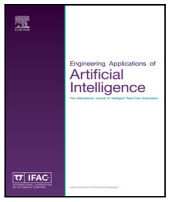
Engineering Applications of Artificial Intelligence 122 (2023) 106057

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Contrastive knowledge integrated graph neural networks for Chinese medical text classification

Ge Lan [a], Mengting Hu [a,*], Ye Li [b], Yuzhi Zhang [a]

[a] College of Software, Nankai University, Tianjin, 300350, China
[b] College of Artificial Intelligence, Nankai University, Tianjin, 300350, China

ARTICLE INFO

Keywords:
Knowledge graph
Graph neural networks
Medical text classification
Supervised contrastive learning

ABSTRACT

This paper aims at medical text classification, where texts describe medicines, diseases, or other medical topics. This field is still challenging since medical texts contain intensive specialization and terminology, which require professional semantic and structured knowledge to classify. Based on our observations, medical knowledge graph (KG) can provide such knowledge although they may be ambiguous. To this end, we propose contrastive knowledge integrated graph neural networks (ConKGNN) to make full use of the above knowledge. Specifically, the proposed method builds two graphs for a medical text, i.e. text graph and text-specific subgraph, containing the text information and relevant KG information, respectively. Two graphs are merged into a united graph, which is jointly modeled by graph neural networks (GNN). In this way, our approach adequately learns interactions between neighbors. Meanwhile, it promotes the mutual influences between text and KG. We further propose graph-based supervised contrastive learning. By randomly cutting off nodes from the text graph, an augmented united graph is obtained. Learning it in a contrastive way could enhance the robustness of introducing KG information. Comprehensive experiments are conducted on five Chinese medical datasets and experimental results show our model outperforms strong baselines remarkably. Consequently, our model can serve as an efficient medical text classifier with excellent performance. We release the code at https://github.com/nolongernome/ConKGNN.

1. Introduction

Medical text classification has a wide range of applications, including medical topic classification (Jelodar et al., 2020; Jiang et al., 2020), medical information index (Li et al., 2017), medical explanatory dialogue (Forcher et al., 2014) and medical diagnosis (Tang et al., 2020; Hosseini et al., 2018) etc. Every year, China publishes tens of thousands of medical journal articles that need to be labeled by libraries. For example, a drug-related document that comes from a medical journal demands to be further classified into a fine-grained drug label, such as anti-cancer drugs. Because these medical text contains intensive specialization and terminology, it has a strong demand for professional domain knowledge. For example, the word `ketamine` (氯胺酮), from the sentence "*The effect of ketamine on the brain* ", may not be observed during training. External knowledge could provide that `ketamine` is a kind of `general anesthetics` (全身性麻醉剂). It will facilitate classifying the sentence into the class *anesthetics*. Several works utilize domain knowledge in diverse applications (Gao et al., 2020; Li and Yang, 2021; Yang et al., 2021). There are some works that have explored the usage of external knowledge in the text classification task (Mishra et al., 2012; Abdollahi et al., 2019; Chen et al., 2019; Tang et al., 2022). However, they only use keywords to retrieve related knowledge. Then the related knowledge is simply concatenated with keywords (Mishra et al., 2012; Abdollahi et al., 2019) or the text representation (Chen et al., 2019; Tang et al., 2022) for classification. This way neglects the mutual influences between text and external knowledge.

Besides, we observe that not all the KG information is beneficial for classification. For instance, in the sentence "*Therapeutic effect of sirolimus on mouse model of skin collagen thesaurosis*" (西罗莫司对小鼠皮肤胶原沉着病模型的治疗作用), a keyword `collagen` (胶原) can retrieve the external knowledge that it is a kind of `topical hemostatic` (局部止血药). Such keyword and retrieval knowledge are misleading that may prevent this text from being classified into the class *immune drugs* (免疫药). This motivates us to enhance the robustness of the model introducing KG information.

In this paper, we argue that external knowledge is not only a supplement for text. The deeper mutual influences can further expand the text feature. Chinese medical text classification is chosen as our target. The external knowledge comes from a medical KG, which is composed

* Corresponding author.
E-mail addresses: grettelan@mail.nankai.edu.cn (G. Lan), mthu@nankai.edu.cn (M. Hu), liyee@mail.nankai.edu.cn (Y. Li), zyz@mail.nankai.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.engappai.2023.106057
Received 13 March 2022; Received in revised form 24 October 2022; Accepted 23 February 2023
Available online 15 March 2023
0952-1976/

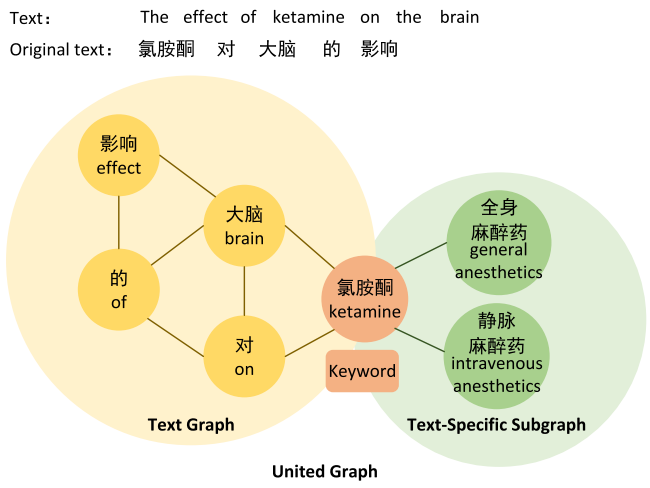


Fig. 1. An illustration for a united graph, combined with a text graph and a text-specific subgraph.

of professional terms and structured relationships. To make full use of the above knowledge, two major problems should be considered.

- How to learn the context of text and external knowledge interactively.
- How to enhance the robustness of the model since the external knowledge may be ambiguous.

To solve these two problems, we propose contrastive knowledge integrated graph neural networks (ConKGNN). For the first problem, we define those words that co-occur in both the medical document and medical KG as keywords. The keywords between text and KG provide a bridge, as `ketamine` in Fig. 1. A simple but effective fusion mechanism is designed to guarantee that the text and KG can affect each other mutually. Concretely, a text graph (see the left side of Fig. 1) is built with a sliding window (Zhang et al., 2020a), where the co-occurrence words in the window are linked with each other. With keywords, a text-specific subgraph (see the right side of Fig. 1) is also retrieved from medical KG. Then the two graphs above are integrated into a united graph. GNN which can catch semantic and structured information is proposed to facilitate learning node interactions among the united graph. In advance, the whole medical KG is pre-trained with graph embedding to preserve global semantic and structured information. As such, the text and KG interact through keywords, where the influences become deeper with the model training.

For the second problem, we propose a united graph supervised contrastive training objective based on contrastive learning (Khosla et al., 2020; Yan et al., 2021; You et al., 2020). Many works have demonstrated that contrastive learning contributes to model robustness (Zeng and Cui, 2022; Xu et al., 2021). We intuitively generalize such ability of contrastive learning to alleviate the misleading information introduced by KG. Specifically, with a random cutting-off strategy (Chen et al., 2020), an augmented united graph is obtained. It is trained to maximize the agreement with the full united graph, and also push away with other classes in a supervised contrastive manner. Cutting off nodes leads to learning a trade-off of how much external knowledge to use. In this way, our model can alleviate the influence of ambiguous information and improve robustness.

The main contributions of our paper are as follows:

- We propose contrastive knowledge integrated GNN (ConKGNN) for Chinese medical text classification. GNN is explored to enhance the mutual influences between text and KG. Bridged by keywords, the text and related KG are transferred into a united graph. Modeling this united graph by GNN tends to learn features in an interactive way and obtain more informative text features, which is remarkably beneficial for classifying medical texts.
- we propose a united graph supervised contrastive training objective to explore the problem of KG knowledge ambiguities. It learns a trade-off of how much external knowledge to use. Thus it enhances the robustness of our model by better utilizing the semantic and structured knowledge.
- We carry out extensive experiments on five real-world datasets. The results demonstrate that our proposed model outperforms strong baselines significantly.

This paper is organized as follows: The background literature on text classification with external knowledge and GNN-based text classification is introduced in Section 2. The structure and formula details of our proposed model are introduced in Section 3. The systematical experiments, including datasets, numerical results, and comprehensive analysis are given in Section 4. The conclusion and future work are summarized in Section 5.

2. Related work

2.1. Text classification with external knowledge

Several works have explored how to incorporate external knowledge for text classification. Traditional methods focus on the keywords that exist both in the text and knowledge base. Abdollahi et al. (2019) utilize a domain-specific dictionary and swarm optimization to select key features as input. Mishra et al. (2012) extract medical entities from a text, retrieve their relations from a database, and then concatenate the entities and relations to a kernel function. Many deep learning methods are also proposed. Yao et al. (2019a) combine the rule-based feature and the text feature to classify disease. Chen et al. (2019) inject conceptual information into attention mechanisms to obtain the conception feature and combine it with the text feature to the classifier. Tang et al. (2022) utilize CNN and attention mechanisms to obtain conception and keywords features as text extensions. Zhang et al. (2020b, 2021a) obtain prior knowledge by pre-trained on the large medical corpus or medical KG based on BERT (Vaswani et al., 2017). Though these works introduce external knowledge for text classification, knowledge is exploited in isolation. The mutual influences between the text and external knowledge may be ignored.

2.2. GNN-based text classification

The performance of traditional text classification methods heavily relies on manual features (Joachims, 2001; Dai et al., 2007; Altınel et al., 2015). With the development of deep learning, convolutional neural networks (CNN) (Zhang et al., 2021b; Chung et al., 2019; Zhang et al., 2022) and recurrent neural networks (RNN) (Mikolov et al., 2010) based methods have been proposed, which help this task by automatically learning features. Compared with the above methods, GNN-based methods can capture non-consecutive structure information (Zhang et al., 2020a; Shang et al., 2021). Defferrard et al. (2016) firstly propose graph convolutional neural networks and employ them in the text classification task. Yao et al. (2019b) utilize text nodes and weighted edges to build a corpus graph based on Defferrard et al. (2016). Huang et al. (2019) convert each text to a graph and support online training. Further, Zhang et al. (2020a) propose a graph-based method for inductive text classification. Wu et al. (2021) propose two windows GNN-based models to take account of the local semantic information and the global co-occurrence information. Although these GNN-based methods can utilize non-consecutive structure information and semantic information of the text, they have little attention to the situation where the model lacks domain knowledge. Based on the characteristic of GNN, we design the united graph and derive the original text and external domain-specific knowledge to learn from each other. Meanwhile, valuable semantic and structured information is obtained by the text feature.

