mean Precision	Mean Recall	Mean F1
Gross Brain Anatomy	96.05%	96.97%	96.52
Cognitive Function	89.89%	93.43%	91.59
Subject	97.87%	98.71%	98.26
Medical Problem	93.27%	88.18%	90.65
Sensory Stimuli or Response	96.74%	99.11%	97.9
Experimental Task	91.34%	82.75%	86.83
Experimental Measurement	99.18%	99.90%	99.55
Analytical Tools and Methods	93.39%	88.19%	90.5
Total	94.50%	95.56%	95.03

on Google, we can find these terms actually belong to the entity type "Experimental Task". The Google query result shows that "The Object-Location Memory task assesses cognition, specifically spatial memory and discrimination, in rodent models of CNS disorders", and "The Object-scene task refers to the task of subjects' recognition judgment and confidence judgment of the old / new scene". Therefore, they should belong to the topic type of "Experimental Task". For the topic type "Medical Problem", 31 recognized new terms, such as "deafness", "hyperglycemia", "eclampsia", "blindness", have the similar situations. Adding these new terms, the final precision is shown in Fig.5.

In addition, false-positive recognized topics include some abstract concepts which have not specific meanings. For examples, "task" was recognized as topics of "Experimental Task" and "cancer" was recognized as topics of "Medical Problem". There are 29 false-positive abstract concepts in "Experimental Task" and 9 false-positive abstract concepts in "Medical Problem". Though these abstract concepts aren't included in the terminology dictionary, they still belong to "Experimental Task" or "Medical Problem" obviously and should not be regarded as false-positive recognized topics. Considering these new terms and abstract concepts, the actual experimental results are shown in Table 6.

Two CRFs-based methods were used as baseline methods. Fig. 6 gives the comparison between the candidate topic recognition layer of proposed model, which is based on CNN-BiLSTM, and baseline methods.

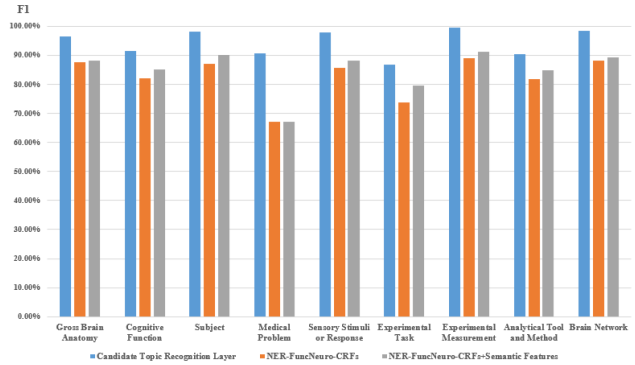

FIGURE 6. The comparison of experimental results on candidate topic recognition.

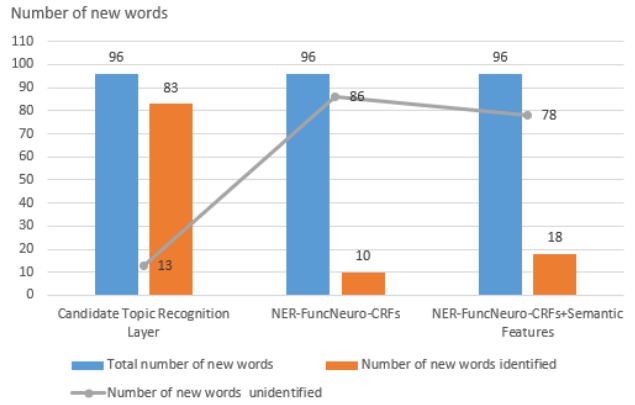

FIGURE 7. The comparison of experimental results on recognizing new words.

The results show that the candidate topic recognition layer achieves better $F1$ value than baseline methods. Furthermore, recognizing new words is an important and bottleneck problem because of rapidly developing brain cognitive researches. The candidate topic recognition layer also achieves better results than two baseline methods. As shown in Fig. 7, the total number of new words is 96, and the number of new words identified by candidate topic recognition layer is 83, which is much higher than the two baseline methods of NER-FuncNeuro-CRFs and NER-FuncNeuro-CRFs+Semantic Features.

