Hindawi
Security and Communication Networks
Volume 2021, Article ID 6635027, 8 pages
https://doi.org/10.1155/2021/6635027

Research Article

Recognition of Disease Genetic Information from Unstructured Text Data Based on BiLSTM-CRF for Molecular Mechanisms

Lejun Gong,[1,2] **Xingxing Zhang**,[1] **Tianyin Chen**,[1] **and Li Zhang**[3]

[1]*Jiangsu Key Lab of Big Data Security & Intelligent Processing, School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China*
[2]*Zhejiang Engineering Research Center of Intelligent Medicine, Wenzhou 325035, China*
[3]*College of Computer Science and Technology, Nanjing Forestry University, Nanjing 210037, China*

Correspondence should be addressed to Lejun Gong; glj98226@163.com

Received 24 December 2020; Revised 20 January 2021; Accepted 5 February 2021; Published 19 February 2021

Academic Editor: Yuan Tian

Disease relevant entities are an important task in mining unstructured text data from the biomedical literature for achieving biomedical knowledge. Autism spectrum disorder (ASD) is a disease related to a neurological and developmental disorder characterized by deficits in communication and social interaction and by repetitive behaviour. However, this kind of disease remains unclear to date. In this study, it identifies entities associated with disease using the machine learning of a computational way from text data collection for molecular mechanisms related to ASD. Entities related to disease are extracted from the biomedical literature related to autism by using deep learning with bidirectional long short-term memory (BiLSTM) and conditional random field (CRF) model. Compared other previous works, the approach is promising for identifying entities related to disease. The proposed approach including five types of molecular entities is evaluated by GENIA corpus to obtain an F-score of 76.81%. The work has extracted 9146 proteins, 145 RNAs, 7680 DNAs, 1058 cell-types, and 981 cell-lines from the autism biomedical literature after removing repeated molecular entities. Finally, we perform GO and KEGG analyses of the test dataset. This study could serve as a reference for further studies on the etiology of disease on the basis of molecular mechanisms and provide a way to explore disease genetic information.

1. Introduction

With the rapid development of intelligent computing and machine learning technology, especially the development of deep learning technology [1], artificial intelligence technology has developed more widely involving algorithms and applications [2–6]. Moreover, it has been widely used in academia and industry such as communication security [7, 8] and opinion and text mining [9–11]. It is also popular in the biomedical field [12, 13]. Abundant experimental data in biomedical research are available [14]. A large number of terminological resources and knowledge bases can also be used in machine learning methods for biomedical text mining [15]. Hassanpour et al. [16] provided a semantic-based method for extracting concept definitions for scientific publications on autism phenotype. Thabtah et al. [17] proposed a new computational intelligence approach based on variable analysis to detect features for autism screening. Spencer et al. [18] found gene associations using frequent pattern mining specific to autism. Bush et al. [19] extracted ASD data from electronic health records for different workflows. Macedoni-Lukšič et al. [20] used ontology construction to identify the main concepts in autism by using the RaJoLink method based on Swanson's ABC model. In our previous work, we extracted candidate genes related to autism based on associated rules [21].

Autism is a neurodevelopmental disorder called autism spectrum disorder (ASD). ASD is a neurological and developmental disorder characterized by deficits in communication, social interaction, and repetitive behaviour. It is also syndrome about neurodevelopment with an as yet unknown unifying pathological or neurobiological etiology.

Zhang et al. [22] conducted genome-wide association study and integrated brain region-related enhancer-gene networks for ASD to explore the roles of chromosomal enhancer region in this disorder. Parr et al. [23] employed Bayesian frameworks to understand brain function formulate perception and action as inferential processes. Sato et al. [24] combined fuzzy spectral clustering and entropy analysis of functional MRI data to identify segregated regions in the functional brain connectome of individuals with autism. They also proved efficiency of this new tool to characterize neuropsychiatric disorders [25]. Rosenberg et al. [26] proposed that the alterations in nonlinear, canonical computations underlie the behavioural characteristics of individuals with autism. They believe that computational perspective on autism may aid in identifying physiological pathways to target in ASD treatment. The abovementioned computational approach can be employed to explore the etiology of autism without the need for expensive and time-consuming experimental validation. Although the etiology of ADS remains unclear, some studies have demonstrated that strong genetic components are involved in ASD development [27–29]. In the present study, we explored the molecular mechanisms related to ASD through computations to understand the etiology of this disorder.

To explore the underlying disease's mechanisms, we identified five disease entities related to autism based on deep learning using the hybrid model containing both bidirectional long short-term memory (BiLSTM) [30] and conditional random field (CRF) [31] model, and explored the molecular mechanism by analysing their relationships among molecular entities.

