IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 October 2022, accepted 22 November 2022, date of publication 24 November 2022,
date of current version 30 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3224802

RESEARCH ARTICLE

Biomedical Word Sense Disambiguation Based on Graph Attention Networks

CHUN-XIANG ZHANG, MING-LEI WANG, AND XUE-YAO GAO
School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China

Corresponding author: Xue-Yao Gao (xueyao_gao@163.com)

This work was supported by the Heilongjiang Provincial Natural Science Foundation of China under Grant LH2022F031.

ABSTRACT Biomedical words have many semantics. Biomedical word sense disambiguation (WSD) is an important research issue in biomedicine field. Biomedical WSD refers to the process of determining meanings of ambiguous word according to its context. It is widely applied to process, translate and retrieve biomedical texts now. In order to improve WSD accuracy in biomedicine, this paper proposes a new WSD method based on graph attention neural network (GAT). Words, parts of speech, and semantic categories in context of ambiguous word are used as disambiguation features. Disambiguation features and the sentence are used as nodes to construct WSD graph. GAT is used to extract discriminative features, and softmax function is applied to determine semantic category of biomedical ambiguous word. MSH dataset is used to optimize GAT-based WSD classifier and test its accuracy. Experiments show that average accuracy of the proposed method is improved. At the same time, majority voting strategy is adopted to optimize GAT-based WSD classifier further.

INDEX TERMS Biomedical word, word sense disambiguation, graph attention neural network, part of speech, semantic category, WSD graph.

I. INTRODUCTION

With the rapid development of biomedicine, the number of biomedical vocabulary is increasing. We need specific tools to process biomedical texts. However, it is very difficult to process biomedical texts in many cases. This is because many biomedical words have multiple meanings, which results in ambiguity of biomedical text. Faced with these challenges, we need to design a novel and effective tool to solve ambiguities of biomedical words. For example, biomedical word 'BLM' has two semantics, including 'Bureau of Land Management' and 'Bleomycin'. So, we need determine correct meanings of biomedical word according to its context. Biomedical WSD is the process of assigning ambiguous word with unambiguous sense.

We extract two sentences containing ADA from the corpus. The first sentence is 'Third dental therapeutics guide debuts at ADA session'. The second sentence is 'We isolated a novel ADA inhibitor from a culture of Bacillus spJ89 and evaluated its anti proliferative activity on human cancer cell lines'. ADA has two semantics. The first one is American Dental Association and the second one is Adenosine deaminase. They are all abbreviated to ADA. The semantics of ADA in the first sentence is American Dental Association. The semantics of ADA in the second sentence is Adenosine deaminase. We explore the full form of ADA and find it to be ambiguous in original text. We can find that ADA is ambiguous and should be disambiguated based on its contexts.

Now, biomedical WSD is widely applied to document classification, information extraction and document retrieval in biomedicine field. Biomedical WSD methods are divided into 3 categories: supervised method, unsupervised one and knowledge-based one.

In supervised WSD method, human-annotated instances are used to train WSD classifier, which assigns semantic category to test instance [1]. In unsupervised method, structural knowledge is learned from unlabeled instances to determine category of biomedical ambiguous word [2]. In knowledge-based method, thesauri and sense inventories are applied to disambiguate biomedical ambiguous words. For example,

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

WordNet and the Unified Medical Language System (UMLS) are important thesaurus which define different senses and corresponding synonyms [3]. We propose a new biomedical WSD method based on GAT. The main innovations and contributions of this paper are summarized as follows:

- Words, parts of speech, semantic categories from contexts and sentence containing biomedical ambiguous word are used as disambiguation features. Word2vec and doc2vec tools are adopted to extract feature vectors from disambiguation features.
- WSD graph is constructed. We construct WSD data as graph and solve WSD problem in graph-structured data. Disambiguation features are used as nodes in graph. Edges are established between word nodes and sentence ones, word nodes and part of speech ones, word nodes and semantic category ones.
- Multi-head graph attention mechanism is adopted to adjust dynamically weight between two neighbor nodes.

This paper is organized as follows. Related work is reported in Section II. WSD feature extraction is given in Section III. WSD based on GAT is described in Section IV. Experimental results are given and analyzed in Section V. Conclusion is described in Section VI.

II. RELATED WORK

Mcinnes firstly researches on Biomedical WSD [4]. WSD method is divided into supervised one, unsupervised one, and knowledge-based one.

In supervised WSD method, labeled data is used to train WSD classifier. Wang gives an interactive learning algorithm with expert labeling instances and features [5]. Experts provide supervision in 3 ways: labeling instances, specifying indicative words of a sense, and highlighting the supporting evidence in a labeled instance. Zhang proposes two supervised WSD models based on deep learning technology [6]. One is based on Bi-directional Long Short-Term Memory (BiLSTM) network, and the other is based on self-attention mechanism. Yepes evaluates several features from contexts of ambiguous word and uses word embeddings to explore global features from MEDLINE [7]. Festag investigates WSD based on word embeddings and recurrent convolutional neural networks [8]. He focuses on terms mapped to multiple concepts of the UMLS. Antunes gives a supervised biomedical WSD method which uses bag-of-words as local features, and utilizes word embeddings as global features [9]. Bis proposes a novel deep neural network for supervised medical WSD based on a layered bidirectional LSTM network which performs a max-pooling along multiple time steps to create dense representation of the context [10]. Lui suggests that the notion of destination is a strong predictor of pedestrian trajectories and proposes a novel enhancement of the data-driven approach for pedestrian tracking in public buildings [11]. Qiao presents DEep contextualized biomedical abbreviation expansion model, which automatically collects substantial and relatively clean annotated contexts for 950 ambiguous abbreviations from PubMed abstracts using a simple heuristic [12]. Supervised methods usually have high accuracy. But, training data need be labeled, and the cost of labeling data is relatively high.