2) COMPARISONS OF TOPIC LEARNING

In this paper, 677 literatures were divided into the training set and the test set according to the ratio of 9: 1. Using the evaluation criterion based on PLoS One, the number of matching words between recognized topics and literature labels of PLoS One was counted. The experimental results are shown in Fig. 8.

When the threshold of correct topic recognition was set to 3, that is, the recognized topics of literature were regarded as correct if the number of matching words between recognized topics and literature labels of PLoS One was greater

TABLE 7. The comparison of topic learning.

Literature Title	Method	Recognized Topic	PLoS One Literature Label
A Model for Visual Memory Encoding	CNN-BiLSTM-CNN-DC	visual memory, fmri, default mode network, independent component analysis, principal component analysis, encoding task, visual cortex, auditory, visual, encoding, adult, memory, cerebellum, vigilance, sulcus, attention	vision, memory, functional magnetic resonance imaging, attention, cognition, recall (memory), emotions, gamete intra-fallopian transfer
	MicroASM	task, memory, visual, network, working, components, cross-modal, information	
	LF-LDA	memory, encoding, using, processed, analyses, studies, multiple, different	
	LF-DMN	encoding, process, directionality, ica, fmri, components, disease, mode	
	LDA-FullText	visual, perceptual, processing, particular, switching, region, study, notion	
	LDA-Abstract	brain, memory, involved, processing, music, process, activity, region	
Fibromyalgia Patients Had Normal Distraction Related Pain Inhibition but Cognitive Impairment Reflected in Caudate Nucleus and Hippocampus during the Stroop Color Word Test	CNN-BiLSTM-CNN-DC	functional magnetic resonance imaging, caudate nucleus, perception, fibromyalgia, pag, pain, hippocampus, nucleus, distraction, hyperactivity disorder, attention deficit	pain sensation, fibromyalgia, functional magnetic resonance imaging, cognitive impairment, caudate nucleus, cognition, hippocampus, analgesia
	MicroASM	patients, pain, motor, fm, pd, group, scwt, stress	
	LF-LDA	fm, patients, rts, scwt, years, cognitive, task, mean	
	LF-DMN	patients, fm, activation, study, cognitive, reduced, task, areas	
	LDA-FullText	patient, ad, impairment, control, healthy, study, cognitive, dysfunction	
	LDA-Abstract	patient, fmri, brain, functional, connectivity, study, control, left	
Decreased Fronto-Temporal Interaction during Fixation after Memory Retrieval	CNN-BiLSTM-CNN-DC	recency judgment task, default mode network, prefrontal cortex, functional mri, reaction time, memory, retrieval, hippocampus	memory, prefrontal cortex, hippocampus, functional magnetic resonance imaging, reaction time, neuroimaging, signal filtering, thumbs
	MicroASM	training, activation, task, brain, equation, trials, regions, dual-task	
	LF-LDA	memory, fixation, trials, periods, mri, relative, order, consisted	
	LF-DMN	functional, temporal, present, lateral, study, mri, relative, order	
	LDA-FullText	trial, button, participant, screen, fixation, presented, cross, press"	
	LDA-Abstract	cortex, prefrontal, anterior, dorsal, activity, cingulate, revealed, region	

than or equal to 3, the CNN-BiLSTM-CNN-DC method proposed in this study can correctly recognize the topics of 56 literatures from 67 test literatures. This number is much higher than that of MicroASM (30 literatures), LF-LDA (36 literatures), LF-DMN (27 literatures), LDA-Abstract (19 literatures) and LDA-FullText (8 literatures). When the threshold of correct topic recognition was set to 2, the proposed CNN-BiLSTM-CNN-DC correctly recognized the topics of 64 literatures from 67 test literatures. This result is still better than MicroASM (51 literatures), LF-LDA (49 literatures), LF-DMN (47 literatures), LDA-Abstract (53 literatures) and LDA-FullText (40 literatures).