2. Materials and Methods

As a large unstructured data repository, the biomedical literature contains abundant biomedical information from which useful knowledge (specific and relevant interest points) can be obtained by subjecting unstructured text to natural language processing. In this study, molecular information related to autism was obtained from the biomedical literature. We first extracted molecular entities from experimental corpus by using a suitable computational model and then explored their relationships among molecular entities. Then, we divided these entities related to autism into confirmed and unknown samples. Finally, we explored known samples related to autism to understand the ethology of the disorder, which could offer a reference for understanding the unknown molecular mechanisms of the other samples related to autism.

Identifying molecular entities is a key factor in this study. Machine learning is the mainstream method. The task is considered a sequence tagging NLP problem. The output of taggers could be used for downstream input in sequence tagging. Some linear statistical models that have been applied in sequence tagging include the Hidden Markov model [32], maximum entropy Markov models [33], and CRF models. Recently, neural networks have been proposed to tackle the sequence tagging problem [34–36]. This study combined a hybrid network both BiLSTM and CRF to form a BiLSTM-CRF model for identifying molecular entities. The network could efficiently use past input features via a BiLSTM layer and sentence level tag information via a CRF layer. The following sections would describe the model of identifications.

2.1. LSTM Model. Long short-term memory (LSTM) [37] networks are similar to recurrent neural networks (RNNs). RNNs could not learn the relevant information of input data with sigma cells or tanh cells. The hidden layer updates are replaced by purpose-built memory cells in LSTM. Thus, LSTM is a special recurrent neural network model which could selectively store contextual information using a specially designed gate structure containing input gate, output gate, and forget gate. LSTM could handle the long-term dependencies well. The LSTM memory cell is illustrated in the works [30, 38]. By forgetting the information in the cell state and memorizing new information, this allows information that is useful for subsequent moments of computation to be transmitted, while useless information is discarded, and the hidden layer state is output at each time step. The values of the forgetting, memory and output are controlled by the state of the hidden layer at the last moment and the values of the memory gate, memory gate, and output gate calculated by the current input.

Generally, LSTM includes five computational processes: (1) calculating the forgetting gate and selecting the information to be forgotten; (2) calculating the memory gate and selecting the information to remember; (3) calculating the current cell state at the moment; (4) calculating the output gate and the state of the hidden layer at the current moment; (5) obtaining a hidden layer state sequence of the same length as the sentence. More details are described in [37]. The threshold mechanism of LSTM can effectively filter and memorize the information of the memory unit to solve the problem of RNN. However, the LSTM only captures the forward information from text. For the named entity identification tasks, the backward propagation information has also important reference values. Therefore, the hybrid network is applied in the work in the following section.

2.2. Hybrid Network. Hybrid network level contains the two parts: both the bidirectional LSTM network (BiLSTM) and CRF. The level of BiLSTM is utilized in the sequence tagging task to access both past and future input features. It mainly depends on forward and backward states resulting in two separate hidden states for capturing past and future information, respectively. In this study, the BiLSTM is used to obtain more contextual information. The input sequences $x = (x_1, x_2, \ldots, x_k)$ are put into the neural network. For each input sequence (x_i) in a sentence, it is converted into word embedding. These words in a given sentence are embedded into a BiLSTM network where the forward and backward representation of each word is computed. The symbol $\overrightarrow{h}_t$ is acted as the output of the forward LSTM at a t time, and the symbol $\overleftarrow{h}_t$ is referred as the output representation of the reverse LSTM at t time. The output representation of BiLSTM at t time is defined as $h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t]$. Thus, this