In unsupervised WSD method, unlabeled corpus is clustered to determine semantic category of ambiguous word. Pesaranghader designs deepBioWSD model which leverages 1 single bidirectional LSTM network to predict sense of ambiguous term [13]. Smalheiser gives similarity metrics for relating two medical subject headings with each other [14]. Li takes account of word order and presents a novel language model based on Bi-LSTM to embed sentential context in continuous space [15]. The proposed model generates contextual representations in an unsupervised manner. Duque presents a graph-based unsupervised biomedical WSD method, in which knowledge base is a graph built with co-occurrence information from medical concepts in scientific abstracts [16]. Ren proposes biomedical WSD method based on Convolutional Neural Network [17]. A large scale of relevant corpus from MEDLINE is crawled for training and contextual feature vectors are obtained. El-Rab applies six relation types of UMLS to build a graph for ambiguous word and gives a graph-based algorithm to disambiguate terms in biomedical text [18]. Moon discusses feature selection for disambiguation of acronyms and abbreviations in clinical domain [19]. Ren predefines the number of senses [20]. He uses kernel fuzzy C-means clustering method to group terms with the same sense into a set. Each set is mapped to a sense. Ahmad proposes the optimized gloss vector relatedness, the adapted gloss vector similarity measures, two enhanced semantic measures [21]. The effectiveness over WSD in biomedical domain is evaluated. Cao proposes an enhanced deep clustering network, which is composed of feature extractor, conditional generator, discriminator and siamese network [22]. The obtained pseudo-labels will be used to generate realistic data by generator. Finally, discriminator is used to model real joint distribution of data and corresponding latent representations for feature extractor enhancement. Ren proposes an abbreviation disambiguation method based on convolutional neural network to solve abbreviation disambiguation problem in biomedical field when no labelled corpus exists [23]. The data of the unsupervised method is unlabeled. We can obtain data easily, but WSD accuracy is not high.

In knowledge-based WSD method, lexical resources are applied including machine-readable dictionaries, thesauri and ontologies. Mohammed presents a simple modified version of SenseRelate algorithm for biomedical WSD, which ignores the distance that terms in contexts have the same distance [24]. Antunes applies results from machine learning and knowledge-based algorithms to biomedical WSD [25]. He represents textual definitions of biomedical concepts from the UMLS as word embeddings, and combines them with concept associations from the MeSH term co-occurrences. Sabbir exploits recent advances in neural word/concept embeddings to improve the performance of biomedical WSD on MSH dataset [26]. Duque gives a biomedical WSD

system based on co-occurrence graphs which contain biomedical concepts and textual information [27]. Rais exploits semantic similarity and relatedness measures from biomedical resources to evaluate the influence of context window size on WSD [28]. Pashuk disambiguates biomedical terms based on word bags from the context, definitions and information on related terms from the UMLS [29]. McInnes uses semantic similarity and relatedness measures to determine semantic category of biomedical term, which does not require human-annotated corpus and yields high accuracy [30]. Garla gives a knowledge-based WSD method that uses semantic similarity from the UMLS and evaluates the contribution of WSD to clinical text classification [31]. Kim suggests the link topic model inspired by latent Dirichlet allocation model, in which each document is perceived as a random mixture of topics, where each topic is characterized by a distribution over words [32]. Knowledge-based methods can mine large-scale data and organize useful information. But, knowledge is more difficult to be obtained.

Kang applies graph attention network to learn heterogeneous information [33]. Wang develops a GAT-based scheduler to learn features of scheduling problems automatically [34]. Wang introduces graph attention network based on syntactic dependency graph into natural language processing tasks [35]. Xie adopts attention architecture to learn representations of single views and uses regularization term to constrain the network's parameters [36]. Long designs graph convolutional network with node-level attention to learn embeddings for microbes and drugs [37].

These 3 methods have their own shortcomings. Although supervised WSD method can achieve the better performance, it needs a lot of annotated corpus. It is time-consuming and laborious. Unsupervised WSD method does not label corpus manually. But, disambiguation accuracy is not high. In knowledge-based WSD method, linguistic resources are used to provide disambiguation information for WSD. But, it is expensive to construct dictionaries. MSH dataset is annotated corpus. Supervised methods are usually more accurate than unsupervised ones. We choose the supervised method for WSD. Previous supervised WSD methods do not use linguistic knowledge such as parts of speech and semantic categories. Semantic category of ambiguous word is closely related to linguistic knowledge of its context in biomedical text. When words, parts of speech and semantic categories from ambiguous word's context are combined to determine its meanings, more discriminative information will be provided. WSD accuracy will be improved. In this paper, we use 2-layer GAT to extract discriminative features from context of biomedical ambiguous word. Softmax function is adopted to determine its semantic category.

III. WSD FEATURE EXTRACTION

Biomedical ambiguous word m has n semantic categories s_1, s_2, …, s_n. Words in context of m are annotated with parts of speech and semantic categories. Words, parts of speech, semantic categories and sentence are used as disambiguation features. For sentence containing ambiguous word 'Milk', the process of extracting disambiguation features is shown as follows:

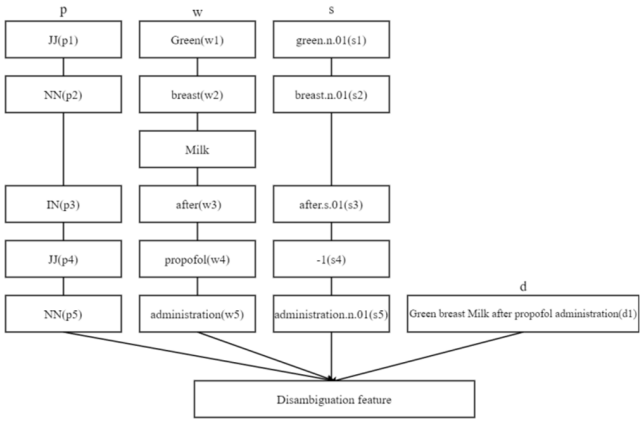


FIGURE 1. Feature extraction.

English sentence: Green breast Milk after propofol administration

Part-of-speech tagging: Green/JJ breast/NN Milk/NN after/IN propofol/JJ administration/NN

Semantic annotation: Green/JJ/green.n.01 breast/NN/breast.n.01 Milk/NN/milk.n.01 after/IN/after.s.01 propofol/JJ/-1 administration/NN/administration.n.01

Here, word is represented by w, part of speech is denoted as p, semantic category is represented by s, and sentence is denoted as d. We extract 16 disambiguation features: p1 = JJ, p2 = NN, p3 = IN, p4 = JJ, p5 = NN, w1 = Green, w2 = breast, w3 = after, w4 = propofol, w5 = administration, s1 = green.n.01, s2 = breast.n.01, s3 = after.s.01, s4 = −1, s5 = administration.n.01, d = Green breast Milk after propofol administration. Use Word2Vec to vectorize word, part of speech, and semantic category. Disambiguation feature x is vectorized as Word2Vec(x). Use Doc2Vec to vectorize sentence d as Doc2Vec(d). Then, 16 disambiguation features are gotten including Word2Vec(p1), …, Word2Vec(p5), Word2Vec(w1), …, Word2Vec(w5), …, Word2Vec(s1), …, Word2Vec(s5), Doc2Vec(d). The process of extracting disambiguation features is shown in Figure 1.

IV. WORD SENSE DISAMBIGUATION BASED ON GAT

Graph is composed of nodes and edges. In graph attention neural network, attention mechanism is introduced into graph neural network to measure the importance of nodes.