Table 7 lists the comparison of topic learning. The topics recognized by LDA-Abstract and LDA-FullText include many irrelevant or poorly semantic words, such as ''processing'' and ''involved''. MicroASM, LF-LDA and LF-DMN can recognize literature topics more accurately, but these topics mainly focus on experimental tasks or analytical results

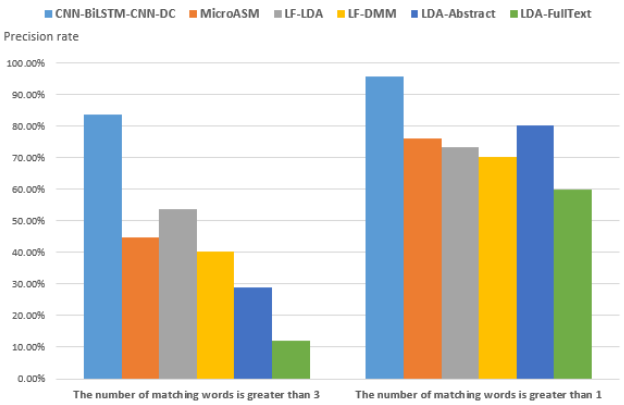

FIGURE 8. The comparison of experimental results on topic learning.

and cannot effectively characterize the whole process of brain cognitive researches. The CNN-BiLSTM-CNN-DC proposed by this study can obtain more accuracy, rich-semantic, and

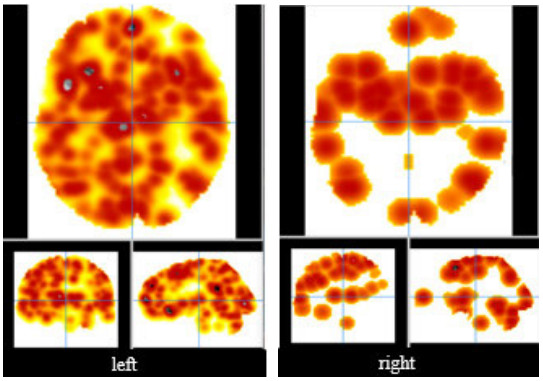

FIGURE 9. The activated brain regions obtained by two meta-analyses of calculation mechanisms (left: Neurosynth-based, right: CNN-BiLSTM-CNN-DC-based).

complete topics for effectively representing the research process of brain cognitive.

As stated above, the Neurosynth-based baseline method was adopted to intuitively prove the validity of the proposed method on brain cognitive researches. This study chose ''calculation'' as the objective cognitive function and performed the meta-analysis of calculation-related brain mechanism for evaluating the results of topic learning.

For fairly comparing two meta-analyses, the overlapping 295 functional neuroimaging articles in the PLoS One data set and the Neurosynth data set were selected as the test data set. From these articles, the following three rules were defined to filter calculation-related articles based on topics from the proposed method or the Neurosynth platform:

a) The rule based on a single word: If the recognized topics of ''Experimental Task'' or ''Cognitive Function'' include ''mental arithmetic'', ''mental calculation'', ''arithmetic'', ''calculation'', ''mathematic'', ''addition'', ''subtraction'', ''multiplication'', ''division', or ''number processing'', the articles are calculation-related.

b) The rule based on multiple words: If the recognized topics of ''Experimental Task'' or ''Cognitive Function'' include ''arithmetic fact + retrieval'' or ''numeric + magnitude'', the articles are calculation-related. The ''A + B'' means that the topics include both ''A'' and ''B''.

c) The rule based on excluding words: If the recognized topics of ''Experimental Task'' or ''Cognitive Function'' include ''cognitive subtractions'', ''cognitive subtraction'', ''numerical stroop'', ''numerical luminance'', or ''additional'', the articles are calculation-unrelated.