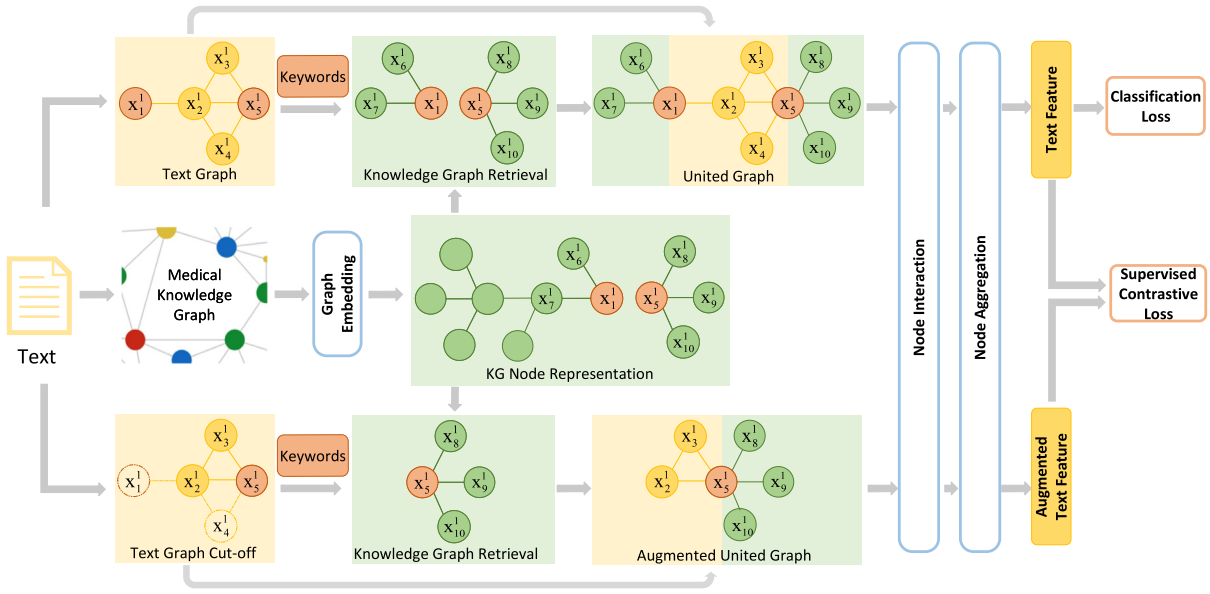


Fig. 2. An overview of ConKGNN. For the convenience of the display, x_i^1 colored in yellow is the node embedding of the text, x_i^1 colored in orange is the node embedding of the keyword, and x_i^1 colored in green is the node embedding of the related KG. The representations of KG nodes and keywords are pre-trained by graph embedding on the whole KG. Best view in color.

2.3. Contrastive learning

Contrastive Learning is proposed to alleviate the drawbacks of cross-entropy loss, such as sensitivity to noisy labels and poor margins. Khosla et al. (2020) extend the self-supervised batch contrastive learning to the supervised setting, allowing the model to effectively leverage label information. Recently, many works apply contrastive learning to obtain robust sentence representation in the domain of NLP (Li et al., 2021; Yan et al., 2021). In the graph domain, Wu et al. (2021) develop contrastive learning for GNN. Song et al. (2022) design two-layer constructive learning, including node-level contrastive learning and graph-level contrastive learning, to improve the performance of knowledge tracing. To find the optimal augmentation method in graph-based contrastive learning, Xu et al. (2021) propose practical principles and argue that the graph encoder should be task-relevant and simple. You et al. (2021) design a unified optimization framework to automatically select augmentation method and Yin et al. (2022) design a learnable graph augmentation generator to automatically generate augmentations for a graph. Motivated by these promising works, we also apply graph-level contrastive learning to deal with the noise from the KG and propose a united graph supervised contrastive training objective to explore this problem.

3. Proposed model

In this section, we propose ConKGNN for Chinese medical text classification. As shown in Fig. 2, utilizing the KG information, each text is processed into a united graph, meanwhile, an augmented united graph. Then the text feature and the augmented text feature will be generated respectively by GNN. The text feature is used to compute the classification objective. Both features are used to compute united graph supervised contrastive objective. Two objectives are jointly exploited to train the model. The proposed ConKGNN mainly includes the following.

- To preserve the global external knowledge, a KG graph embedding module aims to pre-train node representations among the whole KG.
- Bridged by keywords, the united graph is integrated by the text graph and its relevant text-specific subgraph. All nodes in the united graph interact with GNN.
- A united graph supervised contrastive training objective module generates an augmented united graph and maximizes the agreement with the original united graph. The supervised contrastive loss affects the learning of semantic and structured knowledge.

3.1. KG graph embedding

To take advantage of the global medical prior knowledge, we first pre-train graph embedding on the medical KG. Many knowledge graphs have been built (Bosselut et al., 2021; Shao et al., 2021; Borrego et al., 2021) and widely applied in NLP tasks. OMAHA,[1] a large-scale Chinese medical KG, with 1.24 million terms, 2.92 million relations, and 980,000 ontologies, is produced by a professional medical KG group named Huizhi and utilized as external information. To learn the global knowledge, the whole medical KG is pre-trained with node2vec graph embedding algorithm (Grover and Leskovec, 2016). It intends to make nearby nodes, or context-alike nodes, have similar representations. Concretely, upon the OMAHA, node2vec uses biased random walks, which provide a balance with depth-first and breadth-first network searching, to obtain sequences as contexts. And then node2vec employs word2vec (Mikolov et al., 2013) to establish d dimension KG node representation. By training the self-supervised node2vec on OMAHA, we can obtain domain-rich, high quality, and informative node representation. The pre-train of graph embedding is independent of the main model and the representations of part KG nodes will be fed into a KG text-specific subgraph in the next module.

3.2. KG retrieval

We utilize keywords to retrieve text-related nodes from KG. With these nodes, we obtain a text-specific subgraph, which contains the most relevant knowledge about the text. Concretely, the words co-occurring between the text and KG are selected as the keywords and retrieved their related KG terms. Since the term is not directly linked to another term in OMAHA, we use the ontology to obtain related nodes. In OMAHA, ontology is usually an abstractive string ID. A term is connected directly to its ontology and an ontology may connect to another ontology. For example, the term set $\eta_1 = \{\texttt{stomach}, \texttt{tummy}\}$ is

[1] https://hita.omaha.org.cn

connected to ontology1, η_2 = {stomachache, stomach pain} belongs to ontology2, the term stomach is a keyword. Since ontology1 is connected to ontology2, η_1, η_2 and connections from stomach to these terms can be retrieved. There may be more than one keyword in the text. The relevant information can be obtained by repeating the above operations for each keyword. As such, text-specific KG node set S_{sub} and their connections construct the text-specific subgraph, defined as below.

$$\begin{aligned} S_{\text{sub}} &= \{s_i | i \in [1, n]\}, \\ C_{sub} &= \{c_{ij} | i \in [1, n]; j \in \alpha_i\}, \end{aligned} \tag{1}$$

where n is the amount of text-specific subgraph nodes, α_i is the connected node set of the i_{th} subgraph node and C_{sub} is the subgraph edge set. And then, we adopt pre-trained KG node representations to initialize S_{sub}.

$$R_{sub} = \left\{\mathbf{x}_i | i \in [1, n]\right\}, \tag{2}$$

where R_{sub} is the node representation set and $\mathbf{x}_i$ is the embedding of i_{th} node. Then the text-specific subgraph is formulated in $G_{sub} = \{R_{sub}, C_{sub}\}$.

3.3. United graph

In order to preserve the semantic and structured information of the text, the text is processed into a text graph. Then the text graph and text-specific subgraph are combined to interactively enhance each other. Specifically, we first add KG terms into the vocabulary of pyltp (Che et al., 2020), a Chinese segmentation tool, and employ it to obtain Chinese word set S_{text} for the text graph:

$$S_{text} = \{s_i | i \in [1, m]\}, \tag{3}$$

where m is the number of the text graph nodes. In a sliding window (Zhang et al., 2020a) upon the text, text graph edge set C_{text} is formed by co-occurrences between words. The text graph $G_{text} = \{R_{text}, C_{text}\}$ is formulated as below.

$$\begin{aligned} R_{text} &= \left\{\mathbf{x}_i | i \in [1, m]\right\}, \\ C_{text} &= \{c_{ij} | i \in [1, m]; j \in [i - w, i + w]\}, \end{aligned} \tag{4}$$

where $\mathbf{x}_i$ is the embedding of i_{th} word in S_{text} and R_{text} is the node representation set. It is worth noting that the keywords are initialized by KG node representations. General embeddings (Mikolov et al., 2013) are utilized to initialize other nodes. In this way, keywords are enhanced by external knowledge and play a better role in bridging.

Since keywords are shared by the text graph G_{text} and text-specific subgraph G_{sub}, they are utilized to combine such two graphs into a united one G_{uni}.

$$\begin{aligned} S_{uni} &= S_{text} \cup S_{sub} \\ C_{uni} &= C_{text} \cup C_{sub}, \\ R_{uni} &= R_{text} \cup R_{sub}, \end{aligned} \tag{5}$$

where S_{uni}, C_{uni} and R_{uni} indicate the node, edge, and representation set of the united graph, separately. In this way, each original text is processed into a united graph, which consists of the semantic and structured information of both graphs.

3.4. Node interaction

To interactively update the representations of united graph nodes, gated graph neural networks (GGNN) (Li et al., 2016) is adopted. As shown in Fig. 3, during each interaction step, nodes are influenced by all their neighbors in the graph. United graph nodes can obtain better-structured representations, merging both the domain-specific and general information after updating multiple steps. In this way, text nodes and KG nodes can learn the information from each other mutually. The node interaction is defined as:

$$\mathbf{a}^q = \mathbf{A}\mathbf{x}^{q-1}\mathbf{W}_a, \tag{6}$$

where matrix $\mathbf{A} \in \mathbb{R}^{|S_{uni}| \times |S_{uni}|}$ is the adjacency matrix of the united graph. $\mathbf{x}^{q-1} \in \mathbb{R}^{|S_{uni}| \times d}$ is node representation obtained at the previous step and $\mathbf{W}_a$ is trainable parameter matrix.

Gated recurrent mechanism is utilized to update interactive representations:

$$\begin{aligned} \mathbf{z}^q &= \sigma(\mathbf{W}_z\mathbf{a}^q + \mathbf{V}_z\mathbf{x}^{q-1} + \mathbf{b}_z), \\ \mathbf{r}^q &= \sigma(\mathbf{W}_r\mathbf{a}^q + \mathbf{V}_r\mathbf{x}^{q-1} + \mathbf{b}_r), \\ \widetilde{\mathbf{x}}^q &= \tanh(\mathbf{W}_h\mathbf{a}^q + \mathbf{V}_h(\mathbf{r}^q \odot \mathbf{x}^{q-1}) + \mathbf{b}_h), \\ \mathbf{x}^q &= \widetilde{\mathbf{x}}^q \odot \mathbf{z}^q + \mathbf{x}^{q-1} \odot (1 - \mathbf{z}^q), \end{aligned} \tag{7}$$

where σ is the sigmoid function and $\odot$ is the Hadamard product. $\mathbf{z}$ and $\mathbf{r}$ are the update and reset gate, respectively. All $\mathbf{W}$, $\mathbf{V}$ and $\mathbf{b}$ are trainable parameter matrices and biases.