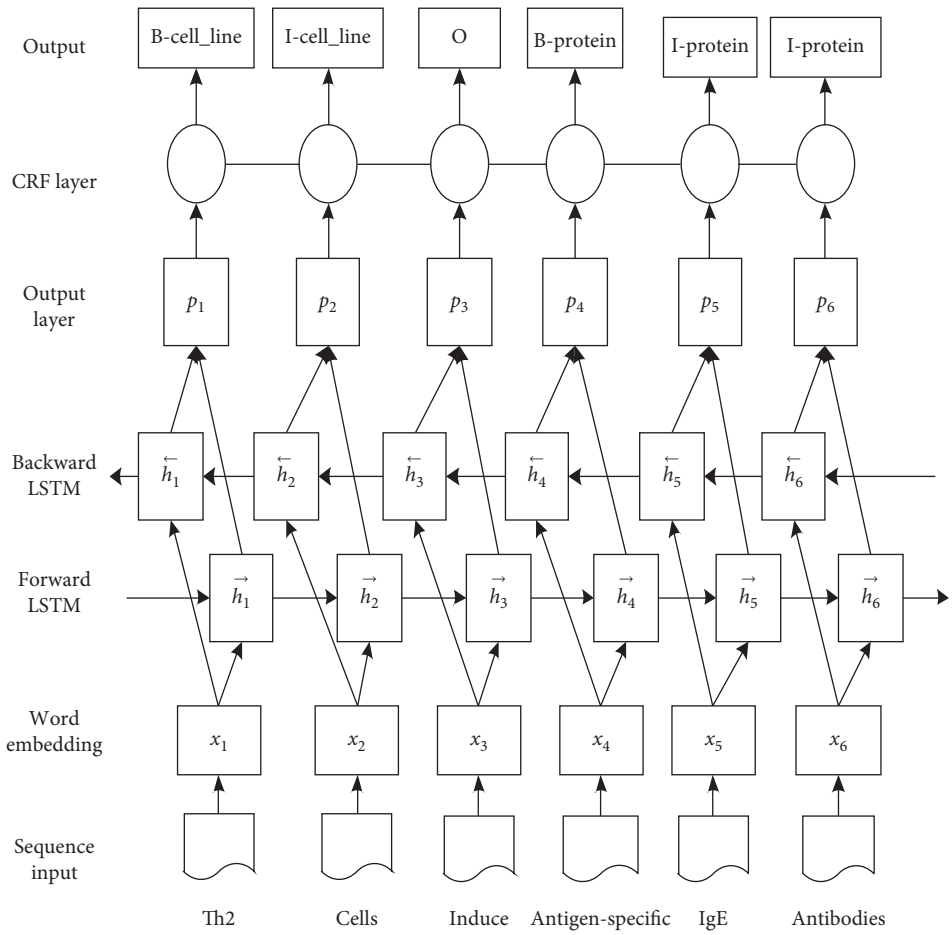

Figure 1: Pipeline of identified entities with the hybrid network.

output contains the more context information. It is used to labelling named entity in the text. The other network level is the condition random field (CRF) model, which focuses on sentence level instead of individual positions in sequence labelling tasks. It makes use of neighbour tag information for predicting the current tags. It is helpful that the correlation between labels in neighbourhoods and jointly decoding the best chain of labels for a given input sentence. Considering the relationship of adjacent labels, the linear CRFs can obtain a globally optimal labelling sequence which could maximize the relationship of adjacent tags. Moreover, it also optimizes the output tag sequences globally and demonstrates enhanced recognition performance for biologically named entities with larger lengths and modified vocabulary. The hybrid network integrated the two network's advantages for more identifying molecular entities.

2.3. Pipeline of Identified Entities. In this study, hybrid network contains both BiLSTM and CRFs. The output of the BiLSTM model is used as the input of the CRFs model to acquire the global optimal marker sequence. Word embedding which is a means of mapping a vocabulary to a real vector for capturing the distributed syntax and semantic information of the words launched by Google is used to switch words to vectors by word2vec. Aiming at the multiword entities, IOB tagging is used to detect entity boundary detection. The label "B" indicates the beginning of the boundary of the entity, the label "I" indicates the intermediate entity, and the label "O" indicates the nonbiological medical entity. Thus, the entity would be tagged as B-entity_category, I-entity_category, and O. For example, when the word is part of protein, it would be tagged as B-protein, I-protein, and O. The pipeline of identified instance "Th2 cells induce antigen-specific IgE antibodies" is shown in Figure 1.

3. Results and Discussion

This study used GENIA [39] corpus to evaluate which is annotated by professional researchers which is a semantically annotated dataset about the biomedical literature to validate the method of entity identification. It also provides the gold standard for the evaluation of text mining systems. GENIA corpus is extracted from the MEDLINE database with MEDLINE ID, title, and abstract encoded in an XML-database. Aiming at the abovementioned approach, we focused on five categories of entities, namely, DNA, protein, RNA, cell-type, and cell-line using three popular measurements which are used the works [40]. The experimental results are illustrated in Table 1. Our approach achieved an

Table 1: Performance of identified molecular entities.

Molecular entity	P (%)	R (%)	F-score (%)
Protein	84.32	80.32	82.27
DNA	76.28	71.33	73.72
RNA	85.71	77.97	81.66
Cell-type	83.67	80.37	81.98
Cell-line	65.22	63.64	64.42
Overall	79.04	74.72	76.81

Compared other previous works, Table 2 illustrates the comparison between our approach and previous works.

Table 2: Comparisons between previous works and our approach.