WSD graph is constructed, in which words, parts of speech, semantic categories and sentence are used as nodes, and their relationships are used as edges between nodes. WSD graph are composed of word set W{w1, w2, w3, …}, part of speech set P{p1, p2, p3, …}, and semantic category set S{s1, s2, s3,…}, and sentence set D{d1, d2, d3,…}. At the same time, WSD graph contains edge set WP{wp1, wp2, wp3, …} between word and part of speech, edge set WS{ws1, ws2, ws3,…} between word and semantic category, edge

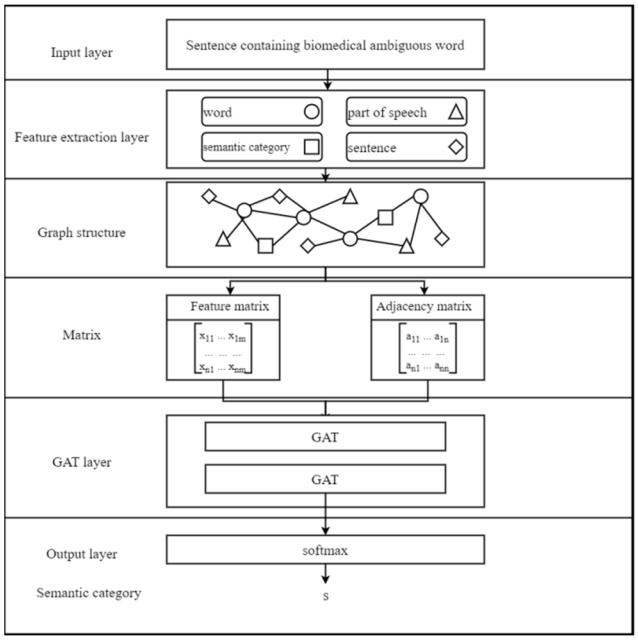


FIGURE 2. Biomedical WSD based on GAT.

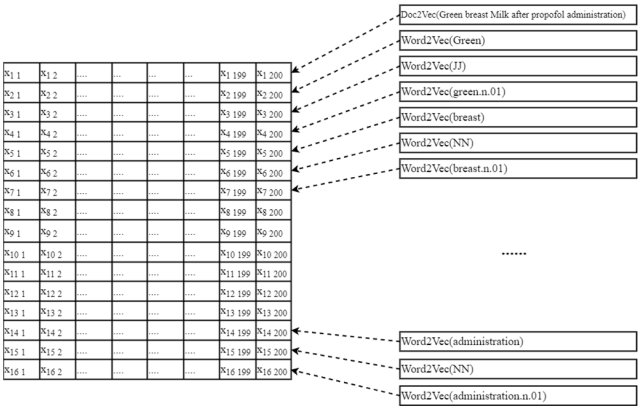


FIGURE 3. Feature matrix X_{16*200}.

set WD{wd1, wd2, wd3, ...} between word and sentence, edge set WW{ww1, ww2, ww3,...} between word and word. Adjacency matrix A is constructed based on WSD graph. When the number of nodes is N, the size of matrix A is N*N. If the dimension of feature vector is M, the scale of feature matrix X is N*M. Adjacency matrix A and feature matrix X are input into GAT to extract discriminative feature, and softmax function is used to determine semantic category of biomedical ambiguous word as shown in Figure 2.

For the above sentence containing ambiguous word 'Milk', 16 disambiguation features are extracted. Use Word2Vec tool and Doc2Vec tool to vectorize disambiguation features. Feature matrix X_{16*200} is gotten as shown in Figure 3 and input into GAT.

We use TF-IDF to determine whether there is edge between word node and sentence node, as shown in formula (1).

$$\mathrm{TF}-\mathrm{IDF}=\mathrm{TF}\left(\mathrm{t_i},\mathrm{d_j}\right)\times\lg\left(\frac{\mathrm{N}}{\mathrm{n_i}}+0.01\right) \tag{1}$$

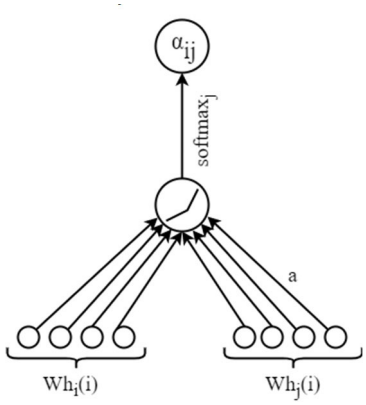


FIGURE 4. Attention weight α_{ij}.

where, $\mathrm{TF}(\mathrm{t_i},\mathrm{d_j})$ represents the frequency of word $\mathrm{t_i}$ appearing in sentence $\mathrm{d_j}$, N is the number of sentences in document, and $\mathrm{n_i}$ represents the number of sentences containing word $\mathrm{t_i}$ in document.

Use PMI to determine whether there are edges between word nodes and word nodes as shown in formula (2).

$$\mathrm{PMI}\left(\mathrm{w_i},\mathrm{w_j}\right)=\lg\frac{\mathrm{p}\left(\mathrm{w_i},\mathrm{w_j}\right)}{\mathrm{p}\left(\mathrm{w_i}\right)\mathrm{p}\left(\mathrm{w_j}\right)} \tag{2}$$

where, $\mathrm{p}(\mathrm{w_i},\mathrm{w_j})$ represents co-occurrence probability of $\mathrm{w_i}$ and $\mathrm{w_j}$, $\mathrm{p}(\mathrm{w_i})$ is occurrence probability of $\mathrm{w_i}$, $\mathrm{p}(\mathrm{w_j})$ denotes occurrence probability of $\mathrm{w_j}$. Each node is regarded as its own neighbor to retain its own information. A closed loop is added into adjacency matrix A, and its diagonal elements are set to 1. Adjacency matrix A is shown in formula (3).

$$\mathrm{A_{ij}}=\begin{cases}1 & \mathrm{w_i}\in\mathrm{W},\mathrm{w_j}\in\mathrm{W},\ \mathrm{PMI}(\mathrm{w_i},\mathrm{w_j})>0\\ 1 & \mathrm{w_i}\in\mathrm{W},\mathrm{d_j}\in\mathrm{D},\mathrm{W}\in\mathrm{D},\mathrm{TF\text{-}IDF}(\mathrm{w_i},\mathrm{d_j})>0\\ 1 & \mathrm{w_i}\in\mathrm{W},\mathrm{p_j}\in\mathrm{P}\\ 1 & \mathrm{w_i}\in\mathrm{W},\mathrm{s_j}\in\mathrm{S}\\ 1 & \mathrm{i}=\mathrm{j}\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