3.5. Node aggregation

This module aggregates the united graph node representations into a text feature $\mathbf{t}$. In view of each united node having different importance, our model employs a soft attention mechanism f_{att} in node-level to emphasize the main nodes. Then the attention weights are used in the projected representations f_{pro}, which are produced by a single hidden layer.

$$\begin{aligned} f_{att}(\mathbf{x}^q) &= \sigma(\mathbf{W}_{att}\mathbf{x}^q + \mathbf{b}_{att}), \\ f_{pro}(\mathbf{x}^q) &= \tanh(\mathbf{W}_{pro}\mathbf{x}^q + \mathbf{b}_{pro}), \\ \mathbf{x} &= f_{att}(\mathbf{x}^q) \odot f_{pro}(\mathbf{x}^q). \end{aligned} \tag{8}$$

Since the united graph is combined by two graphs, node representation $\mathbf{x}$ can be divided into the text graph part, denoted as $\mathbf{x}_{text}$, and the text-specific subgraph part. Our model chooses $\mathbf{x}_{text}$ to calculate the text feature $\mathbf{t}$. With masking the text-specific subgraph part, medical KG deeply introduces external knowledge in the case of not changing the original text structure. We average the node features of $\mathbf{x}_{text}$ and apply a max-pooling function to obtain $\mathbf{t}$ with the motivation of every text node plays a role in the text and the most attention-getting nodes contribute more explicitly (Zhang et al., 2020a). The text feature $\mathbf{t}$ comprises not only general and domain-specific knowledge but also their structured information.

$$\mathbf{t} = \text{Mean}(\mathbf{x}_{text}) + \text{Max}(\mathbf{x}_{text}). \tag{9}$$

Finally, the text feature is input to a fully connected layer to predict all classes.

3.6. United graph supervised contrastive training objective

To explore a trade-off of how much external knowledge to introduce, united graph supervised contrastive training objective is proposed. Concretely, some words are randomly cut off from a text with a rate of μ, which produces a cut-off text graph G_{text}^{aug}. Based on this text graph, a cut-off text-specific subgraph G_{sub}^{aug} is retrieved. The augmented united graph G_{uni}^{aug} is combined with G_{text}^{aug} and G_{sub}^{aug}. It is worth noting that for a text without a keyword, only G_{text}^{aug} is kept.

After data augmentation, both the original united graph G_{uni}^{ori} and the augmentation G_{uni}^{aug} are fed into the node interaction and node aggregation modules, obtaining the corresponding text features, i.e. $\mathbf{t}_{ori}$ and $\mathbf{t}_{aug}$. Graph-based supervised contrastive loss aims to enhance the discriminative ability and robustness of our model. Specifically, for each batch with N samples, we obtain $2N$ text features after augmentation. Each text feature is trained to approach the positive features. Here the positive features are composed of: (1) its corresponding augmented feature; (2) the text features that belong to the same class, from the remaining $2N - 2$ features. Finally, we average all $2N$ classification losses to obtain the final contrastive loss $\mathcal{L}_{sc}$ which is defined as:

$$\mathcal{L}_{sc} = \frac{-1}{|B|} \sum_{i \in B} \frac{1}{|S(i)|} \sum_{s \in S(i)} \log \frac{\exp(sim(t_i, t_s)/\tau)}{\sum_{o \in O(i)} \exp(sim(t_i, t_o)/\tau)}, \tag{10}$$

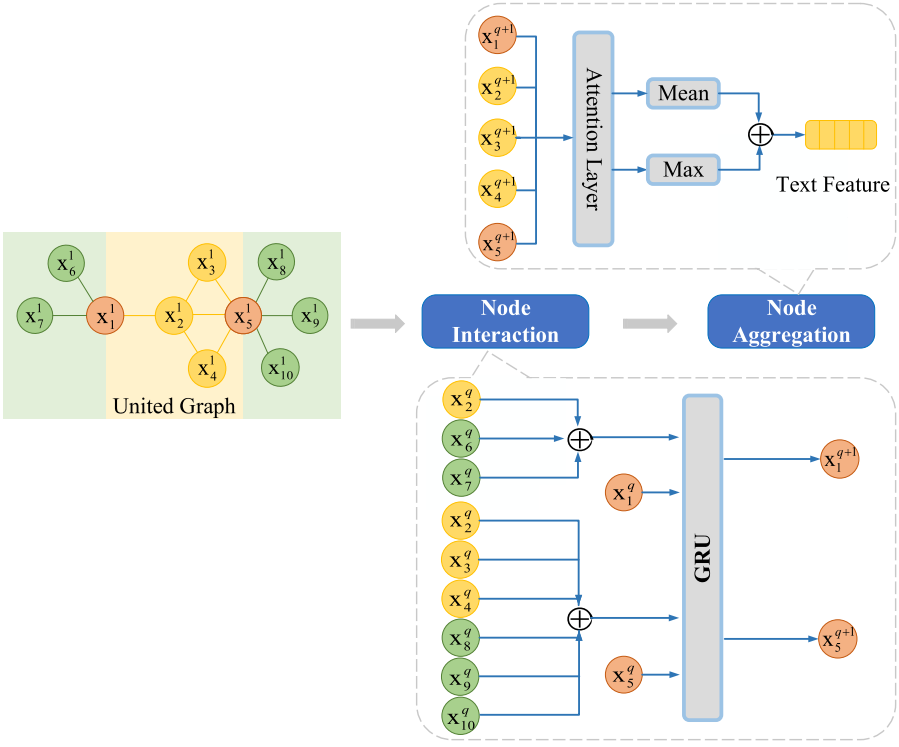


Fig. 3. The structure of node interaction and node aggregation. At q_{th} step, $\mathbf{x}_1^q$ and $\mathbf{x}_5^q$ are taken for example to illustrate that each node is influenced by all its neighbor nodes.

Table 1
Data statistics. AN indicates the average amount of nodes on the dataset.

Datasets	Classes	Docs	Train	Validation	Test	AN of G_{uni}	AN of G_{text}	AN of G_{sub}
Drugs	88	41,386	32,989	4,164	4,233	74.89	73.46	3.45
DA	42	14,482	11,587	1,484	1,411	71.62	68.47	4.49
RC	24	12,616	10,085	1,253	1,278	72.47	68.83	5.00
EM	36	20,099	15,997	2,073	2,029	73.14	70.18	4.58
CHIP-CTC	44	30,644	24,565	3,064	3,015	14.63	13.34	2.00

where B is a batch including $2N$ text features, $i \in B$ is the i_{th} feature of B. $O(i) \equiv B\backslash\{i\}$ and $S(i)$ is the feature set where samples, distinct from i, share the same class with i. $sim(\cdot)$ denotes the cosine similarity function, τ controls the temperature.

Finally, the original text feature $\mathbf{t}_{ori}$ is used to compute cross-entropy category loss $\mathcal{L}_{ce}$ (Rumelhart et al., 1986). The classification loss $\mathcal{L}_{ce}$ and a supervised contrastive loss $\mathcal{L}_{sc}$ are jointly exploited to train the proposed model:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{sc}, \tag{11}$$

where λ is a hyper-parameter to balance two losses.

4. Experiment

4.1. Datasets

To evaluate our proposed method, we access five real-world Chinese medical classification datasets, including 4 datasets build from Chinese medical journals published on HowNet,[2] and a public dataset CHIP-CTC (Zong et al., 2021). Table 1 displays the data statistics. These medical datasets are challenging since some terms are similar. For instance, there are three classes, *atrophic gastritis* (萎缩性胃炎), *superficial gastritis* (浅表性胃炎), and *hypertrophic gastritis* (肥厚性胃炎) all mentioning a high-frequency word `gastritis` (胃炎).

4.1.1. Hownet medical journal datasets

Hownet is one of the most important Chinese comprehensive journal databases, providing industrial, agricultural, economic, educational, and medical journals, etc. The title, keywords, and abstract of each journal are shown on the website. Vocational journal annotators have labeled these journal articles, including medical journals, following the Chinese Library Taxonomy (CLT)[3]. These labels are used to index journals. We collect 4 sufficient medical datasets based on CLT labels and each text is combined with its title, keywords, and abstract. The detailed classes distributions of 4 datasets are displayed in Appendix.

- **Drugs** is one of the most sufficient datasets of CLT, containing 41,386 articles and 88 classes. It includes neurological drugs, cardiovascular drugs, and other kinds of drugs.
- **DA** is the abbreviation of "digestive system and abdominal diseases (消化系统及腹部疾病)", which contains 14,482 articles and 42 classes, includes various intestinal diseases and stomach diseases, etc.
- **RC** is the abbreviation of "respiratory and chest diseases (呼吸系统和胸部疾病)". This dataset contains 12,616 journal articles and 24 classes, such as various lung diseases, bronchial diseases, trachea diseases, etc.
- **EM** is the abbreviation of "endocrine gland and metabolic diseases (内分泌及代谢疾病)". It contains 20,099 journal articles and 36 classes, such as various thyroid diseases, adrenal gland diseases, pituitary system diseases, etc.

[2] https://www.cnki.net/

[3] https://www.clcindex.com

Table 2
Overall performance in terms of accuracy.

Model	Drugs	DA	RC	EM	CHIP-CTC
TextRNN	0.5368 ± 0.0060	0.7125 ± 0.0038	0.7107 ± 0.0042	0.7018 ± 0.0018	0.7579 ± 0.0065
TextRCNN	0.5985 ± 0.0069	0.7325 ± 0.0087	0.7291 ± 0.0048	0.7314 ± 0.0103	0.8167 ± 0.0012
fastText	0.5347 ± 0.0053	0.7033 ± 0.0021	0.7041 ± 0.0034	0.7036 ± 0.0032	0.7641 ± 0.0027
TextGCN	0.6051 ± 0.0013	0.6779 ± 0.0018	0.6740 ± 0.0069	0.6686 ± 0.0006	0.7864 ± 0.0007
STKCA	0.5618 ± 0.0022	0.7151 ± 0.0049	0.7308 ± 0.0075	0.7189 ± 0.0075	0.7737 ± 0.0067
TextCNN	0.5727 ± 0.0043	0.7386 ± 0.0044	0.7557 ± 0.0039	0.7437 ± 0.0007	0.8074 ± 0.0025
TextING	0.5946 ± 0.0024	0.7378 ± 0.0057	0.7528 ± 0.0059	0.7570 ± 0.0032	0.8285 ± 0.0025
BERT	0.6134 ± 0.0019	0.7616 ± 0.0077	0.7563 ± 0.0045	0.7666 ± 0.0086	0.8267 ± 0.0018
PERT	0.6199 ± 0.0035	0.7468 ± 0.0018	0.7571 ± 0.0042	0.7650 ± 0.0130	0.8265 ± 0.0008
MC-BERT	0.6170 ± 0.0032	0.7645 ± 0.0032	0.7613 ± 0.0033	0.7783 ± 0.0043	0.8304 ± 0.0014
SMedBERT	0.6226 ± 0.0024	**0.7661 ± 0.0064**	0.7635 ± 0.0021	**0.7829 ± 0.0038**	0.8308 ± 0.0040
ConKGNN	**0.6400**† **± 0.0014**	0.7563 ± 0.0017	**0.7693**† **± 0.0023**	0.7677 ± 0.0033	**0.8366 ± 0.0010**

Table 3
Overall performance in terms of macro F1.