Approach	P (%)	R (%)	F-score (%)
Zhou et al. [41]	75.99	69.42	72.55
Liao and Wu [42]	72.80	73.60	73.20
Tang et al. [43]	70.78	72.00	71.39
Yao et al. [44]	76.13	66.54	71.01
Li et al. [45]	74.77	70.85	72.76
Li and Guo [46]	**79.58**	69.86	74.40
Our approach	79.04	**74.72**	**76.81**

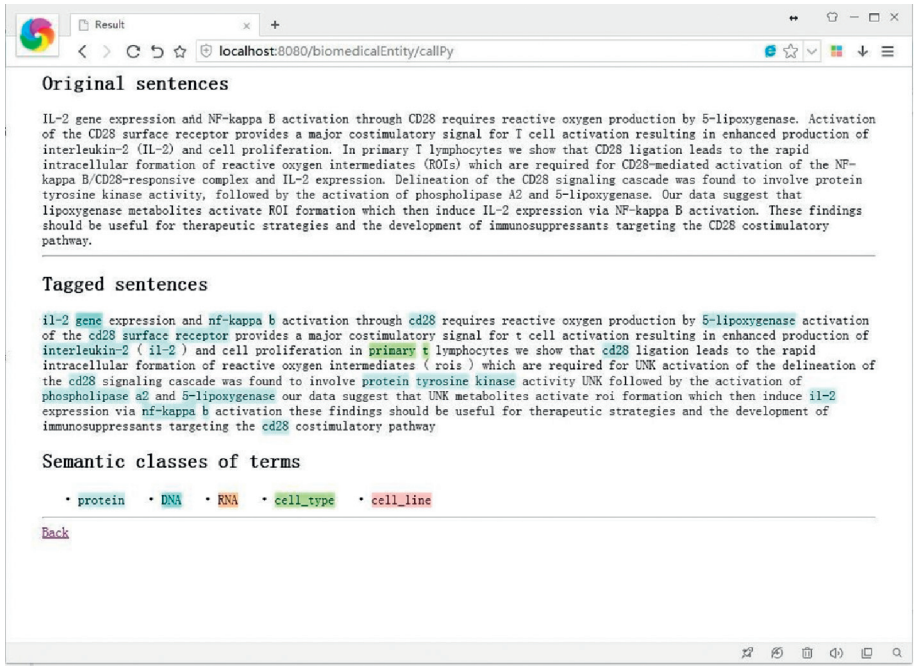

Figure 2: Screen shot of identified molecular entities.

F-score of 76.81%. Table 2 illustrates the comparison between our approach and previous works and previously reported ones.

Zhou et al. [41] identified entities with 72.55% F-score. Liao and Wu [42] used artificial features to construct a skip-chain CRF model that considers long-distance dependencies with an F-score of 73.20% in GENIA corpus. Nevertheless, this paper proposes the BiLSTM-CRF model, which does not use any artificial features but obtains better results in GENIA corpus than the model used by Liao and Wu [42]. Yao et al. [44] used a multilayer neural network learning feature representation and achieved an F-score of 71.01%. Li and Guo [46] constructed a BiLSTM model with word and character vectors and obtained an F-score of 74.40%. Our proposed method obtained an F-score of 76.81%, indicating that our approach is better than those in previous works [42–46]. Thus, it is promising for extracting molecular entities from the biomedical literature.

In this study, we also used the key word "autism" to search the NCBI database, including 29767 literature studies until August 12, 2018. The approach have extracted 9146 proteins, 145 RNAs, 7680 DNAs, 1058 cell-types, and 981 cell-lines after removing repeated molecular entities. In these extracted molecular entities, the MECP2 gene appears

Table 3: The same 70 genes compared to the ripe genes in the work [11].

The same 70 genes compared to the ripe genes in the work [11]						
OT	ERK	RORA	FOXP1	TCF4	CDH13	VEGF
TRPV1	NLGN4	HMGB1	NRG3	UPS	HNF1B	ST8SIA
PAFAH1B1	TNF	FGF22	HDAC4	TLR3	NTK2	CDH8
SCN3A	DIA1	L1CAM	CRK	NOS1	VP	AGC1
CACNA1A	SHOX	ATP8A1	MVP	NR4AL	WNT1	FMR2
SOX5	CRBN	SUSD4	DAT1	MAPT	MTNRLA	ATRNL1
LRRTM3	DLG4	PCDH15	MKL2	RPP25	OGG1	CTCF
SLOS	GLUT1	KIF1A	GRIA1	ID3	BDMR	INS
TSGA14	CRHR1	CD28	GAS	TSC	BF	GATM
MDR1	SOX9	GAP43	ARA	PLA2	FOSB	WMS