Assuming that $\mathrm{v_j}$ is neighbor node of $\mathrm{v_i}$, attention weight α_{ij} between $\mathrm{v_i}$ and $\mathrm{v_j}$ is computed as shown in formula (4).

$$\begin{aligned}\alpha_{ij}&=\mathrm{soft\,max_j}\left(\mathrm{Leaky\,ReLU}\left(\mathrm{a^T}\left[\mathrm{Wh_i(i)}\|\mathrm{Wh_j(i)}\right]\right)\right)\\&=\frac{\exp\left(\mathrm{LeakyReLU}\left(\mathrm{a^T}\left[\mathrm{Wh_i(i)}\|\mathrm{Wh_j(i)}\right]\right)\right)}{\sum_{\mathrm{k}\in\mathrm{N}}\exp\left(\mathrm{Leaky\,ReLU}\left(\mathrm{a^T}\left[\mathrm{Wh_i(i)}\|\mathrm{Wh_k(i)}\right]\right)\right)}\end{aligned} \tag{4}$$

where, W is weight, $\mathrm{h_i(i)}$ and $\mathrm{h_j(i)}$ are respectively feature vector of node $\mathrm{v_i}$ and $\mathrm{v_j}$ in the ith layer, N_i is the set containing all adjacent nodes of $\mathrm{v_i}$, || represents the splicing operation, LeakyReLU is activation function, a is determined by weight vector.

Attention weight α_{ij} is computed as shown in Figure 4.

We use K-head attention to stabilize self-attention learning process. Here, $\mathrm{h_i(i+1)}$ is feature vector of $\mathrm{v_i}$ in the i+1th layer

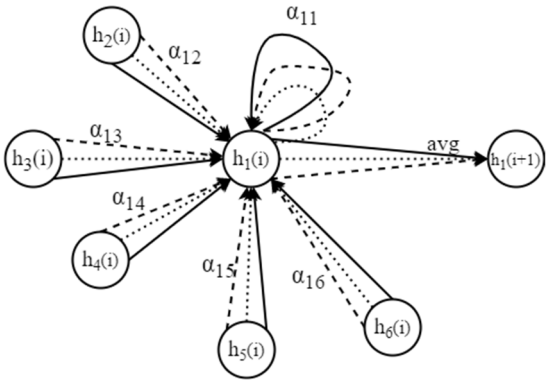


FIGURE 5. Compute $h_1(i+1)$ with 3 head attentions.

as shown in formula (5).

$$h_i(i+1) = \sigma\left(\frac{1}{K}\sum_{K=1}^{K}\sum_{j\in N_i}\alpha_{ij}^{k}W^{k}h_j(i)\right) \tag{5}$$

where, σ represents nonlinear activation function.

The process of computing $h_1(i+1)$ under 3 head attentions is shown in Figure 5.

We extract disambiguation features from sentence containing biomedical ambiguous word m. WSD graph is constructed. Feature matrix and adjacency matrix are constructed. They are input into GAT layer to extract discriminative features. Softmax function is used to calculate probability $p(s_i|m)$ of m under semantic category s_i. Then, semantic category s of ambiguous word m is determined as shown in formula (6).

$$s = \arg\max_{i=1,2,\ldots,n} P(s_i|m) \tag{6}$$

V. EXPERIMENTAL RESULTS AND ANALYSIS

MSH data set in biomedical field is used to train and testify the proposed method. The US National Library of Medicine has developed a medical language system. The data generated by this system is integrated as MSH dataset. MSH dataset consists of 203 ambiguous words, including 106 ambiguous abbreviations, 88 ambiguous terms and 9 words which are a combination of both. MSH dataset is used as training corpus and test corpus to testify the proposed method. 28 acronyms are selected from MSH dataset, which have a lot of sentences, are representative, have classification significance and are often used. There are 18 abbreviations with two semantic categories including ADA, ALS, BLM, Cement, Cilia, DI, drinking, EMS, Fish, HHV8, IP, JP, MBP, Milk, Moles, Nurse, Root and Wasp. There are 10 abbreviations with 3 semantic categories. They are respectively Cold, Cortical, CP, DDS, Lens, Lupus, PCP, RA, TAT and THYMUS. 5 groups of experiments are carried out. The first group of experiments are performed to compare GCN-based WSD method, CNN-based WSD method and the proposed method. The second group of experiments are conducted to testify the influence of head number on WSD. The third group of experiments are performed to investigate the impact of corpus size on WSD. The fourth group of experiments are conducted to testify the influence of attention layer number on WSD. In the fifth group of experiments, majority voting strategy is adopted to optimize GAT-based WSD classifier in which CNN-based classifier and GCN-based classifier are used.

Average accuracy is used to evaluate WSD classifier as shown in formula (7).

$$p_i = \frac{m_i}{n_i}, \quad p_{avg} = \frac{\sum_{i=1}^{N} p_i}{N} \tag{7}$$

where, N is the number of ambiguous words, m_i is the number of test sentences correctly classified for the ith ambiguous word, n_i is the number of test sentences containing the ith ambiguous word, p_i is disambiguation accuracy of the ith ambiguous word, p_{avg} is average accuracy.

The first group of experiments include Experiment 1, Experiment 2, and Experiment 3. In these 3 experiments, the learning rate is 0.01, the dropout rate is 0.5, and the number of training epochs is 100. In Experiment 1 and Experiment 2, words, parts of speech and semantic categories are extracted from contexts of ambiguous word as disambiguation features. Disambiguation features and sentence containing ambiguous word are used as nodes to construct WSD graph. GAT and GCN are respectively used to determine semantic category of ambiguous word on WSD graph. In Experiment 3, words, parts of speech and semantic categories are extracted as disambiguation features from two left and right units around ambiguous word. CNN is used to determine semantic category of ambiguous word. Activation function of GAT, GCN and CNN layer is Relu. Softmax layer is adopted to determine semantic category of ambiguous word. Training corpus is used to optimize GAT, GCN, and CNN. Test corpus is adopted to evaluate the optimized GAT, GCN, and CNN as shown in Table 1.

It can be seen from Table 1 that average accuracy of Experiment 1 is the best and achieves better than Experiment 2. GCN and GAT aggregate neighbor nodes' features to the center one. But, GCN uses Laplacian matrix and GAT uses attention coefficient. GAT can extract more effective features than GCN. At the same time, the correlation between nodes is better integrated into WSD model. Experiment 2 achieves better than Experiment 3 at average accuracy. This is because that disambiguation features are extracted from all left and right units of ambiguous word in Experiment 2. But, disambiguation features of Experiment 3 are extracted from two left and right units around ambiguous word. More linguistic knowledge is integrated into WSD classifier in Experiment 2. So, average accuracy of Experiment 2 is better than that of Experiment 3.