Model	Drugs	DA	RC	EM	CHIP-CTC
TextRNN	0.1331 ± 0.0120	0.3499 ± 0.0327	0.2433 ± 0.0116	0.2194 ± 0.0198	0.3814 ± 0.0313
TextRCNN	0.2524 ± 0.0127	0.5299 ± 0.0197	0.3690 ± 0.0264	0.4282 ± 0.0231	0.6442 ± 0.0082
fastText	0.1924 ± 0.0073	0.4752 ± 0.0093	0.3394 ± 0.0197	0.3466 ± 0.0127	0.4415 ± 0.0198
TextGCN	0.2052 ± 0.0048	0.4819 ± 0.0148	0.3803 ± 0.0126	0.3753 ± 0.0146	0.6819 ± 0.0119
STKCA	0.1776 ± 0.0149	0.3993 ± 0.0267	0.3963 ± 0.0227	0.3733 ± 0.0244	0.5781 ± 0.0206
TextCNN	0.2622 ± 0.0307	0.5186 ± 0.0185	0.3835 ± 0.0137	0.3642 ± 0.0135	0.6607 ± 0.0183
TextING	0.2777 ± 0.0059	0.5760 ± 0.0120	**0.5264 ± 0.0082**	0.5143 ± 0.0083	0.7553 ± 0.0100
BERT	0.1756 ± 0.0036	0.5962 ± 0.0335	0.3385 ± 0.0393	0.4385 ± 0.0811	0.6913 ± 0.0164
PERT	0.1682 ± 0.0033	0.4793 ± 0.0243	0.3405 ± 0.0371	0.3750 ± 0.0966	0.6435 ± 0.0561
MC-BERT	0.1765 ± 0.0044	0.6007 ± 0.0114	0.3778 ± 0.0237	0.4508 ± 0.0629	0.6787 ± 0.0204
SMedBERT	0.1842 ± 0.0106	0.5982 ± 0.0176	0.3654 ± 0.0393	0.5013 ± 0.0545	0.6688 ± 0.0405
ConKGNN	**0.3055**† **± 0.0067**	**0.6012 ± 0.0040**	0.5153 ± 0.0227	**0.5366**† **± 0.0221**	**0.7676**† **± 0.0057**

4.1.2. CHIP-CTC dataset

CHIP-CTC is a short text classification dataset focusing on clinical trial screening standards. This dataset contains 30,644 articles, classified into 44 classes, such as age, therapy or surgery, laboratory examination, etc.

4.2. Evaluation metrics

To evaluate the models systematically, we adopt the accuracy, kappa coefficient, macro F1, and weighted F1 as evaluation metrics. Considering the imbalance of class distribution, macro F1, kappa coefficient, and weighted F1 are utilized to evaluate our model. The kappa coefficient is widely utilized to measure imbalanced tasks (Zhu et al., 2021). This metric is based on the confusion matrix and is the indicator of consistency inspection i.e. whether the model predictions and labels are consistent.

4.3. Implementation details

At the ratio of 8:1:1, each dataset is randomly split into train, validation, and test sets. The validation set is used for hyper-parameter tuning. The text words embeddings are initialized with 300 dimensional pre-trained skip-gram (Mikolov et al., 2013) vectors with Sogou News corpus.[4] The KG nodes are initialized with 300 dimension node representations pre-trained upon the whole medical KG. Unknown text nodes are initialized randomly. The step of united graph node interaction is 2 and hidden units of GGNN are 96. The size of the sliding window is 3. The dimension of the text feature is 96 and the dropout rate is 0.5. The loss balancing hyper-parameter λ is set to 0.3 on the RC, Drugs, and DA datasets, 0.01 on the CHIP-CTC dataset, and 0.005 on the EM dataset. The temperature τ is set to 0.1. The ratio of cut-off is set as 0.15 on the RC, Drugs, and DA datasets, 0.1 on the CHIP-CTC dataset, and 0.2 on the EM dataset. The batch size is 2,048. Adam (Kingma and Ba, 2015) is the optimizer of our model with a learning rate of 0.0045. We obtain the mean and standard deviation from the results of 5 runs, and re-run all baselines.

[4] https://github.com/Embedding/Chinese-Word-Vectors

4.4. Compared models

To evaluate our model extensively, we compare it with three kinds of strong baselines:

- **Text classification models**: (1) TextRNN (Mikolov et al., 2010) which has advantage of capturing longer and bidirectional sequence information. (2) TextCNN (Kim, 2014) which can extract key n-gram information in sentences and is widely used in the medical domain (Li et al., 2019). (3) TextRCNN (Lai et al., 2015) which uses RNN and the max-pooling method to establish a text classification model. (4) fastText (Joulin et al., 2017) which utilizes the average of all word features and the N-Gram features of the text to get the text feature. (5) BERT (Vaswani et al., 2017) which is a powerful and large-scale bidirectional language representation model. (6) PERT (Cui et al., 2022), which shuffles the word order of the text and learns to predict the location of the original token based on BERT.
- **GNN-based models**: (1) TextGCN (Yao et al., 2019b), which proposes to build heterogeneous graphs based on the texts and words, and apply GCN to text classification. (2) TextING (Zhang et al., 2020a), which builds a graph for each text and utilizes GGNN to learn the structured information of the text.
- **Merging external knowledge models**: (1) STCKA (Chen et al., 2019), which utilizes external knowledge bases to enhance the semantics of the text and introduces attention mechanisms to measure the importance of each concept of knowledge bases. (2) MC-BERT (Zhang et al., 2020b), which is pre-trained on a variety of Chinese medical corpora via masking different granularity tokens. (3) SMedBERT (Zhang et al., 2021a), which is a pre-trained language model trained on a large-scale medical corpus with relation information of linked entities based on KG.

4.5. Experimental results

The evaluation results by accuracy, macro F1, kappa coefficient, and weighted F1 are displayed in Table 2, Table 3, Table 4, and

Table 4
Overall performance in terms of kappa coefficient.

Model	Drugs	DA	RC	EM	CHIP-CTC
TextRNN	0.4907 ± 0.0066	0.6431 ± 0.0109	0.6075 ± 0.0092	0.6171 ± 0.0089	0.7284 ± 0.0072
TextRCNN	0.5613 ± 0.0082	0.7121 ± 0.0085	0.6724 ± 0.0073	0.6600 ± 0.0197	0.7926 ± 0.0014
fastText	0.4839 ± 0.0056	0.6506 ± 0.0028	0.5974 ± 0.0155	0.6247 ± 0.0092	0.7335 ± 0.0028
TextGCN	0.5630 ± 0.0015	0.6570 ± 0.0034	0.5799 ± 0.0025	0.5910 ± 0.0016	0.7594 ± 0.0007
STKCA	0.5169 ± 0.0017	0.6623 ± 0.0067	0.6345 ± 0.0058	0.6310 ± 0.0123	0.7466 ± 0.0071
TextCNN	0.5342 ± 0.0049	0.6991 ± 0.0067	0.6789 ± 0.0043	0.6774 ± 0.0037	0.7832 ± 0.0030
TextING	0.5515 ± 0.0035	0.7138 ± 0.0058	0.6777 ± 0.0077	0.6905 ± 0.0044	0.8097 ± 0.0020
BERT	0.5743 ± 0.0022	0.7424 ± 0.0083	0.6858 ± 0.0053	0.7013 ± 0.0105	0.8059 ± 0.0020
PERT	0.5824 ± 0.0041	0.7262 ± 0.0021	0.6865 ± 0.0050	0.7009 ± 0.0163	0.8053 ± 0.0009
MC-BERT	0.5775 ± 0.0035	0.7455 ± 0.0034	0.6921 ± 0.0044	0.7165 ± 0.0065	0.8099 ± 0.0017
SMedBERT	0.5849 ± 0.0029	**0.7473 ± 0.0068**	0.6950 ± 0.0038	**0.7236 ± 0.0052**	0.8101 ± 0.0044
ConKGNN	**0.6020**† **± 0.0034**	0.7316 ± 0.0042	**0.6957 ± 0.0049**	0.7019 ± 0.0039	**0.8151**† **± 0.0030**

Table 5
Overall performance in terms of weighted F1.

Model	Drugs	DA	RC	EM	CHIP-CTC
TextRNN	0.5063 ± 0.0088	0.6486 ± 0.0098	0.6724 ± 0.0069	0.6873 ± 0.0073	0.7482 ± 0.0085
TextRCNN	0.5804 ± 0.0076	0.7212 ± 0.0094	0.7286 ± 0.0054	0.7255 ± 0.0147	0.8086 ± 0.0010
fastText	0.5036 ± 0.0045	0.6666 ± 0.0038	0.6712 ± 0.0100	0.6973 ± 0.0073	0.7505 ± 0.0034
TextGCN	0.5651 ± 0.0020	0.6684 ± 0.0041	0.6572 ± 0.0023	0.6711 ± 0.0018	0.7840 ± 0.0010
STKCA	0.5243 ± 0.0046	0.6629 ± 0.0082	0.6990 ± 0.0033	0.7017 ± 0.0081	0.7726 ± 0.0062
TextCNN	0.5604 ± 0.0071	0.7081 ± 0.0074	0.7338 ± 0.0038	0.7372 ± 0.0033	0.8014 ± 0.0044
TextING	0.5605 ± 0.0033	0.7221 ± 0.0019	0.7421 ± 0.0051	0.7506 ± 0.0027	0.8266 ± 0.0025
BERT	0.5671 ± 0.0020	0.7468 ± 0.0079	0.7350 ± 0.0049	0.7550 ± 0.0133	0.8226 ± 0.0021
PERT	0.5741 ± 0.0045	0.7237 ± 0.0038	0.7358 ± 0.0044	0.7539 ± 0.0157	0.8200 ± 0.0029
MC-BERT	0.5656 ± 0.0037	**0.7498 ± 0.0037**	0.7410 ± 0.0033	0.7675 ± 0.0084	0.8263 ± 0.0017
SMedBERT	0.5760 ± 0.0032	0.7498 ± 0.0066	0.7436 ± 0.0058	**0.7758 ± 0.0053**	0.8255 ± 0.0053
ConKGNN	**0.6100**† **± 0.0013**	0.7413 ± 0.0019	**0.7561**† **± 0.0034**	0.7633 ± 0.0026	**0.8339**† **± 0.0014**

Table 5. The best performances of each dataset are marked in bold. The marker † suggests p-values<0.05 comparing with SMedBERT. We can observe that ConKGNN outperforms all compared models on Drugs, RC, and CHIP-CTC datasets. And it can achieve comparable performance on the DA and EM datasets compared with SMedBERT. ConKGNN demonstrates an obvious advantage according to macro F1. It shows that ConKGNN successfully makes the text and KG knowledge mutually influence each other, meanwhile, improving the effectiveness of introducing external knowledge.

Among all the baselines, SMedBERT outperforms the others on accuracy, except for the CHIP-CTC dataset, which indicates the effect of prior knowledge pre-trained on the medical corpus and KG. The results of macro F1 imply that the introduced knowledge may make SMedBERT more biased towards certain classes, while ConKGNN does not. Besides, according to the computation complexity (Section 4.6.2), the performance of our model has a significant advantage and is more suitable for practical engineering. Except for transformer-based models, TextING outperforms the others on macro F1 in all datasets and weighted F1 in most datasets, which indicates the importance of structure information. Following this research direction, our method ConKGNN also pays attention to structures, effectively utilizing the semantic and structured KG information. Compared with the best baseline on the Drugs dataset, ConKGNN improves the macro F1 from 27.77% to 30.55%, accuracy from 62.26% to 64.00%, kappa coefficient from 58.49% to 60.20% and weighted F1 from 58.04% to 61.00%, respectively. This indicates the effectiveness of ConKGNN. Particularly, compared with STKCA, ConKGNN outperforms it on all datasets. This demonstrates that simply concatenating external knowledge is not enough. ConKGNN successfully learns deeper mutual influences between text and external knowledge, which helps generate more informative text features. Moreover, graph-based supervised contrastive learning can effectively alleviate the influence of ambiguous information and improve robustness, which further improves the performance of our model.