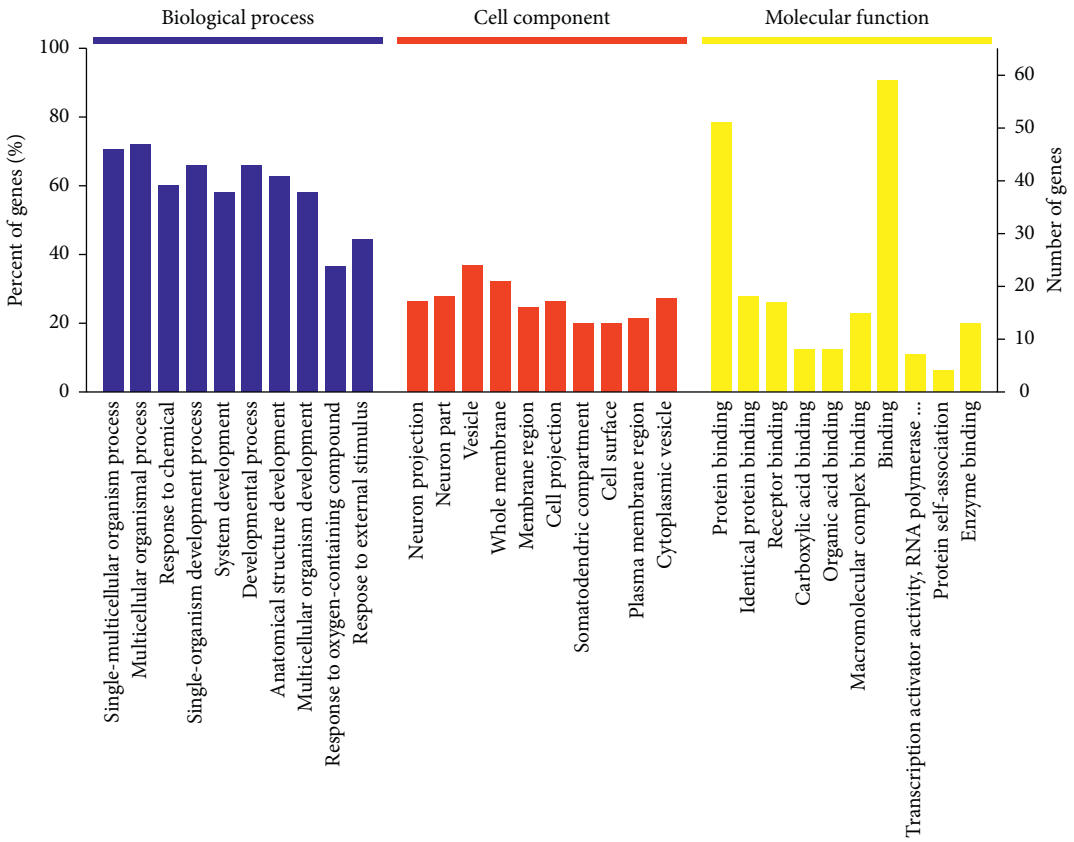

Figure 3: GO analysis related to 70 genes.

most frequently, followed by gene the OXYTOCIN gene in the experimental dataset. The two genes are confirmed as autism susceptibility genes. We used Python to extract molecular entities related to autism and developed an identification system. The screen shot is shown in Figure 2.

Compared to the ripe genes in the work [11], there are the same 70 genes in the extracted entities. They are shown in Table 3. GO and KEGG analyses of the 70 genes are shown in Figures 3 and 4, respectively.

GO analysis showed that about 70% of the genes participate in developmental process and nearly 50% of the genes participate in response to external stimulus as shown in Figure 3. Nearly 30 genes are located in neuron projection and partly in cell component. Finally, about 90% of the genes show binding and protein binding molecular functions.

KEGG analysis indicated that some of these genes are associated with long-term depression, glutamatergic synapse, dopaminergic synapse, and circadian entrainment in the nervous system as shown in Figure 4. About 9% of the genes participate in the MAPK signaling pathway. Both GO and KEGG analyses of the known genes related to autism provide a reference for understanding the molecular