Ambiguous words of Experiment 1, Experiment 2 and Experiment 3 in Table 1 are respectively classified according to category number. Average accuracy of ambiguous words with the same category number is calculated as shown in Figure 6.

TABLE 1. Disambiguation accuracies in the first group of experiments.

Ambiguous word	Exp 1	Exp 2	Exp 3
ADA	0.9300	0.9670	0.8670
ALS	0.9210	0.8950	0.8950
BLM	1.0000	0.9830	0.7860
Cement	0.8800	0.8100	0.6760
Cilia	0.9140	0.8290	0.8910
Cold	0.7620	0.4280	0.5650
Cortical	0.5600	0.6200	0.7730
CP	0.9160	0.8920	0.9340
DDS	0.6800	0.5460	0.8330
DI	0.8970	0.6410	0.9400
drinking	0.7950	0.8980	0.9000
EMS	0.9700	1.0000	0.8270
Fish	0.8050	0.8320	0.7770
HHV8	0.7100	0.6700	0.5000
IP	0.9620	0.8540	0.8890
JP	0.5650	0.8700	0.8200
Lens	0.8960	0.7640	0.6500
Lupus	0.8930	0.7590	0.8240
MBP	0.8240	0.8500	0.7180
Milk	0.8950	0.8330	0.8620
Moles	0.8570	0.8210	0.7500
Nurse	0.8160	0.8500	0.7830
PCP	0.9160	0.8540	0.7560
RA	0.8420	0.7860	0.8080
Root	0.7000	0.7160	0.7500
TAT	0.8570	0.8250	0.7930
THYMUS	0.8850	0.8360	0.9010
Wasp	0.6500	0.5830	0.6660
Average accuracy	0.8320	0.7930	0.7900

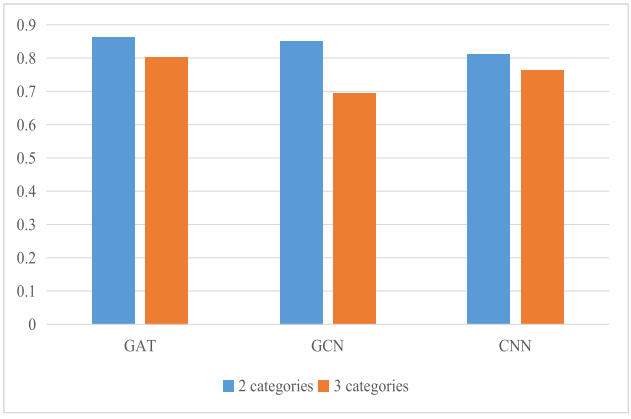


FIGURE 6. Average accuracy under different WSD classifiers and category number.

From Figure 6, it can be seen that average accuracy of WSD classifier decreases when category number increases. The reason is that the predicted results have more possibilities with category number increasing. It makes error rate of WSD classifier higher. Average accuracy of GAT is higher than that of GCN for 2 categories and 3 ones. This is because that GAT has better ability of feature extraction than GCN. GAT assigns different weights to neighbor nodes with the same order. Information from neighbor nodes is aggregated and scaled according to attention weight. The correlation between nodes is better integrated into WSD classifier. GCN assigns the same weights to neighbor nodes with the same order. So, the way that GCN fuses adjacent nodes' features is related to graph structure. GCN WSD classifier has poor ability of being generalized to graphs with different structure. Average accuracy of GAT is higher than that of CNN for 2 categories and 3 ones. The reason is that GAT extracts disambiguation features from context of ambiguous word. But, CNN only extracts disambiguation features from 2 left and right units of ambiguous word. Average accuracy of GCN is higher than that of CCN for 2 categories. This is because that information of all units is used in GCN. But, information of 2 left and right units is adopted in CNN. Average accuracy of GCN is lower than that of CCN for 3 categories. The reason is that accuracy of GCN is considerably lower than that of CCN for some biomedical ambiguous words. For example, DDS and DI.

TABLE 2. The influence of head number on WSD.

Ambiguous word	3 heads	4 heads	5 heads	6 heads
ADA	0.9300	0.9300	0.9300	1
ALS	0.9210	0.9210	0.9210	0.8940
BLM	0.8660	1.0000	1.0000	0.930
Cement	0.8090	0.8570	0.8800	0.8330
Cilia	0.8850	0.8280	0.9140	0.9140
Cold	0.7210	0.7340	0.7620	0.7530
Cortical	0.6800	0.8200	0.5600	0.4390
CP	0.6380	0.8050	0.9160	0.7770
DDS	0.4310	0.4500	0.6800	0.5680
DI	0.8460	0.7690	0.8970	0.820
drinking	0.9800	0.7950	0.7950	0.8770
EMS	0.9800	0.9580	0.9700	0.8530
Fish	0.8610	0.8900	0.8050	0.8610
HHV8	0.6880	0.6400	0.7100	0.7520
IP	0.9430	0.9430	0.9620	0.9620
JP	0.4340	0.3040	0.5650	0.4780
Lens	0.9650	0.9650	0.8960	0.8620
Lupus	0.5750	0.8930	0.8930	0.9540
MBP	0.7750	0.8240	0.8240	0.7000
Milk	0.9580	0.8540	0.8950	0.9370
Moles	0.7500	0.8210	0.8570	0.8920
Nurse	0.7830	0.8330	0.8160	0.7830
PCP	0.9370	0.7910	0.9160	0.9790
RA	0.8420	0.8980	0.8420	0.8200
Root	0.6160	0.6500	0.7000	0.6830
TAT	0.7930	0.8250	0.8570	0.8730
THYMUS	0.7860	0.8190	0.8850	0.8520
Wasp	0.6300	0.6160	0.6500	0.7000
Average accuracy	0.7860	0.8010	0.8320	0.8120

Attention head number affects the performance of the proposed network. The second group of experiments are performed to investigate the influence of head number on biomedical WSD. In these 4 experiments, the learning rate is 0.01, the dropout rate is 0.5, and the number of training epochs is 100. Activation function of GAT layer is Relu and softmax layer is adopted to determine semantic category of ambiguous word. Head number is respectively set to 3, 4, 5 and 6. Training corpus is used to optimize the proposed network. Test corpus is adopted to evaluate the optimized network as shown in Table 2.

It can be seen from Table 2 that average accuracy of the proposed network first increases and then decreases with

TABLE 3. Discrimination accuracies of the proposed network under different ratio.