Based on the above rules, 35 calculation-related articles were chosen based on topics of the Neurosynth platform and 4 articles were chosen based on the topics of the proposed CNN-BiLSTM-CNN-DC method. The peak coordinates of activated brain regions reported in these two groups of articles were used to perform the meta-analysis respectively, by using the activation likelihood estimation (ALE) [26]. The results are shown in Fig. 9. The activated brain regions in the left sub-figure almost cover the whole brain. It denotes that the frequency-based Neurosynth topics cannot precisely characterize brain cognitive researches to create an accurate detailed map of brain functions [66]. Compared with the Neurosynth-based method, the activated brain regions in the right sub-figure cover generally accepted calculation-related brain regions. It denotes that the topics from CNN-BiLSTM-CNN-DC can better characterize brain cognitive researches.

V. CONCLUSION

For decoding complex brain cognition, large-scale modeling needs to characterize numerous entities. Therefore, the literature-based knowledge-driven modeling approach is necessary. Its basic is knowledge extraction from neuroimaing literature. At present, the researches on neuroimaging text mining have caused widely attentions, but there are still some shortcomings, including poor topic semantics and topic independent results. This paper designs the new task definition of neuroimaging text mining for curating brain cognitive researches. A topic learning pipeline CNN-BiLSTM-CNN-DC is proposed to extract key research information of the whole research process from open-access neuroimaging literature. The experimental results on actual data sets show that the proposed method can obtain more accurate and complete literature topics, and provide an automatic and reliable collection method of neuroimaging research knowledge for supporting the current extensive research on brain science, brain computer interface and artificial intelligence. With the rapid growth of open-access literature, the proposed method can also be expanded to realize domain knowledge extraction in other active science fields, such as geography, astronomy.

There are still some limitations in this study. The proposed method only aims at the topic extraction in a closed domain and cannot learn open domain topics beyond prior definitions. Furthermore, this proposed method adopts an asynchronous pipeline mode, which brings error propagation. The further work will focus on the topic learning in an open domain and developing the joint model of topic learning.

ACKNOWLEDGMENT

(Ying Sheng and Jianhui Chen contributed equally to this work.)

REFERENCES

[1] H. Kuai, X. Zhang, Y. Yang, J. Chen, B. Shi, and N. Zhong, ''THINKING-LOOP: The semantic vector driven closed-loop model for brain computing,'' *IEEE Access*, vol. 8, pp. 4273–4288, Jan. 2020.

[2] M. Wegrzyn, J. Aust, L. Barnstorf, M. Gippert, M. Harms, A. Hautum, S. Heidel, F. Herold, S. M. Hommel, A.-K. Knigge, D. Neu, D. Peters, M. Schaefer, J. Schneider, R. Vormbrock, S. M. Zimmer, F. G. Woermann, and K. Labudda, ''Thought experiment: Decoding cognitive processes from the fMRI data of one individual,'' *PLoS ONE*, vol. 13, no. 9, Sep. 2018, Art. no. e0204338.

[3] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, and K. Ugurbil, ''The WU-minn human connectome project: An overview,'' *NeuroImage*, vol. 80, pp. 62–79, Oct. 2013.

[4] B. J. Casey, T. Cannonier, M. I. Conley, A. O. Cohen, M. Deanna, M. M. Heitzeg, and A. M. Dale, "The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites," *Develop. Cognit. Neurosci.*, vol. 32, pp. 43–54, Mar. 2018.

[5] K. L. Miller, F. Alfaroalmagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, and S. M. Smith, "Multimodal population brain imaging in the UK biobank prospective epidemiological study," *Nature Neurosci.*, vol. 19, no. 11, pp. 1523–1536, Nov. 2016.

[6] J. D. Van Horn, J. S. Grethe, P. J. Kostelec, J. B. Woodward, J. A. Aslam, D. Rus, and M. S. Gazzaniga, "The functional magnetic resonance imaging data center (fMRIDC): The challenges and rewards of large-scale databasing of neuroimaging studies," *Phil. Trans. Roy. Soc. B*, vol. 356, no. 1412, pp. 1323–1339, Aug. 2001.

[7] J. D. Van Horn and A. W. Toga, "Is it time to re-prioritize neuroimaging databases and digital repositories?" *NeuroImage*, vol. 47, no. 4, pp. 1720–1734, Oct. 2009.