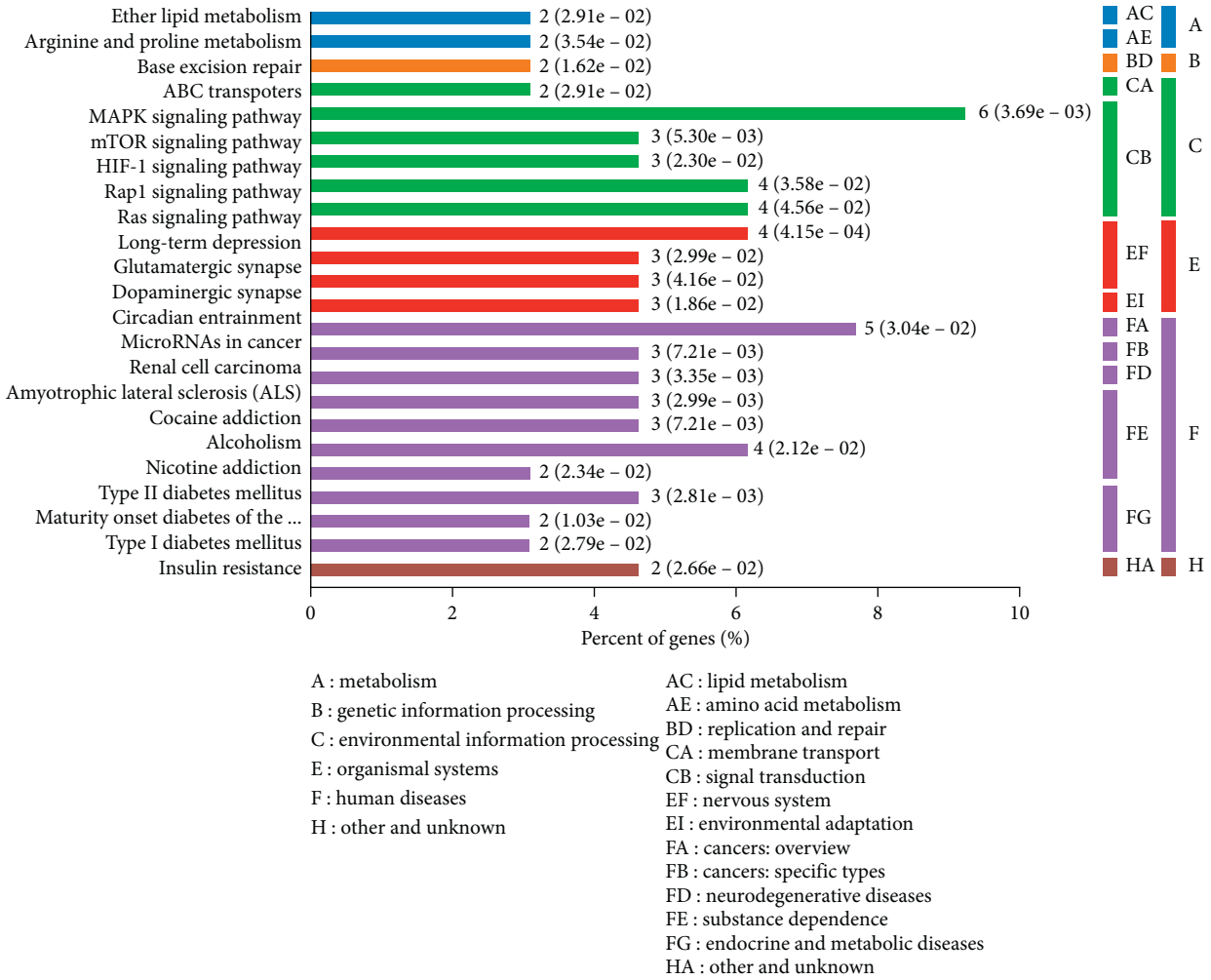

Figure 4: KEGG analysis related to 70 genes.

mechanism of the unknown samples, which could find new genes related to autism.

4. Conclusions

Entities related to disease were identified using the BiLSTM-CRF model, and the approach was evaluated with an F-score of 76.81%. To the best of our knowledge, the provided approach is state-of-art compared the previous works. Based on the approach, we also develop an identified system. Meanwhile, this study also analyses the extracted genes by GO and KEGG analyses. The proposed approach will be applied to explore other molecular mechanisms related to other neurological-diseases, such as Parkinson. This study can serve as a reference for understanding disease etiology, which is promising for identifying disease entities.

Data Availability

The experiment dataset related to the autism biomedical literature was extracted from the PubMed database with E-utilities (http://eutils.ncbi.nlm.nih.gov/corehtml/query/static/eutils_help.html) by using the key word "autism." The biomedical corpus plays an important role in biomedical text mining for achieving the biomedical knowledge domain. It promoted the blossom of text mining technology based on machine learning. GENIA corpus provides a reference material using natural language processing techniques for biomedical text mining. It is a semantically annotated dataset that provides evaluation criteria for text mining approaches. It is also annotated by authoritative domain experts for biological terms encoded in an XML-based markup scheme. This study applied GENIA corpus to build a method about the identification of molecular entities.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (grant nos. 61502243 and 61802193), Natural Science Foundation of Jiangsu Province (BK20170934), Zhejiang Engineering Research Center of

Intelligent Medicine under 2016E10011, China Postdoctoral Science Foundation (2018M632349), NUPTSF (NY217136), and Natural Science Foundation of the Higher Education Institutions of Jiangsu Province in China (16KJD520003).

References

[1] R. Huan, T. Ma, J. Cao, Y. Tian, and A.-D. Abdullah, "Deep rolling: a novel emotion prediction model for a multi-participant communication context," *Information Sciences*, vol. 488, pp. 158–180, 2019.