Ambiguous word	0.75	0.7	0.65	0.6
ADA	0.9300	0.9300	0.8420	0.8740
ALS	0.9670	0.9210	0.8830	1.0000
BLM	0.9180	1.0000	0.9420	0.8610
Cement	0.7350	0.8800	0.7820	0.7510
Cilia	0.9650	0.9140	0.8730	0.8360
Cold	0.7620	0.7620	0.8250	0.7780
Cortical	0.8810	0.5600	0.7110	0.7910
CP	0.6370	0.9160	0.7640	0.6830
DDS	0.7510	0.6800	0.6030	0.7720
DI	0.8750	0.8970	0.8550	0.7680
drinking	0.7530	0.7950	0.8590	0.9300
EMS	0.8500	0.9700	0.8130	0.8280
Fish	0.9000	0.8050	0.9260	0.8900
HHV8	0.7720	0.7100	0.7540	0.6620
IP	0.8530	0.9620	0.8850	0.9570
JP	0.6310	0.5650	0.7580	0.4260
Lens	0.8570	0.8960	0.8310	0.7200
Lupus	0.6570	0.8930	0.8420	0.7190
MBP	0.7150	0.8240	0.7740	0.7620
Milk	0.9250	0.8950	0.9810	0.9680
Moles	0.7820	0.8570	0.8120	0.8050
Nurse	0.8570	0.8160	0.8400	0.8730
PCP	0.8460	0.9160	0.8920	0.9300
RA	0.7800	0.8420	0.8050	0.7620
Root	0.6120	0.7000	0.6810	0.7340
TAT	0.8070	0.8570	0.8350	0.8190
THYMUS	0.9200	0.8850	0.8710	0.8620
Wasp	0.7140	0.6500	0.7390	0.7590
average accuracy	0.8090	0.8320	0.8200	0.8040

TABLE 4. The influence of layer number on WSD.

Ambiguous word	1 layer	2 layers	3 layers	4 layers
ADA	0.8740	0.9300	0.8460	0.8460
ALS	0.7470	0.9210	0.6970	0.5960
BLM	0.8650	1.0000	0.9850	0.9850
Cement	0.7340	0.8800	0.7550	0.8600
Cilia	0.9140	0.9140	0.8780	0.8420
Cold	0.6840	0.7620	0.6420	0.6220
Cortical	0.6380	0.5600	0.6190	0.7000
CP	0.8550	0.9160	0.7740	0.7360
DDS	0.6210	0.6800	0.5780	0.7020
DI	0.7350	0.8970	0.7530	0.6990
drinking	0.9510	0.7950	0.8070	0.8790
EMS	0.7480	0.9700	0.9370	0.8740
Fish	0.6970	0.8050	0.8420	0.6430
HHV8	0.9210	0.7100	0.9210	0.8570
IP	0.9830	0.9620	0.8990	0.8360
JP	0.6430	0.5650	0.5130	0.6950
Lens	0.8350	0.8960	0.7970	0.8350
Lupus	0.8130	0.8930	0.8510	0.8130
MBP	0.8720	0.8240	0.7520	0.8240
Milk	0.9450	0.8950	0.9200	0.9740
Moles	0.8210	0.8570	0.7500	0.8210
Nurse	0.8330	0.8160	0.7830	0.7330
PCP	0.9610	0.9160	0.8950	0.9390
RA	0.8080	0.8420	0.8310	0.7750
Root	0.6160	0.7000	0.6500	0.6830
TAT	0.7930	0.8570	0.8250	0.8570
THYMUS	0.8190	0.8850	0.8850	0.8520
Wasp	0.7160	0.6500	0.6160	0.6500
average accuracy	0.8010	0.8320	0.7850	0.7900

the increase of head number. The proposed network with 5 head attentions achieves the best and its average accuracy reaches 0.8424. Different attentions consider the relevance in different levels and calculate independently. When head number is larger, information in more levels can be considered to obtain effective features. When head number is smaller, the dimension of feature vector for each attention is larger. The network's structure is complicated, and the overfitting phenomenon occurs easily. Disambiguation feature is divided into 5 parts, and 5 attentions calculate independently. Then, they are concatenated to obtain effective discriminative features.

The scale of training corpus influences the performance of the proposed network. The third group of experiments are performed where the ratio of training corpus and test one is respectively set to 4:3, 7:3, 20:13, 5:3. In these 4 experiments, the learning rate is 0.01, the dropout rate is 0.5, and the number of training epochs is 100. Activation function of GAT layer is Relu and softmax layer is adopted to determine semantic category of ambiguous word. Head number of the proposed network is set to 5. Training corpus is used to optimize the proposed network. Test corpus is adopted to evaluate the optimized network as shown in Table 3.

It can be seen from Table 3 that average accuracy of the proposed network first increases and then decreases with the scale of training corpus increasing. The proposed network is optimized adequately when the ratio of training corpus and test one is 7:3. It achieves the best and its average accuracy reaches 0.8424. The reason is that the proposed network is optimized adequately and it can extract more discriminative features. So, WSD classifier performs better. When the ratio is less than 7:3, there are less training data and GAT is optimized inadequately. When the ratio is larger than 7:3, there are more training data and more noise is introduced into the process of optimizing GAT.

Layer number of GAT influences the performance of the proposed network. The fourth group of experiments are conducted where layer number is respectively set to 1, 2, 3, and 4 respectively. Head number of the proposed network is set to 5. The learning rate is 0.01, the dropout rate is 0.5, and the number of training epochs is 100. Activation function of GAT layer is Relu and softmax layer is adopted to determine semantic category of ambiguous word. Training corpus is used to optimize the proposed network. Test corpus is adopted to evaluate the optimized network as shown in Table 4.

It can be seen from Table 4 that the proposed network achieves the best, whose layer number is 2. Its average accuracy reaches 0.8424. This is because that information between GAT nodes cannot be fused well when layer number is too small. When there are more GAT layers, extensive information will be collected. But, when layer number is too high, information will be diffused excessively in graph's nodes. Each node's representation is smoothed and its discriminative ability decreases. Information in high-order neighbor nodes are fused, which results in excessive fusion of information and reduces the performance of biomedical WSD classifier.

Ambiguous words in Table 4 are respectively classified according to category number and layer number. Average accuracy of ambiguous words with the same category number and layer number is calculated as shown in Figure 7.

TABLE 5. The influence of majority voting strategy on GAT-based WSD classifier.