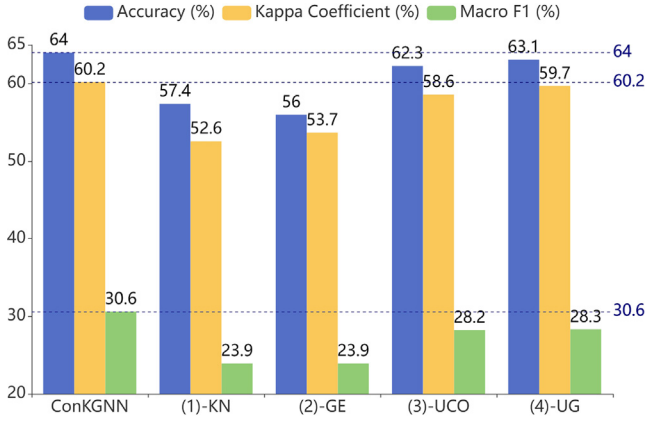


Fig. 4. Ablation studies on the Drugs dataset.

4.6. Analysis

4.6.1. Ablation test

We conduct the ablation study to explore the contributions of individual components of ConKGNN. Fig. 4 suggests the results.

(1) Without KG nodes (-KN): only exploits KG nodes to enrich the vocabulary for word segmentation but not merge them into the united graph. The results present that without KG nodes, all evaluation metrics drop significantly. It shows the effectiveness of mutual influence between the text and KG information, suggesting that text-specific subgraphs really provide relevant knowledge for the text, and improve the capability of our model.

(2) Without graph embedding (-GE): initializes all united graph nodes with general word embeddings. Removing graph embedding makes the performance drop. General embeddings lack professional medical knowledge and structured information like graph embedding pre-trained on KG. Besides, most KG nodes are uncommon. These medical terms appear low-frequency and some of them are even nonexistent

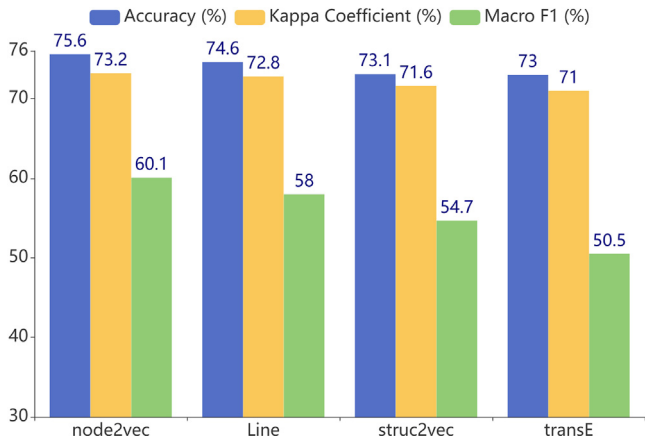


Fig. 5. The results of KG nodes initialized with graph embedding methods in our model on the DA dataset.

Table 6
FLOPs and trainable params of ConKGNN, TextCNN and SMedBERT.

Model	FLOPs	Trainable params
ConKGNN	**22.23394G**	**0.1026M**
TextCNN	48.0070G	1.4515M
SMedBERT	190.2736G	85.6420M

in the general corpus. Since then, KG graph embedding is essential for the process of knowledge fusion in our model.

(3) Without united graph supervised contrastive training objective (-UCO). The results indicate that the performances are decreased. United graph supervised contrastive objective tends to utilize augmented united graph and label information to improve the robustness of our model. It maximizes the agreement between features of the original and augmented united graphs, meanwhile, pushes them away from other classes. The results denote that with the improvement of robustness benefiting from united graph supervised contrastive training objective, the performance of our model is enhanced.

(4) Without united graph augmentation (-UG). In this variant, only the label information can be utilized to influence the feature space of classes. We can observe that it is less performed compared to the full version and outperforms the version without united graph supervised contrastive objective. The result indicates that both united graph augmentation and label information benefit our model. Especially, united graph augmentation is really valuable to learn a trade-off of the KG information. Employing this augmentation can alleviate the influence of ambiguous information and improve robustness.

4.6.2. Computation complexity

For practical purposes, we use FLOPs (floating point operations) and trainable parameters to discuss the computational complexity of the models. FLOPs are widely used to measure the computational complexity of the model. We utilize THOP[5] or tf.profiler.profile[6] to obtain such messages for different models. As shown in Table 6, ConKGNN maintains outstanding performance yet both its FLOPs and trainable parameters are significantly smaller than the traditional model TextCNN and transformer-based model SMedBERT. This demonstrates that ConKGNN has an obvious advantage in practical engineering.

[5] https://gitcode.net/mirrors/Lyken17/pytorch-OpCounter?utm_source=csdn_github_accelerator

[6] https://tensorflow.google.cn/versions/r1.15/api_docs/python/tf/profiler/profile

4.6.3. Graph embedding study

KGs have recently been applied in the development of novel explanation interface techniques in the field of explainable AI due to their graph structure and conceptual information (Holzinger et al., 2021). To exploit such information, graph embedding is utilized in our model. In this section, the impact of several graph embedding methods is compared on the DA dataset. For a fair comparison, we only change the pre-train algorithm on ConKGNN.

- **transE** (Bordes et al., 2013): It explains relationships as translation operating to model embedding entities and relationships of multi-relational data.
- **LINE** (Tang et al., 2015): This approach utilizes an approximate factorization of the adjacency matrix and preserves both first order and second proximities as the embedding.
- **node2vec** (Grover and Leskovec, 2016): This is a kind of random walk-based method, which preserves higher-order proximity between nodes by maximizing the probability of occurrence of subsequent nodes in fixed-length biased-random walks.
- **struc2vec** (Ribeiro et al., 2017): It uses a hierarchical structure to measure the relationship of nodes at different scales. It builds a multi-layer graph to encode architectural similarity and generates an architectural context for nodes.

The results of different graph embedding methods are reported in Fig. 5. It can be seen that the proposed method with node2vec consistently outperforms the model variants of using other graph embedding methods. A possible explanation for these achievements is that node2vec utilizes the skip-gram to pre-train KG node representations, which has the same way as the word embeddings for text. This makes the representations of the two sources nodes closer in semantics and logic space. Hence, it facilitates the knowledge "flowing" between the text and KG.

4.6.4. Graph-based augmentation study

Graph-based augmentation aims at creating rational and novel graphs by certain transformations. The augmentation can generalize the augmented united graph and enhance the robustness of our model on some sides. Since universally appropriate graph-based augmentation has not existed, we compare the effect of 3 kinds of random augmentation methods and 2 kinds of sorting augmentation methods on the DA dataset.

- **Cut-off**: This strategy will randomly discard a portion of nodes accompanied by their connections on the text graph. The augmented united graph is built based on such text graph. The implied prior enforced by cut-off is that discarded part does not alter the semantics of the original united graph. Our model ConKGNN uses this strategy.
- **Embedding Shuffle** (You et al., 2020): The node embeddings of the united graph will be perturbed by randomly exchanging. It is assumed that changing certain positions of nodes does not influence the predictions of the model much.
- **Edge Shuffle** (You et al., 2020): The connections of the united graph will be perturbed by randomly adding or cutting. The implied assumption is that the semantics of the united graph has certain robustness against the connection variances.
- **Top-K KG Node Sorting** (Zhang et al., 2021a): Since the medical KG nodes construct a graph structure, this augmentation method utilizes PageRank (Page et al., 1999) on the KG to calculate the node scores and filter low-score nodes. As a high connection frequency means the node is important in the graph, this method assumes a high connection frequency means the KG node is also important for classification.
- **Bottom-K KG Node Sorting** (Zhang et al., 2021a): PageRank is utilized to calculate the node scores on medical KG and high-score nodes are removed. This method assumes a low connection frequency means the KG node is rare and important for classification.

Table 7
Graph-based data augmentation studies on the DA.

Augmentation	Accuracy	Kappa Coefficient	Macro F1	Weighted F1
Cut-off	**0.7563**	0.7316	**0.6012**	**0.7413**
Embedding Shuffle	0.7536	0.7285	0.5920	0.7366
Edge Shuffle	0.7503	0.7257	0.5928	0.7342
Top-K KG Node Sorting	0.7528	**0.7322**	0.6007	0.7373
Bottom-K KG Node Sorting	0.7479	0.7274	0.6004	0.7333

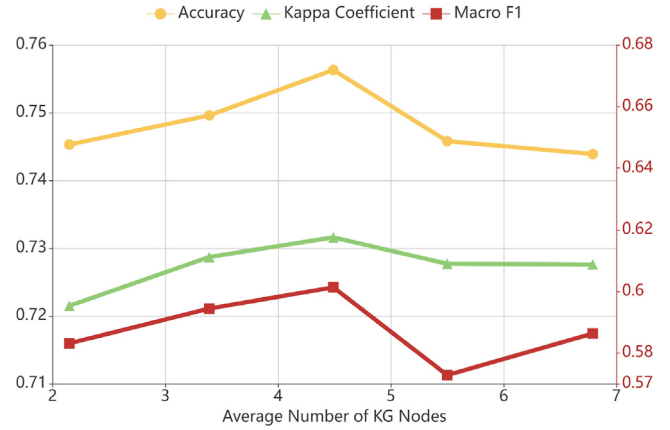


Fig. 6. Evaluation results on the test set of the DA dataset according to the different average number of terms from KG. The Y-axis of macro F1 is on the right.

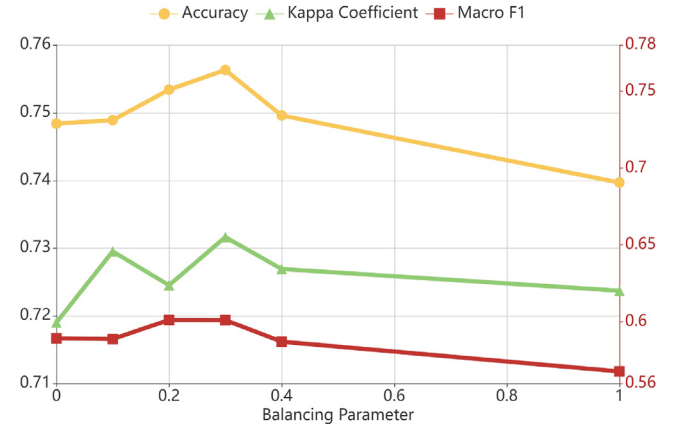


Fig. 7. Evaluation results on the test set of the DA dataset according to the balancing parameter. The Y-axis of macro F1 is on the right.

Table 7 displays the results of graph-based augmentations. It can be observed that the performance of the variants with embedding shuffle or edge shuffle is consistently poorer than the cut-off. This implied that node positions and edges are valuable and essential for our model. Both KG node sorting methods are not better than the cut-off. This means that KG nodes with high or low connection frequency are not always important to our classification method. Compared with sorting methods, the random cut-off has no sorting standard which makes the model encounter more diverse situations. This enhances the model robust. The experimental results demonstrate the robustness benefiting from random cut-off improves the performance of our model. Meanwhile, the cut-off is also simple. Thus ConKGNN adopts a cut-off strategy.