[8] R. A. Poldrack, D. M. Barch, J. P. Mitchell, T. D. Wager, A. D. Wagner, J. T. Devlin, C. Cumba, O. Koyejo, and M. P. Milham, "Toward open sharing of task-based fMRI data: The OpenfMRI project," *Frontiers Neuroinf.*, vol. 7, p. 12, 2013.

[9] K. J. Gorgolewski, O. Esteban, G. Schaefer, B. A. Wandell, and R. A. Poldrack, "OpenNeuro—A free online platform for sharing and analysis of neuroimaging data," in *Proc. Org. Hum. Brain Mapping*, Vancouver, BC, Canada, Jun. 2017, p. 1677.

[10] M. Shardlow, M. Ju, M. Li, C. O'Reilly, E. Iavarone, J. McNaught, and S. Ananiadou, "A text mining pipeline using active and deep learning aimed at curating information in computational neuroscience," *Neuroinformatics*, vol. 17, no. 3, pp. 391–406, Jul. 2019.

[11] F. Nielsen, L. K. Hansen, and D. Balslev, "Mining for associations between text and brain activation in a functional neuroimaging database," *Neuroinform.*, vol. 2, no. 4, pp. 369–379, Dec. 2004.

[12] K. J. Gorgolewski, G. Varoquaux, G. Rivera, Y. Schwartz, V. V. Sochat, S. S. Ghosh, C. Maumet, T. E. Nichols, J.-B. Poline, T. Yarkoni, D. S. Margulies, and R. A. Poldrack, "NeuroVault.Org: A repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain," *NeuroImage*, vol. 124, pp. 1242–1244, Jan. 2016.

[13] A. Badhwar, D. N. Kennedy, J. Poline, and R. Toro, "Distributed collaboration: The case for the enhancement of Brainspell's interface," *GigaSci.*, vol. 5, no. suppl_1, p. s13742-016, Nov. 2016.

[14] L. Moreau, B. Ludascher, I. Altintas, R. Barga, S. Bowers, S. P. Callahan, and Y. Zhao, "Special issue: The first provenance challenge," *Concurrency Comput. Pract. Exper.*, vol. 20, no. 5, pp. 409–418, Nov. 2007.

[15] D. B. Keator, K. Helmer, J. Steffener, J. A. Turner, T. G. M. Van Erp, S. Gadde, N. Ashish, G. A. Burns, and B. N. Nichols, "Towards structured sharing of raw and derived neuroimaging data across existing resources," *NeuroImage*, vol. 82, pp. 647–661, Nov. 2013.

[16] C. Maumet, T. Auer, A. Bowring, G. Chen, S. Das, G. Flandin, S. Ghosh, T. Glatard, K. J. Gorgolewski, K. G. Helmer, M. Jenkinson, D. B. Keator, B. N. Nichols, J.-B. Poline, R. Reynolds, V. Sochat, J. Turner, and T. E. Nichols, "Sharing brain mapping statistical results with the neuroimaging data model," *Scientific Data*, vol. 3, no. 1, Dec. 2016, Art. no. 160102.

[17] D. Keator, K. Helmer, C. Maumet, and J. Poline, "Tools for FAIR neuroimaging experiment metadata annotation with NIDM experiment," in *Proc. 25th Annu. Meeting Org. Hum. Brain Mapping (OHBM)*, Rome, Italy, 2019, pp. 1–5.

[18] T. Bolt, J. S. Nomi, R. Arens, S. G. Vij, M. C. Riedel, T. Salo, and L. Q. Uddin, "Ontological dimensions of cognitive-neural mappings," *Neuroinformatics*, vol. 10, pp. 1–13, Feb. 2020.

[19] T. N. Rubin, O. Koyejo, K. J. Gorgolewski, M. N. Jones, R. A. Poldrack, and T. Yarkoni, "Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition," *PLOS Comput. Biol.*, vol. 13, no. 10, Oct. 2017, Art. no. e1005649.