[2] L. Fu, Z. Li, Q. Ye et al., "Learning robust discriminant subspace based on joint L2,p- and l2,s-norm distance metrics," *IEEE Transactions on Neural Networks and Learning Systems*, 2020, Early Access.

[3] L. Fu, D. Zhang, and Q. Ye, "Recurrent thrifty attention network for remote sensing scene recognition," *IEEE Transactions on Geoscience and Remote Sensing*, 2020, Early Access.

[4] Q. Ye, Z. Li, L. Fu, Z. Zhang, W. Yang, and G. Yang, "Nonpeaked discriminant analysis for data representation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 12, pp. 3818–3832, 2019.

[5] Q. Ye, J. Yang, F. Liu, C. Zhao, N. Ye, and T. Yin, "L1-Norm distance linear discriminant analysis based on an effective iterative algorithm," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 1, pp. 114–129, 2018.

[6] Q. Ye, H. Zhao, Z. Li et al., "L1-Norm distance minimization-based fast robust twin support vector k -plane clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 4494–4503, 2018.

[7] B. Al-Otibi, N. Al-Nabhan, and Y. Tian, "Privacy-preserving vehicular rogue node detection scheme for fog computing," *Sensors*, vol. 19, no. 4, p. 965, 2019.

[8] Y. Tian, M. M. Kaleemullah, M. A. Rodhaan et al., "A privacy preserving location service for cloud-of-things system," *Journal of Parallel and Distributed Computing*, vol. 123, p. 215, 2019.

[9] Z. Pan, C.-N. Yang, S. Sheng Victor, N. Xiong, and W. Meng, "Machine learning for wireless multimedia data security," *Security and Communication Networks*, vol. 2019, Article ID 7682306, 2019.

[10] T. Ma, R. Huan, Y. Hao, J. Cao, Y. Tian, and Al-R. Mznah, "A novel sentiment polarity detection framework for Chinese," *IEEE Transactions on Affective Computing*, 2019.

[11] L. Gong, R. Yang, and X. Sun, "Prioritization of disease susceptibility genes using LSM/SVD," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 12, pp. 3410–3417, Article ID 000327554000020, 2013.

[12] L. Gong, Y. Yan, J. Xie, H. Liu, and X. Sun, "Prediction of autism susceptibility genes based on association rules," *Journal of Neuroscience Research*, vol. 90, no. 6, pp. 1119–1125, Article ID 000302536300002, 2012.

[13] L. Gong, X. Sun, D. Jiang, and S. Gong, "AutMiner: a system for extracting ASD-related genes using text mining," *Journal of Biological Systems*, vol. 19, no. 1, pp. 113–125, Article ID 000288809600007, 2011.

[14] W. W. M. Fleuren and W. Alkema, "Application of text mining in the biomedical domain," *Methods*, vol. 74, pp. 97–106, 2015.

[15] A. Jimeno Yepes and R. Berlanga, "Knowledge based word-concept model estimation and refinement for biomedical text mining," *Journal of Biomedical Informatics*, vol. 53, pp. 300–307, 2015.

[16] S. Hassanpour, M. J. O'Connor, and A. K. Das, "A semantic-based method for extracting concept definitions from scientific publications: evaluation in the autism phenotype domain," *Journal of Biomedical Semantics*, vol. 4, no. 1, p. 14, 2013.

[17] F. Thabtah, F. Kamalov, and K. Rajab, "A new computational intelligence approach to detect autistic features for autism screening," *International Journal of Medical Informatics*, vol. 117, pp. 112–124, 2018.

[18] M. Spencer, N. Takahashi, S. Chakraborty, J. Miles, and C.-R. Shyu, "Heritable genotype contrast mining reveals novel gene associations specific to autism subgroups," *Journal of Biomedical Informatics*, vol. 77, pp. 50–61, 2018.

[19] R. A. Bush, C. D. Connelly, A. Pérez, H. Barlow, and G. J. Chiang, "Extracting autism spectrum disorder data from the electronic health record," *Applied Clinical Informatics*, vol. 8, no. 3, pp. 731–741, 2017.

[20] M. Macedoni-Lukšič, I. Petrič, B. Cestnik, and T. Urbančič, "Developing a deeper understanding of autism: connecting knowledge through literature mining," *Autism Research and Treatment*, vol. 2011, Article ID 307152, 10 pages, 2011.

[21] L. Gong, Y. Yan, J. Xie, H. Liu, and X. Sun, "Prediction of autism susceptibility genes based on association rules," *Journal of Neuroscience Research*, vol. 90, no. 6, pp. 1119–1125, 2012.

[22] L. Zhang, L. Liu, Y. Wen et al., "Genome-wide association study and identification of chromosomal enhancer maps in multiple brain regions related to autism spectrum disorder," *Autism Research*, vol. 12, no. 1, p. 26, 2018.