4.6.5. Different number of nodes from KG

Given a keyword, the related terms from KG are retrieved to build the joint graph. However, keywords may have a different number of neighbors. To obtain the balance, we set an upper bound for the number of terms retrieved from KG. The experimental results are displayed in Fig. 6. Note that the x-axis indicates the average of terms, thus it is not an integer number.

It can be observed that as the number grows, the performances increase. More KG nodes introduce richer knowledge for our model, which enhances the text features and improves the discriminative ability of the model. When the scale is too large, the performances tend to decline. A possible reason is that though more terms from KG introduce more external knowledge, it might be dominant for the final representation. Yet the major information, including the structure and semantics of the text, is weakened.

4.6.6. Influence of balancing parameter

Hyper-parameter λ controls the influence of united graph supervised contrastive training objective. Fig. 7 reports the results of fine-tuning the balancing parameter λ. A too-small λ will extremely reduce the performance of supervised contrastive loss $\mathcal{L}_{sc}$. Since the effect of the augmented united graph depends on $\mathcal{L}_{sc}$, when the model pays less attention to $\mathcal{L}_{sc}$, the model may ignore its effect. It can be seen that after a small optional range, as the λ grows, the performance gradually decreases. We argue that though a more large weight of $\mathcal{L}_{sc}$ makes the model more robust, the model may ignore the performance of classification.

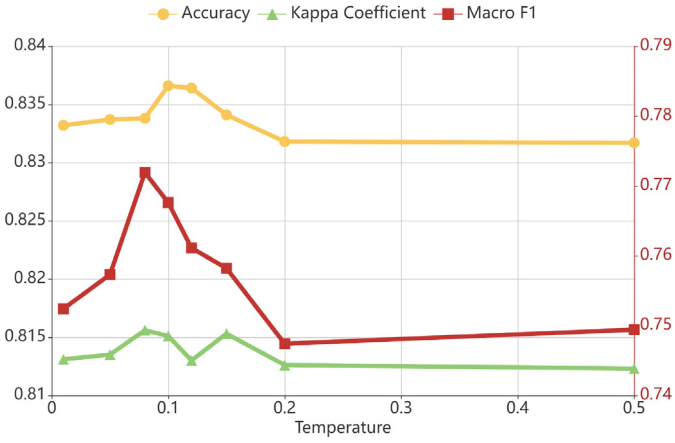
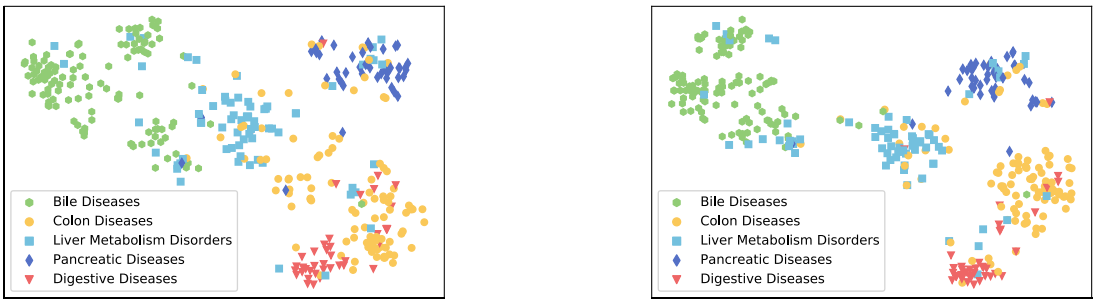


Fig. 8. The influence of the temperature on the CHIP-CTC dataset. The Y-axis of macro F1 is on the right.

4.6.7. Influence of temperature

Some previous works (Li et al., 2021; Yan et al., 2021) have discovered that the temperature τ can influence gradients during backpropagation by controlling normalized distribution' smoothness. To explore the influence of τ in our model, which introduces KG, we conduct experiments and the result is presented in Fig. 8. The result denotes a too-small temperature makes the distribution too sharp and the model performs badly. A too-large temperature makes the distribution too smooth and the model also performs badly. Especially, it can be found that macro F1 is more sensitive to the temperature among the performance measures.



(a) ConKGNN-UCO (b) ConKGNN

Fig. 9. Visualization of text features on the DA dataset.

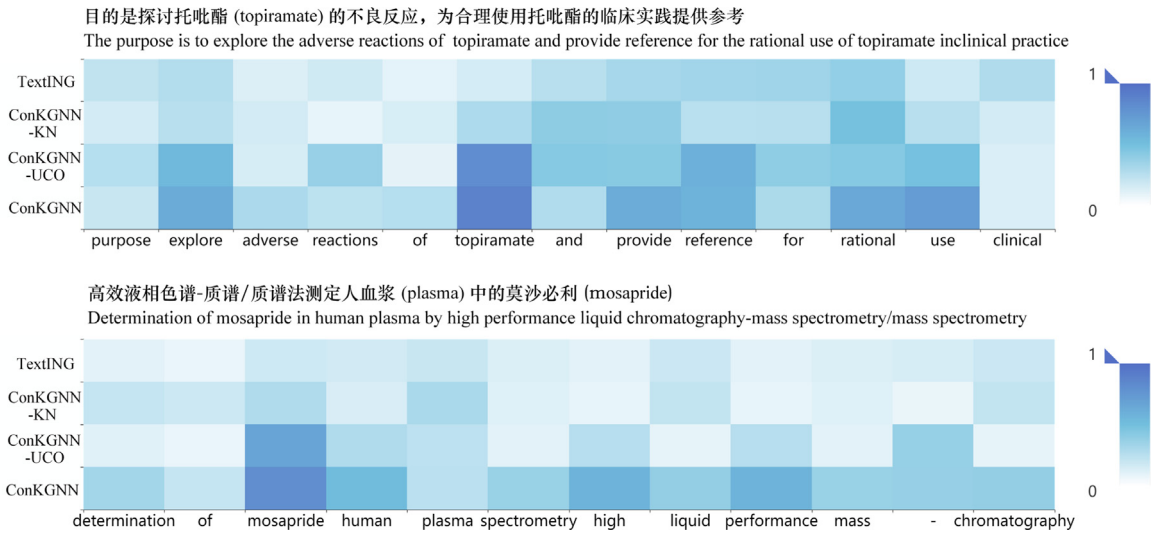


Fig. 10. Attention visualization on the Drugs dataset.

4.6.8. Feature visualization

In Fig. 9, text features of 5 classes from the DA dataset are visualized using t-SNE (Van der Maaten and Hinton, 2008). The visualization shows that our model without united graph supervised contrastive training objective can slightly cluster the text features. And then, united graph supervised contrastive training objective further enhances the ability of clustering. ConKGNN tightly clusters the text features, maximizes the agreement of text features belonging to the same class, and preserves them distant from other classes. According to the improvement of performance, a suitable tight cluster can benefit the classification task.

4.6.9. Attention visualization

In Fig. 10, attention weights for two example texts from the Drugs dataset are visualized. The deeper color indicates that the words have a larger attention score. For the first sentence, the strong baseline method TextING allocates similar attention weights for all words. As a result, when removing KG nodes from our method (ConKGNN-KN), the attention weights are also similar among words. Removing united graph supervised contrastive learning objective from our model (ConKGNN-UCO) and the full version of ConKGNN can highlight the words that are discriminative for medical text classification. For example, `topiramate` (托吡酯) is a keyword. We can obtain its related node `antiepileptic drug` (抗癫痫药) from the medical KG. With the help of mutual learning, `topiramate` can receive a larger attention score in our method. Besides, compared with ConKGNN-UCO, the full version of ConKGNN pays more attention to `topiramate`, which means contrastive KG further improves the attention to the important word. It is also worth noting that the first two methods predict this text wrongly, while our models predict correctly. This also indicates the effectiveness of our models.

For the second sentence, we can get similar observations as the first one. Besides, there are two keywords in this sentence: `mosapride` (莫沙必利) and `plasma` (血浆). Though they all get enhanced by mutual learning, for ConKGNN, the first keyword `mosapride` which is discriminative for medical text classification gets a large attention weight. And less discriminative keyword `plasma` gets small attention. This indicates that our proposed ConKGNN can keep robustness for introducing KG knowledge.

4.6.10. Error analysis

A deeper error analysis is provided about some hard cases, which are struggled to deal with for the proposed model. For example, the text "*Study on amifostine induced apoptosis of K562 cells*" (氨磷汀诱导K562细胞凋亡的研究) is predicted wrongly because the term `amifostine` (氨磷汀) does not exist in our large scale medical KG. Although this term is not included in the medical KG, 注射用氨磷汀 (amifostine for injection) is included. Since then, the lack of some terms can be divided into two situations, one is that the KG does not contain any related nodes; the other is that the KG contains related professional nodes. The first situation points out that a more comprehensive KG is required for the medical field and a good medical KG will not only contribute to text classification but also broader tasks or applications. The second situation shows that some common terms may be ignored by professional KG while related specialized terms are collected. How to utilize these related terms in a situation without some

common terms is a challenge. In addition, another text "*Top 10 public health achievements of the United States from 2001 to 2010*" (2001–2010 年美国十大公共卫生成就) has no medical terms, but annotated to the other drugs class. This sentence is a noise case in the dataset. Thus high-quality dataset is also demanding for supervised learning models. Lastly, a few sentences have keywords with KG. Yet such keywords in KG are isolated nodes, in which the domain-specific knowledge is hard to learn. These cases also trigger challenges in future research.

5. Conclusion

In this paper, we propose contrastive knowledge integrated graph neural networks (ConKGNN) for Chinese medical text classification. It utilizes medical expertise from a large medical KG and mutually learns the influences between text and KG. For the united graph node, GNN capacitates its interaction and facilitates information transfer. Furthermore, united graph supervised contrastive training objective improves the robustness of introducing KG knowledge by reducing the ambiguities of KG. Extensive experimental results demonstrate that our proposed model outperforms strong baselines significantly. All of these demonstrate the efficacy of our model. In future work, relation information of KG can be further explored to enhance the text. And more KGs can be applied to other classification tasks.

CRediT authorship contribution statement

Ge Lan: Methodology, Writing – original draft, Software, Visualization, Investigation. **Mengting Hu:** Methodology, Writing – review & editing, Supervision. **Ye Li:** Visualization, Investigation, Writing – review & editing. **Yuzhi Zhang:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my code and data.

Acknowledgment

This work is supported by National Major Science and Technology Project of China (2021YFB0300104).

Appendix A. Additional statistics of four Hownet datasets

To better display, the details of four HowNet datasets, Tables A.8–A.11 are added to describe their class distribution separately.

Appendix B. Cross-validation

We conduct 5-fold cross-validation on the Drugs and CHIP-CTC datasets. As shown in Tables B.12 and B.13, the results of cross-validation are consistent with the experiment results in Section 4.5. According to evaluation metrics of two datasets, ConKGNN outperforms all baselines, and SMedBERT outperforms the strong baselines TextCNN and TextING. The cross-validation results demonstrate that there is no potential bias in our experiments.

Table A.8
The statistics of classes on the DA dataset.