[20] S. Genon, A. Reid, H. Li, L. Fan, V. I. Müller, E. C. Cieslik, F. Hoffstaedter, R. Langner, C. Grefkes, A. R. Laird, P. T. Fox, T. Jiang, K. Amunts, and S. B. Eickhoff, "The heterogeneity of the left dorsal premotor cortex evidenced by multimodal connectivity-based parcellation and functional characterization," *NeuroImage*, vol. 170, pp. 400–411, Apr. 2018.

[21] R. A. Poldrack, J. A. Mumford, T. Schonberg, D. Kalar, B. Barman, and T. Yarkoni, "Discovering relations between mind, brain, and mental disorders using topic mapping," *PLoS Comput. Biol.*, vol. 8, no. 10, Oct. 2012, Art. no. e1002707.

[22] L. French, S. Lane, L. Xu, C. Siu, C. Kwok, Y. Chen, C. Krebs, and P. Pavlidis, "Application and evaluation of automated methods to extract neuroanatomical connectivity statements from free text," *Bioinformatics*, vol. 28, no. 22, pp. 2963–2970, Sep. 2012.

[23] F. H. Alhazmi, D. Beaton, and H. Abdi, "Semantically defined subdomains of functional neuroimaging literature and their corresponding brain regions," *Hum. Brain Mapping*, vol. 39, no. 7, pp. 2764–2776, Mar. 2018.

[24] A. B. Abacha, A. G. S. de Herrera, K. Wang, L. R. Long, S. Antani, and D. Demner-Fushman, "Named entity recognition in functional neuroimaging literature," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Kansas City, MO, USA, Nov. 2017, pp. 2218–2220.

[25] M. C. Riedel, T. Salo, J. Hays, M. D. Turner, M. T. Sutherland, J. A. Turner, and A. R. Laird, "Automated, efficient, and accelerated knowledge modeling of the cognitive neuroimaging literature using the ATHENA toolkit," *Frontiers Neurosci.*, vol. 13, p. 494, May 2019.

[26] A. Teghil, M. Boccia, F. D'Antonio, A. Di Vita, C. de Lena, and C. Guariglia, "Neural substrates of internally-based and externally-cued timing: An activation likelihood estimation (ALE) meta-analysis of fMRI studies," *Neurosci. Biobehavioral Rev.*, vol. 96, pp. 197–209, Jan. 2019.

[27] B. Gong, S. Naveed, D. M. Hafeez, K. I. Afzal, S. Majeed, J. Abele, S. Nicolaou, and F. Khosa, "Neuroimaging in psychiatric disorders: A bibliometric analysis of the 100 most highly cited articles," *J. Neuroimag.*, vol. 29, no. 1, pp. 14–33, Jan. 2019.

[28] R. A. Poldrack, A. Kittur, D. Kalar, E. Miller, C. Seppa, Y. Gil, D. S. Parker, F. W. Sabb, and R. M. Bilder, "The cognitive atlas: Toward a knowledge foundation for cognitive neuroscience," *Frontiers Neuroinform.*, vol. 5, p. 17, Sep. 2011.

[29] T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager, "Large-scale automated synthesis of human functional neuroimaging data," *Nature Methods*, vol. 8, no. 8, pp. 665–670, Jun. 2011.

[30] J. A. Turner and A. R. Laird, "The cognitive paradigm ontology: Design and application," *Neuroinformatics*, vol. 10, no. 1, pp. 57–66, Jan. 2012.

[31] H. Larochelle and S. Lauly, "A neural autoregressive topic model," in *Proc. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, 2012, pp. 2708–2716.

[32] V. Rakesh, W. Ding, A. Ahuja, N. Rao, Y. Sun, and C. K. Reddy, "A sparse topic model for extracting aspect-specific summaries from online reviews," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, Lyon, France, 2018, pp. 1573–1582.

[33] Y. Xu, J. Yin, J. Huang, and Y. Yin, "Hierarchical topic modeling with automatic knowledge mining," *Expert Syst. Appl.*, vol. 103, pp. 106–117, Aug. 2018.

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, May 2003.

[35] C. Guo, M. Lu, and W. Wei, "An improved LDA topic modeling method based on partition for medium and long texts," *Ann. Data Sci.*, vol. 7, pp. 1–14, Apr. 2019.