[23] T. Parr, G. Rees, and K. J. Friston, "Computational neuropsychology and bayesian inference," *Frontiers in Human Neuroscience*, vol. 12, p. 61, 2018.

[24] J. R. Sato, J. Balardin, M. C. Vidal, and A. Fujita, "Identification of segregated regions in the functional brain connectome of autistic patients by a combination of fuzzy spectral clustering and entropy analysis," *Journal of Psychiatry & Neuroscience*, vol. 41, no. 2, pp. 124–132, 2016.

[25] J. R. Sato, M. Calebe Vidal, S. de Siqueira Santos, K. Brauer Massirer, and A. Fujita, "Complex network measures in autism spectrum disorders," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 2, pp. 581–587, 2018.

[26] A. Rosenberg, J. S. Patterson, and D. E. Angelaki, "A computational perspective on autism," *Proceedings of the National Academy of Sciences*, vol. 112, no. 30, pp. 9158–9165, 2015.

[27] S. Jamain, H. Quach, H. Quach et al., "Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism," *Nature Genetics*, vol. 34, no. 1, pp. 27–29, 2003.

[28] A. M. Persico and T. Bourgeron, "Searching for ways out of the autism maze: genetic, epigenetic and environmental clues," *Trends in Neurosciences*, vol. 29, no. 7, pp. 349–358, 2006.

[29] J. F. Abelson, K. Y. Kwan, B. J. O'Roak et al., "Sequence variants in SLITRK1 are associated with Tourette's syndrome," *Science*, vol. 310, no. 5746, pp. 317–320, 2005.

[30] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, http://arxiv.org/abs/1508.01991.

[31] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," *Proceedings of ICML*, vol. 28, 2001.

[32] L. Patel, N. Gustafsson, Y. Lin, R. Ober, R. Henriques, and E. Cohen, "A hidden markov model approach to

characterizing the photo-switching behavior OF fluorophores," *The Annals of Applied Statistics*, vol. 13, no. 3, pp. 1397–1429, 2019.

[33] R. Cofré, C. Maldonado, and F. Rosas, "Large deviations properties of Maximum entropy Markov chains from spike trains," *Entropy*, vol. 20, no. 8, p. 573, 2018.

[34] M. Yin, C. Mou, K. Xiong, and J. Ren, "Chinese clinical named entity recognition with radical-level feature and self-attention mechanism," *Journal of Biomedical Informatics*, vol. 98, Article ID 103289, 2019.

[35] M. Basaldella, L. Furrer, C. Tasso, and F. Rinaldi, "Entity recognition in the biomedical domain using a hybrid approach," *Journal of Biomedical Semantics*, vol. 8, no. 1, p. 51, 2017.

[36] X. Wang, Y. Zhang, X. Ren et al., "Cross-type biomedical named entity recognition with deep multi-task learning," *Bioinformatics*, vol. 35, no. 10, pp. 1745–1752, 2019.

[37] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[38] X. Yang, Y. Li, L. Gong et al., "Bidirectional LSTM-CRF for biomedical named entity recognition," in *Proceedings of the 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Huangshan, China, July 2018.

[39] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus--a semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. 1, pp. i180–i182, 2003.

[40] L. Gong, R. Yang, Q. Liu, Z. Dong, H. Chen, and G. Yang, "A dictionary-based approach for identifying biomedical concepts," in *Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery*, Changsha, China, August 2005.

[41] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan, "Recognizing names in biomedical texts: a machine learning approach," *Bioinformatics*, vol. 20, no. 7, pp. 1178–1190, 2004.

[42] Z. Liao and H. Wu, "Biomedical named entity recognition based on skip-chain CRFS," in *Proceedings of the 2012 International Conference on Industrial Control and Electronics Engineering*, pp. 1495–1498, Xi'an, China, August 2012.

[43] B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu, "Evaluating word representation features in biomedical named entity recognition tasks," *BioMed Research International*, vol. 2014, Article ID 240403, 2 pages, 2014.

[44] L. Yao, H. Liu, Y. Liu, X. Li, and M. W. Anwar, "Biomedical named entity recognition based on deep neutral network," *International Journal of Hybrid Information Technology*, vol. 8, no. 8, pp. 279–288, 2015.

[45] L. Li, L. Jin, Y. Jiang et al., "Recognizing biomedical named entities based on the sentence vector/twin word embeddings conditioned bidirectional LSTM," in *Proceedings of the China National Conference on Chinese Computational Linguistics*, pp. 165–176, Kunming, China, October 2019.

[46] L. Li and Y. Guo, "Biomedical named entity recognition based on CNN-BLSTM-CRF model," *Chinese Journal of Information*, vol. 1, pp. 116–122, 2018.