Class	Number
Cirrhosis	2191
Liver and gallbladder diseases	1865
Colon diseases	1466
Bowel diseases	1132
Stomach diseases	1105
Liver metabolism disorder	1069
Pancreatic diseases	972
Digestive and abdominal diseases	658
Esophageal diseases	437
Peptic ulcer diseases	394
Liver failure	382
Hepatitis	357
Upper gastrointestinal bleeding	273
Small bowel diseases	207
Colorectal diseases	186
Bile duct diseases	182
Gallbladder diseases	144
Bowel dysfunction	142
Peritoneal and retroperitoneal diseases	126
Intestinal obstruction	125
Gastritis	110
Cholelithiasis	97
Esophageal varices	95
Cecal disease, appendix diseases	87
Mesenteric diseases	81
Biliary cirrhosis	72
Atrophic gastritis	67
Liver abscess	60
Cholecystitis	59
Duodenal diseases	56
Coeliac diseases	54
Pylori diseases	51
Rectal diseases	47
Cardia diseases	44
Hepatolenticular degeneration liver cirrhosis	28
Other stomach diseases	25
Anal diseases	23
Portal cirrhosis	19
Gastrolithiasis	18
Esophageal stricture	17
Ileal diseases	12
Intussusception	11

Table A.9
The statistics of classes on the RC dataset.

Class	Number
Lung diseases	4932
Bronchial asthma	1944
Respiratory and chest diseases	1526
Respiratory failure	1067
Pulmonary embolism	957
Pneumonia (lung infection)	739
Bronchial diseases	270
Other lung diseases	170
Trachea and bronchial diseases	161
Bronchitis	139
Hydrothorax, pleural effusion	130
Interstitial pneumonia	119
Emphysema	96
Pneumothorax, hydropneumothorax	82
Diseases of the pleura and thoracic cavity	57
Airway diseases	55
Tracheal stenosis	55
Bronchiectasis	36
Atelectasis	27
Mediastinal diseases	24
Tracheitis	22
Bronchopneumonia	16
Empyema	15
Pulmonary hemosiderosis	11

Table A.10
The statistics of classes on the EM dataset.

Class	Number
Diabetes	7656
Diabetic coma and other complications	4725
Lipodystrophy	2275
Metabolic diseases	1046
Thyroid diseases	998
Islet diseases	627
Hyperthyroidism	584
Purine metabolism disorder	436
Hypothyroidism	252
Thyroiditis	205
Hyperparathyroidism	147
Endocrine and metabolic diseases	130
Goiter	128
Hyperadrenocorticism (cushing's disease)	112
Protein intermediate metabolism disorder	102
Adrenal diseases	101
Primary aldosteronism	89
Pituitary and diencephalitic-pituitary disorders	87
Gonadal diseases	74
Hyperinsulinemia, hypoglycemia	60
Pineal disease, early puberty (precocious puberty)	50
Carbohydrate metabolism disorder	44
Calcium and phosphorus metabolism disorders	39
Hypopituitarism	38
Hypoparathyroidism (tetany)	32
Posterior pituitary diseases	32
Acid–base imbalance	31
Other metabolic diseases	29
Porphyrin metabolism disorder	28
Water and salt metabolism disorder	23
Parathyroid diseases	17
Adrenal insufficiency (addison's disease)	16
Female gonadal (ovarian) disorders	14
Acromegaly	14
Other adrenal diseases	10
Thymus diseases	10

Table A.11
The statistics of classes on the Drugs dataset.

Class	Number
Antibiotics	7801
Antitumor and anticancer drugs	7026
Anesthetics	2895
Nervous system drugs	2464
Cardiovascular rugs	2234
Insulin and hypoglycemic drugs	1244
Drugs	1241
Antihypertensives, vasodilators	1117
Immune enhancers, immunosuppressants	1038
Anti-viral drugs	989
Antipyretic analgesics	905
Hypolipidemic, anti-atherosclerotic drugs	818
Hormone drugs	785
Enzyme	715
Digestive system drugs	636
Other drugs	624
Proteins, conjugated sugars, lipids	596
Antifungal drugs	575
Beta-lactenamines	568
Anticoagulants	488
Antipsychotics	469
Anticonvulsants, antiepileptic drugs	439
Respiratory medicine	398
Antimanic depressive drugs	333
Drugs that affect growth and metabolism	328
Anti-tuberculosis drugs, anti-leprosy drugs	324
Drugs for the infectious	278
Blood and hematopoietic drugs	261
Gallbladder, liver supporters	237
Macrolides	230
Antiarrhythmic drugs	170
Vitamin drugs	170
Sedative tranquilizer hypnotics	145
Amino acids and their derivatives	144
Anti-ulcer drug	142
Anti-schizophrenia drugs	142
Antiallergic drugs	131
Asthma medicine	126
Other metabolic drugs	123
Sex hormones	116
Contraceptives	108
Corticotropin and adrenocortical hormone	103
Antianginal drugs	100
Other drugs for the infectious	99
Disinfectant and antiseptics	93
Antiparasitic drugs	92
Thyroxine and antithyroid drugs	91
Occupational medicines, antidotes	90
Skeletal muscle relaxants	77
Antiprotozoal drugs	73
Aminoglycosides	71
Radiation sickness drugs	68
Antitumor antibiotics	68
Hemostatic	67
Emetics and antiemetics	65
Vitamin d	64
Other antineoplastic drugs	62
Peptides	61
Central stimulant	60
Family planning drugs	60
Cough medicine	58
Vitamin b	53
Autonomic nervous system drugs	52
Anti tremor paralysis drugs	50
Cardiotonic	49
Blood substitute	49
Vitamin e	48
Vitamin c	47
Sulfonamides	41
Antacids	39
Four-ring	38
Expectorant	36
Anticholinergics	34
Laxatives and antidiarrheals	32
Antimetabolites	32
Stomach medicine	32
Anti-anemia drugs	27
Insecticide, rodenticide	23
Minerals	20
Other cardiovascular system drugs	19
Antihelminthic drugs	18
Chloramphenicol and its derivatives	17
Anti-physical damage drugs	16
Vitamin k	14
Gastrointestinal antispasmodics	14
Medications that regulate water or electrolytes	14
Vitamin a	10
Cholinergics	10

Appendix C. Generalization study

To verify the generalization of ConKGNN, we use a real-world English medical classification dataset HowMed from HowNet. This dataset contains 5,665 articles and 40 classes. It includes drugs, heart diseases, and other kinds of medical classes. The dataset is divided into the train and test set at the ratio of 9:1 randomly. SNOMED CT is utilized as English medical KG. Different from Chinese, English text does not need word segmentation. To fit our method, we first use the KG to extract the KG nodes existing in the text. For nested nodes, the longer node is preferred. For the other words of the text, a word is a node. Therefore,

Table B.12
Cross-validation on the Drugs dataset.

Model	Accuracy	Kappa Coefficient	Macro F1	Weighted F1
TextCNN	0.5804 ± 0.0057	0.5303 ± 0.0072	0.2061 ± 0.0114	0.5290 ± 0.0089
TextING	0.5928 ± 0.0018	0.5507 ± 0.0013	0.2636 ± 0.0069	0.5575 ± 0.0026
SMedBERT	0.6207 ± 0.0073	0.5815 ± 0.0089	0.1913 ± 0.0192	0.5760 ± 0.0121
ConKGNN	**0.6429 ± 0.0009**	**0.6118 ± 0.0019**	**0.3352 ± 0.0071**	**0.6180 ± 0.0018**

Table B.13
Cross-validation on the CHIP-CTC dataset.

Model	Accuracy	Kappa Coefficient	Macro F1	Weighted F1
TextCNN	0.7926 ± 0.0024	0.7638 ± 0.0028	0.4624 ± 0.0113	0.7735 ± 0.0031
TextING	0.8026 ± 0.0023	0.7782 ± 0.0026	0.7345 ± 0.0100	0.8004 ± 0.0024
SMedBERT	0.8207 ± 0.0030	0.7977 ± 0.0033	0.5214 ± 0.0363	0.8078 ± 0.0046
ConKGNN	**0.8317 ± 0.0014**	**0.8110 ± 0.0016**	**0.7731 ± 0.0037**	**0.8303 ± 0.0016**

Table C.14
Generalization study on the HowMed dataset.

Model	Accuracy	Kappa Coefficient	Macro F1	Weighted F1
TextCNN	0.6784 ± 0.0049	0.6547 ± 0.0053	0.5490 ± 0.0121	0.6638 ± 0.0055
TextING	0.6846 ± 0.0081	0.6627 ± 0.0088	0.5390 ± 0.0107	0.6658 ± 0.0085
BioBERT	0.7113 ± 0.0048	0.6923 ± 0.0051	0.5471 ± 0.0132	0.6960 ± 0.0057
ConKGNN	**0.7186 ± 0.0016**	**0.6992 ± 0.0019**	**0.6063 ± 0.0065**	**0.7027 ± 0.0026**

for a node in the text graph, if it is not in the KG, it will be a word. Otherwise, it may be a word or multiple words according to KG.

For the HowMed dataset, AN of G_{uni} is 56.51, AN of G_{text} is 52.57, and AN of G_{sub} is 6.14. The KG nodes are initialized with 300-dimension node representations pre-trained upon the KG. The loss balancing hyper-parameter is set to 0.001. The temperature is set to 0.1. The ratio of cut-off is set as 0.1. The batch size is 512. We utilize 3 strong models, TextING, TextCNN, and BioBERT as baselines. The results is shown in Table C.14. We can observe that our proposed ConKGNN outperforms all compared models on the HowMed dataset. It is implied that our model has the potential to be used in the English language.

References

Abdollahi, M., Gao, X., Mei, Y., Ghosh, S., Li, J., 2019. An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimisation. In: 2019 IEEE Congress on Evolutionary Computation. CEC, IEEE, pp. 119–126. http://dx.doi.org/10.1109/CEC.2019.8790259.

Altınel, B., Ganiz, M.C., Diri, B., 2015. A corpus-based semantic kernel for text classification by using meaning values of terms. Eng. Appl. Artif. Intell. 43, 54–66. http://dx.doi.org/10.1016/j.engappai.2015.03.015.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O., 2013. Translating embeddings for modeling multi-relational data. Adv. Neural Inf. Process. Syst. 26, 2787–2795.

Borrego, A., Ayala, D., Hernández, I., Rivero, C.R., Ruiz, D., 2021. CAFE: Knowledge graph completion using neighborhood-aware features. Eng. Appl. Artif. Intell. 103, 104302. http://dx.doi.org/10.1016/j.engappai.2021.104302.

Bosselut, A., Le Bras, R., Choi, Y., 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, pp. 4923–4931. http://dx.doi.org/10.1609/aaai.v35i6.16625.

Che, W., Feng, Y., Qin, L., Liu, T., 2020. N-LTP: A open-source neural Chinese language technology platform with pretrained models. pp. 5812–5823, arXiv preprint arXiv:2009.11616.

Chen, J., Hu, Y., Liu, J., Xiao, Y., Jiang, H., 2019. Deep short text classification with knowledge powered attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. 33, pp. 6252–6259. http://dx.doi.org/10.1609/aaai.v33i01.33016252.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E., 2020. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. In: Proceedings of Machine Learning Research, 119, PMLR, pp. 1597–1607. http://dx.doi.org/10.48550/arXiv.2002.05709.

Chung, T., Xu, B., Liu, Y., Ouyang, C., Li, S., Luo, L., 2019. Empirical study on character level neural network classifier for Chinese text. Eng. Appl. Artif. Intell. 80, 1–7. http://dx.doi.org/10.1016/j.engappai.2019.01.009.