[36] G. Balikas, M. R. Amini, and M. Clausel, "On a topic model for sentences," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Pisa, Italy, 2016, pp. 921–924.

[37] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 299–313, Dec. 2015.

[38] S. Li, T. S. Chua, J. Zhu, and C. Miao, "Generative topic embedding: A continuous representation of documents," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 666–675.

[39] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Leveraging multi-domain prior knowledge in topic models," in *Proc. 23rd Int. Joint Conf. Artif. Intell. (IJCAI)*, Beijing, China, 2013, pp. 2071–2077.

[40] L. Yao, Y. Zhang, B. Wei, H. Qian, and Y. Wang, "Incorporating probabilistic knowledge into topic models," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, Ho Chi Minh City, Vietnam, 2015, pp. 586–597.

[41] R. K. Amplayo and S.-W. Hwang, "Aspect sentiment model for micro reviews," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, New Orleans, LA, USA, Nov. 2017, pp. 727–732.

[42] J. Zhu, "Research on topic modeling method based on deep learning," (in Chinese), M.S. thesis, Comput. School, Wuhan Univ., Wuhan, China, Tech. Rep. 10486, 2017.

[43] A. B. Dieng, C. Wang, J. Gao, and J. Paisley, "TopicRNN: A recurrent neural network with long-range semantic dependency," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, 2017, pp. 1–13.

[44] W. Zhang, Y. Li, and S. Wang, "Learning document representation via topic-enhanced LSTM model," *Knowl.-Based Syst.*, vol. 174, pp. 194–204, Jun. 2019.
[45] F. Yang, X. Zhao, and M. Zhang, "Research on topic mining algorithm based on deep learning extension," *J. Phys. Conf. Ser.*, vol. 1345, Nov. 2019, Art. no. 042034.
[46] C. Bogler, J. Mehnert, J. Steinbrink, and J.-D. Haynes, "Decoding vigilance with NIRS," *PLoS ONE*, vol. 9, no. 7, Jul. 2014, Art. no. e101729.
[47] M.-Y. Hsiao, C.-C. Chen, and J.-H. Chen, "Using UMLS to construct a generalized hierarchical concept-based dictionary of brain functions for information extraction from the fMRI literature," *J. Biomed. Informat.*, vol. 42, no. 5, pp. 912–922, Oct. 2009.
[48] R. A. Poldrack and K. J. Gorgolewski, "Making big data open: Data sharing in neuroimaging," *Nature Neurosci.*, vol. 17, no. 11, pp. 1510–1517, Oct. 2014.
[49] H. Zhong, J. Chen, T. Kotake, J. Han, N. Zhong, and Z. Huang, "Developing a brain informatics provenance model," in *Proc. Int. Conf. Brain Health Informat.*, Maebashi, Japan, 2013, pp. 439–449.
[50] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.
[51] W. Shalaby and W. Zadrozny, "Mined semantic analysis: A new concept space model for semantic representation of textual data," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Boston, MA, USA, Dec. 2017, pp. 2122–2131.
[52] K. Xu, Z. Yang, P. Kang, Q. Wang, and W. Liu, "Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition," *Comput. Biol. Med.*, vol. 108, pp. 122–132, May 2019.
[53] S. Zhang, Y. Sheng, J. Gao, J. Chen, J. Huang, and S. Lin, "A multi-domain named entity recognition method based on part-of-speech attention mechanism," in *Proc. CCF Conf. Comput. Supported Cooperat. Work Social Comput.*, Kunming, China, 2019, pp. 631–644.
[54] P. Song, C. Geng, and Z. Li, "Research on text classification based on convolutional neural network," in *Proc. Int. Conf. Comput. Netw., Electron. Autom. (ICCNEA)*, Xi'an, China, Sep. 2019, pp. 229–232.
[55] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. 25th Int. Conf. Comput. Linguistics Tech. Papers (COLING)*, Dublin, Ireland, 2014, pp. 2335–2344.
[56] B. He, Y. Guan, and R. Dai, "Classifying medical relations in clinical text via convolutional neural networks," *Artif. Intell. Med.*, vol. 93, pp. 43–49, Jan. 2019.
[57] Y. Cui, D. Liu, Q. Li, Z. Qiu, and X. Yang, "DTR: A novel topic generate algorithm based on dbscan and TextRank," in *Proc. The Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery*, Kunming, China, 2019, pp. 425–433.
[58] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. AFNLP: Volume 2 - ACL-IJCNLP*, Suntec, Singapore, 2009, pp. 1003–1011.
[59] Y. Huang, J. Hullfish, D. De Ridder, and S. Vanneste, "Meta-analysis of functional subdivisions within human posteromedial cortex," *Brain Struct. Function*, vol. 224, no. 1, pp. 435–452, Jan. 2019.
[60] P. Dino, S. Kumar, B. A. Ali, and H. Raj, "Bio-NER: Biomedical named entity recognition using rule-based and statistical learners," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 12, pp. 163–170, 2017.
[61] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinf.*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
[62] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102034.
[63] M. Xu, R. Yang, S. Ranshous, S. Li, and N. F. Samatova, "Leveraging external knowledge for phrase-based topic modeling," in *Proc. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Taipei, Taiwan, Dec. 2017, pp. 29–32.
[64] N. Loukachevitch, K. Ivanov, and B. Dobrov, "Thesaurus-based topic models and their evaluation," in *Proc. 8th Int. Conf. Web Intell., Mining Semantics (WIMS)*, New York, NY, USA, 2018, pp. 1–9.
[65] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from k-Fold cross validation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1586–1594, Aug. 2020.
[66] J. D. Lusher, J. X. Ji, and J. M. Orr, "Implementation of high-performance correlation and mapping engine for rapid generation of brain connectivity networks from big fMRI data," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Honolulu, HI, USA, Jul. 2018, pp. 1032–1036.