Ambiguous word	GAT	Majority voting
ADA	0.9300	0.9160
ALS	0.9210	0.9480
BLM	1.0000	1.0000
Cement	0.8800	0.8570
Cilia	0.9140	0.9420
Cold	0.7620	0.7340
Cortical	0.5600	0.6800
CP	0.9160	0.9440
DDS	0.6800	0.7640
DI	0.8970	0.8460
drinking	0.7950	0.8160
EMS	0.9700	0.9370
Fish	0.8050	0.8330
HHV8	0.7100	0.8220
IP	0.9620	0.9240
JP	0.5650	0.6520
Lens	0.8960	0.8270
Lupus	0.8930	0.9240
MBP	0.8240	0.8240
Milk	0.8950	0.8750
Moles	0.8570	0.8920
Nurse	0.8160	0.8660
PCP	0.9160	0.9370
RA	0.8420	0.8310
Root	0.7000	0.7830
TAT	0.8570	0.9200
THYMUS	0.8850	0.8360
Wasp	0.6500	0.7160
average accuracy	0.8320	0.8510

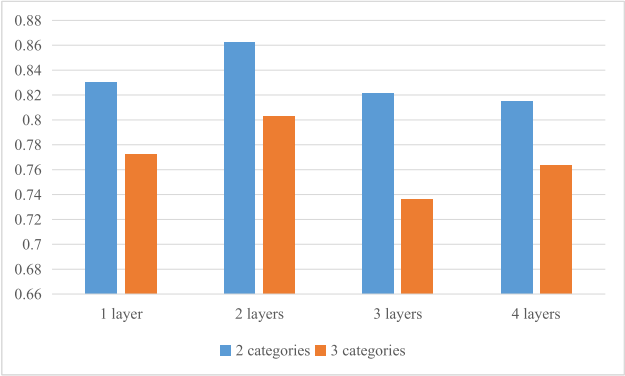


FIGURE 7. Average accuracy under different layer number and category number.

From Figure 7, it can be seen that average accuracy of the proposed network decreases with layer number increasing. The is because that the predicted results have more possibilities when layer number increases. It makes error rate of the proposed network higher. Its average accuracy first increases and then decreases with the increase of layer number. The proposed network with 2 layers achieves the best for 2 categories and 3 ones. If layer number is too small, information between nodes can not be well fused. If there is more GAT layers, information in high-order nodes is fused and effective features can not be extracted.

In the fifth group of experiments, the optimized CNN-based classifier, GCN-based classifier and GAT-based classifier are respectively used to determine semantic categories of ambiguous word m in test sentence. Then, 3 semantic categories can be gotten. We select category with the highest frequency as semantic category of m. Disambiguation accuracies of ambiguous words are shown in Table 5.

From Table 5, we can see that WSD method based on majority voting strategy achieves better than GAT-based WSD method on average accuracy. Its average accuracy reaches 0.8510. The reason is that CNN, GCN and GAT have their own advantages and disadvantages on feature extraction. When CNN, GCN are used to help GAT for determining semantic category of ambiguous word under majority voting strategy, average accuracy is improved.

VI. CONCLUSION AND FUTURE WORKS

In this paper, a biomedical WSD network is proposed, including input layer, feature extraction layer, graph construction layer, GAT layer and output layer. Words, parts of speech and semantic categories are extracted as disambiguation features from context of biomedical ambiguous word. WSD graph is constructed, in which disambiguation features and sentence are used as nodes. Relationships between word and sentence, word and part of speech, word and semantic category are viewed as edges. Then, adjacency matrix and feature matrix are built. GAT is used to extract discriminative features and softmax function is adopted to determine semantic category of biomedical ambiguous word. Experiments are conducted on MSH dataset and results show that average accuracy of the proposed method reaches 0.8424. When majority voting strategy is adopted to optimize GAT-based WSD classifier, its average accuracy is improved.

In the future, more linguistic and biomedicine knowledge will be introduced into biomedical WSD. For example, parsing information of context and common knowledge in biomedicine field. At the same time, we apply GAT to WSD graph for disambiguating biomedical ambiguous word. So, WSD graph is key to the proposed network in this paper. When adjacency matrix is constructed, PMI is adopted to evaluate the relevance of two words. It is not precise. In the future, we will try more methods to compute the relevance of two words.

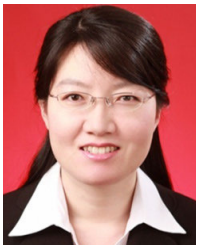
REFERENCES

[1] A. M. Hisham and G. Sandeep, "A learning approach for word sense disambiguation in the biomedical domain," in *Proc. 3rd Int. Conf. Bioinf. Comput. Biol.*, New Orleans, LO, USA, Mar. 2011, pp. 104–109.
[2] M. Rais and A. Lachkar, "Evaluation of disambiguation strategieson biomedical text categorization," in *Proc. 4th Int. Work-Conf. Bioinf. Biomed. Eng.* Grenada, Caribbean, Apr. 2016, pp. 790–801.
[3] J. Preiss and M. Stevenson, "The effect of word sense disambiguation accuracy on literature based discovery," in *Proc. ACM 9th Int. Workshop Data Text Mining Biomed. Informat.*, Melbourne, VIC, Australia, Oct. 2015, p. 1.
[4] B. T. McInnes and M. Stevenson, "Determining the difficulty of word sense disambiguation," *J. Biomed. Informat.*, vol. 47, pp. 83–90, Feb. 2014.
[5] Y. Wang, K. Zheng, H. Xu, and Q. Mei, "Interactive medical word sense disambiguation through informed learning," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 7, pp. 800–808, Jul. 2018.
[6] C. Zhang, D. Bis, X. Liu, and Z. He, "Biomedical word sense disambiguation with bidirectional long short-term memory and attention-based neural networks," *BMC Bioinf.*, vol. 20, no. S16, pp. 1–15, Dec. 2019.