Cui, Y., Yang, Z., Liu, T., 2022. PERT: pre-training BERT with permuted language model. CoRR 1–14. http://dx.doi.org/10.48550/arXiv.2203.06906, arXiv:2203.06906.

Dai, W., Xue, G.-R., Yang, Q., Yu, Y., 2007. Transferring naive bayes classifiers for text classification. In: AAAI. 7, pp. 540–545.

Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering. Adv. Neural Inf. Process. Syst. 29, 3844–3852. http://dx.doi.org/10.48550/arXiv.1606.09375.

Forcher, B., Roth-Berghofer, T., Agne, S., Dengel, A., 2014. Intuitive justifications of medical semantic search results. Eng. Appl. Artif. Intell. 30, 1–17. http://dx.doi.org/10.1016/j.engappai.2014.01.013.

Gao, L., Gan, Y., Zhou, B., Dong, M., 2020. A user-knowledge crowd sourcing task assignment model and heuristic algorithm for Expert Knowledge Recommendation Systems. Eng. Appl. Artif. Intell. 96, 103959. http://dx.doi.org/10.1016/j.engappai.2020.103959.

Grover, A., Leskovec, J., 2016. Node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 855–864. http://dx.doi.org/10.1145/2939672.2939754.

Holzinger, A., Malle, B., Saranti, A., Pfeifer, B., 2021. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Inf. Fusion 71, 28–37. http://dx.doi.org/10.1016/j.inffus.2021.01.008.

Hosseini, A., Chen, T., Wu, W., Sun, Y., Sarrafzadeh, M., 2018. Heteromed: Heterogeneous information network for medical diagnosis. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 763–772. http://dx.doi.org/10.1145/3269206.3271805.

Huang, L., Ma, D., Li, S., Zhang, X., Wang, H., 2019. Text level graph neural network for text classification. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. Association for Computational Linguistics, pp. 3442–3448. http://dx.doi.org/10.18653/v1/D19-1345.

Jelodar, H., Wang, Y., Orji, R., Huang, S., 2020. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. IEEE J. Biomed. Health Inf. 24 (10), 2733–2742. http://dx.doi.org/10.1109/JBHI.2020.3001216.

Jiang, L., Sun, X., Mercaldo, F., Santone, A., 2020. DECAB-LSTM: Deep Contextualized Attentional Bidirectiona LSTM for cancer hallmark classification. Knowl.-Based Syst. 210, 106486. http://dx.doi.org/10.1016/j.knosys.2020.106486.

Joachims, T., 2001. A statistical learning learning model of text classification for support vector machines. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 128–136. http://dx.doi.org/10.1145/383952.383974.

Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2017. Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. pp. 427–431. http://dx.doi.org/10.18653/v1/e17-2068.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. Adv. Neural Inf. Process. Syst. 33, 18661–18673. http://dx.doi.org/10.48550/arXiv.2004.11362.

Kim, Y., 2014. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a Meeting of SIGDAT, a Special Interest Group of the ACL. ACL, pp. 1746–1751. http://dx.doi.org/10.3115/v1/d14-1181.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. pp. 1–15. http://dx.doi.org/10.48550/arXiv.1412.6980.

Lai, S., Xu, L., Liu, K., Zhao, J., 2015. Recurrent convolutional neural networks for text classification. In: Twenty-Ninth AAAI Conference on Artificial Intelligence. pp. 2267–2273. http://dx.doi.org/10.1609/aaai.v29i1.9513.

Li, D., Azoulay, P., Sampat, B.N., 2017. The applied value of public investments in biomedical research. Science 356 (6333), 78–81. http://dx.doi.org/10.1126/science.aal0010.

Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R.S., 2016. Gated graph sequence neural networks. In: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings. pp. 1–20. http://dx.doi.org/10.48550/arXiv.1511.05493.

Li, Q., Weng, L., Ding, X., 2019. A novel neural network-based method for medical text classification. Future Internet 11 (12), 255. http://dx.doi.org/10.3390/fi11120255.

Li, Y., Yang, C., 2021. Domain knowledge based explainable feature construction method and its application in iron making process. Eng. Appl. Artif. Intell. 100, 104197. http://dx.doi.org/10.1016/j.engappai.2021.104197.

Li, Z., Zou, Y., Zhang, C., Zhang, Q., Wei, Z., 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. Association for Computational Linguistics, pp. 246–256. http://dx.doi.org/10.18653/v1/2021.emnlp-main.22.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings. pp. 1–12.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., Khudanpur, S., 2010. Recurrent neural network based language model. In: Interspeech. 2, Makuhari, pp. 1045–1048.

Mishra, M., Huan, J., Bleik, S., Song, M., 2012. Biomedical text categorization with concept graph representations using a controlled vocabulary. In: Proceedings of the 11th International Workshop on Data Mining in Bioinformatics. pp. 26–32. http://dx.doi.org/10.1145/2350176.2350181.

Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The PageRank citation ranking: Bringing order to the web. Technical Report 1999–66, Stanford InfoLab, pp. 0–17.

Ribeiro, L., Saverese, P., Figueiredo, D.R., 2017. Struc2vec: Learning node representations from structural identity. In: The 23rd ACM SIGKDD International Conference. pp. 385–394. http://dx.doi.org/10.1145/3097983.3098061.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back propagating errors. Nature 323 (6088), 533–536. http://dx.doi.org/10.1038/323533a0.

Shang, R., Meng, Y., Zhang, W., Shang, F., Jiao, L., Yang, S., 2021. Graph convolutional neural networks with geometric and discrimination information. Eng. Appl. Artif. Intell. 104, 104364. http://dx.doi.org/10.1016/j.engappai.2021.104364.

Shao, B., Li, X., Bian, G., 2021. A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph. Expert Syst. Appl. 165, 113764. http://dx.doi.org/10.1016/j.eswa.2020.113764.

Song, X., Li, J., Lei, Q., Zhao, W., Chen, Y., Mian, A., 2022. Bi-CLKT: Bi-graph contrastive learning based knowledge tracing. Knowl.-Based Syst. 241, 108274. http://dx.doi.org/10.1016/j.knosys.2022.108274.

Tang, Z., Dai, D., Chen, Z., Chen, T., 2022. Short text classification combining keywords and knowledge. In: 2022 2nd International Conference on Consumer Electronics and Computer Engineering. ICCECE, pp. 662–665. http://dx.doi.org/10.1109/ICCECE54139.2022.9712673.

Tang, C., Ji, J., Tang, Y., Gao, S., Tang, Z., Todo, Y., 2020. A novel machine learning technique for computer-aided diagnosis. Eng. Appl. Artif. Intell. 92, 103627. http://dx.doi.org/10.1016/j.engappai.2020.103627.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q., 2015. Line: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1067–1077. http://dx.doi.org/10.1145/2736277.2741093.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11), 2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008. http://dx.doi.org/10.48550/arXiv.1706.03762.

Wu, X., Luo, Z., Du, Z., Wang, J., Gao, C., Li, X., 2021. TW-TGNN: Two windows graph-based model for text classification. In: 2021 International Joint Conference on Neural Networks. IJCNN, pp. 1–8. http://dx.doi.org/10.1109/IJCNN52387.2021.9534150.

Xu, D., Cheng, W., Luo, D., Chen, H., Zhang, X., 2021. Infogcl: Information-aware graph contrastive learning. Adv. Neural Inf. Process. Syst. 34, 30414–30425. http://dx.doi.org/10.48550/arXiv.2110.15438.

Yan, Y., Li, R., Wang, S., Zhang, F., Wu, W., Xu, W., 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, pp. 5065–5075. http://dx.doi.org/10.18653/v1/2021.acl-long.393.

Yang, W., Dong, Y., Du, Q., Qiang, Y., Wu, K., Zhao, J., Yang, X., Zia, M.B., 2021. Integrate domain knowledge in training multi-task cascade deep learning model for benign–malignant thyroid nodule classification on ultrasound images. Eng. Appl. Artif. Intell. 98, 104064. http://dx.doi.org/10.1016/j.engappai.2020.104064.

Yao, L., Mao, C., Luo, Y., 2019a. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med. Inf. Decis. Mak. 19 (3), 31–39. http://dx.doi.org/10.1186/s12911-019-0781-4.

Yao, L., Mao, C., Luo, Y., 2019b. Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. 33, pp. 7370–7377. http://dx.doi.org/10.1609/aaai.v33i01.33017370.

Yin, Y., Wang, Q., Huang, S., Xiong, H., Zhang, X., 2022. AutoGCL: Automated graph contrastive learning via learnable view generators. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, the Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. AAAI Press, pp. 8892–8900. http://dx.doi.org/10.1609/aaai.v36i8.20871.

You, Y., Chen, T., Shen, Y., Wang, Z., 2021. Graph contrastive learning automated. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. In: Proceedings of Machine Learning Research, 139, PMLR, pp. 12121–12132. http://dx.doi.org/10.48550/arXiv.2106.07594.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y., 2020. Graph contrastive learning with augmentations. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual. 33, Curran Associates, Inc., pp. 5812–5823. http://dx.doi.org/10.48550/arXiv.2010.13902.

Zeng, H., Cui, X., 2022. SimCLRT: A simple framework for contrastive learning of rumor tracking. Eng. Appl. Artif. Intell. 110, 104757. http://dx.doi.org/10.1016/j.engappai.2022.104757.

Zhang, T., Cai, Z., Wang, C., Qiu, M., Yang, B., He, X., 2021a. SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, pp. 5882–5893. http://dx.doi.org/10.18653/v1/2021.acl-long.457.

Zhang, Z., He, N., Li, Q., Wang, K., Gao, H., Gao, T., 2022. DetectPMFL: Privacy-preserving momentum federated learning considering unreliable industrial agents. IEEE Trans. Ind. Inform. 1. http://dx.doi.org/10.1109/TII.2022.3140806.

Zhang, N., Jia, Q., Yin, K., Dong, L., Gao, F., Hua, N., 2020b. Conceptualized representation learning for Chinese biomedical text mining. CoRR 1–4. http://dx.doi.org/10.48550/arXiv.2008.10813, arXiv:2008.10813.

Zhang, Z., Xu, X., Gong, W., Chen, Y., Gao, H., 2021b. Efficient federated convolutional neural network with information fusion for rolling bearing fault diagnosis. Control Eng. Pract. 116, 104913. http://dx.doi.org/10.1016/j.conengprac.2021.104913.

Zhang, Y., Yu, X., Cui, Z., Wu, S., Wen, Z., Wang, L., 2020a. Every document owns its structure: Inductive text classification via graph neural networks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 334–339. http://dx.doi.org/10.18653/v1/2020.acl-main.31.

Zhu, Q., Deng, W., Zheng, Z., Zhong, Y., Guan, Q., Lin, W., Zhang, L., Li, D., 2021. A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification. IEEE Trans. Cybern. 1–15. http://dx.doi.org/10.1109/TCYB.2021.3070577.

Zong, H., Yang, J., Zhang, Z., Li, Z., Zhang, X., 2021. Semantic categorization of Chinese eligibility criteria in clinical trials using machine learning methods. BMC Med. Inf. Decis. Mak. 21 (1), 128. http://dx.doi.org/10.1186/s12911-021-01487-w.