YING SHENG was born in Shanxi, China, in 1997. She received the B.S. degree in software engineering from the Taiyuan Institute of Technology, Taiyuan, China, in 2018. She is currently pursuing the master's degree in software engineering with the Beijing University of Technology, Beijing, China. Her current research interests include natural language processing, text mining, and big data.

JIANHUI CHEN received the Ph.D. degree from the Beijing University of Technology, China, in 2011. He then joined the Hong Kong Polytechnic University and Tsinghua University, China, as a Postdoctoral Fellow, from 2013 to 2015. He is currently an Assistant Professor with the International WIC Institute, Beijing University of Technology, China. His research interests include brain informatics, web intelligence, text mining, and semantic web with more than 30 journal and conference publications.

XIAOBO HE was born in Shanxi, China, in 1990. He received the B.E. degree from Southwest University, Chongqing, China, in 2016. He is currently pursuing the master's degree with the Beijing University of Technology, Beijing, China. His current interest includes automatic control theory and its applications.

ZHE XU was born in Shandong, China, in 1998. He received the B.E. degree in software engineering from Qingdao University, Shandong, China, in 2019. He is currently pursuing the master's degree in software engineering with the Beijing University of Technology, Beijing, China. His current research interest is natural language processing.

JIANGFAN GAO was born in Hebei, China, in 1994. He received the B.E. degree from North Minzu University, Ningxia, China, in 2017. He is currently pursuing the master's degree in software engineering with the Beijing University of Technology, Beijing, China. His research interest is natural language processing.

SHAOFU LIN received the B.S. degree from the University of Science and Technology of China, in 1990, and the Ph.D. degree from Peking University, China, in 2002. He is currently the Executive Vice Dean of the Beijing Institute of Smart City, Faculty of Information Technology, Beijing University of Technology, has memberships of the Beijing Institute of Big Data Research, the Expert Committee of the China Big Data Industry Ecological Alliance, the Cyber Technology Expert Committee of the China Artificial Intelligence Industry Alliance, and the China Computer Federation (CCF). His research interests focus on spatial-temporal computing, big data, the Internet of Things (IoT), and blockchain technologies.