[7] A. J. Yepes, "Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation," *J. Biomed. Inform.*, vol. 73, pp. 137–147, Sep. 2017.
[8] S. Festag and C. Spreckelsen, "Word sense disambiguation of medical terms via recurrent convolutional neural networks," in *Proc. 11th Annu. Conf. Health Informat. Meets eHealth*, May 2017, pp. 8–15.
[9] R. Antunes and S. Matos, "Supervised learning and knowledge-based approaches applied to biomedical word sense disambiguation," *J. Integrative Bioinf.*, vol. 14, no. 4, Dec. 2017.
[10] D. Bis, C. Zhang, X. Liu, and Z. He, "Layered multistep bidirectional long short-term memory networks for biomedical word sense disambiguation," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 313–320.
[11] A. K.-F. Lui, Y.-H. Chan, and M.-F. Leung, "Modelling of destinations for data-driven pedestrian trajectory prediction in public buildings," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Orlando, FL, USA, Dec. 2021, pp. 1709–1717.
[12] Q. Jin, J. Liu, and X. Lu, "Deep contextualized biomedical abbreviation expansion," in *Proc. 18th BioNLP Workshop Shared Task*, Florence, Italy, 2019, pp. 88–96.
[13] A. Pesaranghader, S. Matwin, M. Sokolova, and A. Pesaranghader, "DeepBioWSD: Effective deep neural word sense disambiguation of biomedical text data," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 5, pp. 438–446, May 2019.
[14] N. Smalheiser and G. Bonifield, "Two similarity metrics for medical subject headings (MeSH): An aid to biomedical text mining and author name disambiguation," *J. Biomed. Discovery Collaboration*, vol. 7, Apr. 2016.
[15] Z. Li, F. Yang, and Y. Luo, "Context embedding based on bi-LSTM in semi-supervised biomedical word sense disambiguation," *IEEE Access*, vol. 7, pp. 72928–72935, 2019.
[16] A. Duque, M. Stevenson, J. Martinez-Romo, and L. Araujo, "Co-occurrence graphs for word sense disambiguation in the biomedical domain," *Artif. Intell. Med.*, vol. 87, pp. 9–19, May 2018.
[17] K. Ren and S. W. Wang, "Improved convolutional neural network for biomedical word sense disambiguation with enhanced context feature modeling," *J. Digit. Inf. Manag.*, vol. 14, no. 6, pp. 342–350, 2016.
[18] W. G. El-Rab, O. R. Zaiane, and M. El-Hajj, "Unsupervised graph-based word sense disambiguation of biomedical documents," in *Proc. IEEE 15th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Lisbon, Portugal, Oct. 2013, pp. 649–652.
[19] S. Moon, B. McInnes, and G. B. Melton, "Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain," *Healthcare Informat. Res.*, vol. 21, no. 1, pp. 35–42, 2015.
[20] K. Ren and Y. F. Ren, "Kernel fuzzy C-means clustering for word sense disambiguation in BioMedical texts," *J. Digit. Inf. Manag.*, vol. 13, no. 6, pp. 411–420, 2015.
[21] A. Pesaranghader, A. Pesaranghader, and N. Mustapha, "Word sense disambiguation for biomedical text mining using definition-based semantic relatedness and similarity measures," *Int. J. Bioscience, Biochemistry Bioinf.*, vol. 4, no. 4, pp. 280–283, 2014.
[22] W. Cao, Z. Zhang, C. Liu, R. Li, Q. Jiao, Z. Yu, and H.-S. Wong, "Unsupervised discriminative feature learning via finding a clustering-friendly embedding space," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108768.
[23] K. Ren and S. W. Wang, "Applying convolutional neural network model and auto-expanded corpus to biomedical abbreviation disambiguation," *J. Eng. Sci. Technol. Rev.*, vol. 9, pp. 178–184, Dec. 2016.
[24] R. Mohammed and L. Abdelmonaime, "An empirical study of word sense disambiguation for biomedical information retrieval system," in *Proc. 6th Int. Work-Conf. Bioinf. Biomed. Eng.*, Granada, Spain, Apr. 2018, pp. 314–326.
[25] R. Antunes and S. Matos, "Biomedical word sense disambiguation with word embeddings," in *Proc. Adv. Intell. Syst. Comput.*, vol. 616, 2017, pp. 273–279.
[26] A. Sabbir, A. Jimeno-Yepes, and R. Kavuluru, "Knowledge-based biomedical word sense disambiguation with neural concept embeddings," in *Proc. IEEE 17th Int. Conf. Bioinf. Bioengineering (BIBE)*, Oct. 2017, pp. 163–170.
[27] A. Duque, J. Martinez-Romo, and L. Araujo, "Can multilinguality improve biomedical word sense disambiguation?" *J. Biomed. Informat.*, vol. 64, pp. 320–332, Dec. 2016.
[28] M. Rais and A. Lachkar, "Biomedical word sense disambiguation context-based: Improvement of SenseRelate method," in *Proc. Int. Conf. Inf. Technol. for Organizations Develop. (IT OD)*, Mar. 2016.
[29] A. V. Pashuk, A. B. Gurinovich, N. A. Volorova, and A. P. Kuznetsov, "Analysis of the methods of word sense disambiguation in the biomedical domain," *Doklady BGUIR*, no. 5, pp. 60–65, Jul. 2019.
[30] B. T. McInnes and T. Pedersen, "Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text," *J. Biomed. Informat.*, vol. 46, no. 6, pp. 1116–1124, Dec. 2013.
[31] V. N. Garla and C. Brandt, "Knowledge-based biomedical word sense disambiguation: An evaluation and application to clinical document classification," *J. Amer. Med. Informat. Assoc. Jamia*, vol. 20, no. 5, pp. 83–90, 2013.
[32] S. Z. Kang, L. X. Ji, and J. P. Zhang, "Heterogeneous information network representation learning framework based on graph attention network," *J. Electron. Inf. Technol.*, vol. 43, pp. 915–922, Sep. 2021.
[33] S. Kang, L. Ji, and J. Zhang, "Heterogeneous information network representation learning framework based on graph attention network," *J. Electron. Inf. Technol.*, vol. 43, no. 3, pp. 915–922, 2021.
[34] Z. Wang and M. Gombolay, "Learning scheduling policies for multi-robot coordination with graph attention networks," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4509–4516, Jul. 2020.
[35] Z. Y. Wang and M. Gombolay, "Causal relation extraction based on graph attention networks," *Comput. Res. Develop.*, vol. 57, no. 1, pp. 159–174, 2020.
[36] Y. Xie, Y. Zhang, M. Gong, Z. Tang, and C. Han, "MGAT: Multi-view graph attention networks," *Neural Netw.*, vol. 132, pp. 180–189, Dec. 2020.
[37] Y. Long, M. Wu, Y. Liu, C. K. Kwoh, J. Luo, and X. Li, "Ensembling graph attention networks for human microbe–drug association prediction," *Bioinformatics*, vol. 36, no. 2, pp. i779–i786, Dec. 2020.

CHUN-XIANG ZHANG received the Graduate and Ph.D. degrees from the MOE-MS Key Laboratory of Natural Language Processing and Speech, School of Computer Science and Technology, Harbin Institute of Technology, in 2007. He is currently a Professor at the School of Computer Science and Technology, Harbin University of Science and Technology. His research interests include natural language processing, machine translation, machine learning, computer graphics and CAD, and 3D model retrieval. He has authored or coauthored more than 60 journals and conference papers in these areas.

MING-LEI WANG received the B.S. degree from the Qilu University of Technology, in 2019. He is currently pursuing the master's degree with the School of Computer Science and Technology, Harbin University of Science and Technology. His research interests include natural language processing and word sense disambiguation.



XUE-YAO GAO received the Graduate and Ph.D. degrees from the School of Computer Science and Technology, Harbin University of Science and Technology, in 2009. She is currently a Professor at the School of Computer Science and Technology, Harbin University of Science and Technology. Her research interests include computer graphics and CAD, 3D model retrieval, natural language processing, and machine learning. She has authored or coauthored more than 50 journals and conference papers in these areas.