Reducing Knowledge Noise for Improved Semantic Analysis in Biomedical Natural Language Processing Applications

Usman Naseem[1], Surendrabikram Thapa[2], Qi Zhang[3,4], Liang Hu[3,4], Anum Masood[5], Mehwish Nasim[6,7]
[1]University of Sydney, Australia [2]Virginia Tech, USA
[3]Tongji University, China [4]DeepBlue Academy of Sciences, China
[5]Norwegian University of Science and Technology, Norway
[6]University of Western Australia [7]Flinders University, Australia

Abstract

Graph-based techniques have gained traction for representing and analyzing data in various natural language processing (NLP) tasks. Knowledge graph-based language representation models have shown promising results in leveraging domain-specific knowledge for NLP tasks, particularly in the biomedical NLP field. However, such models have limitations, including knowledge noise and neglect of contextual relationships, leading to potential semantic errors and reduced accuracy. To address these issues, this paper proposes two novel methods. The *first method* combines knowledge graph-based language model with nearest-neighbor models to incorporate semantic and category information from neighboring instances. The *second method* involves integrating knowledge graph-based language model with graph neural networks (GNNs) to leverage feature information from neighboring nodes in the graph. Experiments on relation extraction (RE) and classification tasks in English and Chinese language datasets demonstrate significant performance improvements with both methods, highlighting their potential for enhancing the performance of language models and improving NLP applications in the biomedical domain.

1 Introduction

Language models (LM) have become increasingly popular in a wide range of applications (Adhikari et al., 2023; Shah et al., 2023). LMs have also shown great promise in assisting clinical decision-making. LMs for clinical decision-making are based on massive amounts of text data, including medical literature, clinical notes, and patient records (Lewis et al., 2020; Adhikari et al., 2021). By analyzing medical records and identifying patterns, language models can help doctors identify potential health risks and recommend appropriate treatment options (Kalyan et al., 2022; Naseem et al., 2022a). Additionally, LMs can be used to assist with drug discovery by analyzing vast amounts of scientific literature and identifying potential drug targets (Naseem et al., 2022b).

LMs can also improve patient outcomes by assisting with patient education. LMs can analyze patient records and suggest personalized educational resources, such as videos or articles, to help patients better understand their conditions and treatments. Likewise, language models can help clinicians communicate more effectively with patients by providing real-time translation services for patients who speak different languages. However, these language models are not without their shortcomings. For example, the cross-entropy loss used in fine-tuning language models can lead to poor generalization of performance, as pointed out by Liu et al. (2016). Similarly, LMs can be prone to overfitting as the predictions are made through linear classifiers added directly to the top of pre-trained LMs (Li et al., 2021). Moreover, they may neglect the relationship between textual contexts, impacting performance. Given the importance of the medical field, errors in language models can have significant consequences. Thus, efforts to improve the performance of language models are being made. A possible scope for improvement can be using knowledge from nearest neighbors.

Khandelwal et al. (2019) employed a k-nearest neighbor (kNN) approach to enhance the performance of LMs. The kNN-LM approach uses nearest-neighbor models to improve language modeling by explicitly memorizing rare patterns and improving performance, indicating that the representation learning problem is easier than the prediction problem. Similarly, Kassner and Schütze (2020) approached an open-domain question-answering problem with kNN-BERT, a combination of BERT's prediction for a given question with a kNN search. The authors demonstrated through their experimental results and evaluation that incorporating kNN into the BERT-based question-answering model was effective in retriev-

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 272–277

ing accurate factual information. This model particularly excelled in providing answers to less frequent and difficult questions and was able to handle even recent events that were not included in the training data of the BERT model. Unlike other studies that use kNN to generate augmented samples based on pre-trained language models, Li et al. (2021) utilized a kNN classifier as the decision maker. It was demonstrated that incorporating kNNs with traditional fine-tuning of BERT-like models can significantly improve accuracy in both rich-source and few-shot settings and improve robustness against adversarial attacks. Nearest neighbors have shown immense success in improving the interpretability of models as well (Wallace et al., 2018). The use of kNN in explaining the model behavior in language models is a topic that has been attracting the research community (Rajani et al., 2020). Interpretable machine learning is key to building trustworthiness in AI systems used in healthcare. Interpretability can provide clinicians and patients with insights into the reasoning behind a particular decision made by a machine learning model, making it easier to understand and trust. This can help to improve patient outcomes and increase the acceptance of AI systems in healthcare. Additionally, interpretability can help identify and address biases in machine learning models.

While LMs like BERT (Devlin et al., 2018) have limited ability to capture global information, Graph Neural Networks (GNNs) have proved to be better at it (Wu et al., 2023). Language models like BERT are good at capturing contextual information. In order to make the best out of a language model, it needs to have global information as well as proper contextual information. Utilizing the strengths of both graph neural networks and language models, Lu et al. (2020) proposed Vocabulary Graph Convolutional Network (VGCN)-BERT. The motivation for combining VGCN and BERT is to allow them to collectively build an optimal representation while performing tasks such as classification. VGCN models the relationships between words in a text by representing them as nodes in a graph, while BERT is a state-of-the-art language model that can effectively capture the contextual information of text data. This enhanced representation led to improved performance in text classification tasks, as demonstrated by the results of the study.

Global knowledge is very important in the medical domain because medical data is inherently complex and interconnected. In order to effectively analyze medical data, it is essential to capture the relationships between different medical concepts and understand their context within the larger medical knowledge network (Rasmy et al., 2021). For example, in the field of medical diagnosis, a patient's symptoms and medical history need to be considered in the context of the larger medical knowledge network to accurately diagnose their condition (Lin et al., 2021). By leveraging global knowledge, medical professionals can better understand the relationships between different medical concepts and make more well-informed decisions regarding patient care. Furthermore, medical research often involves analyzing large datasets containing vast medical information (Naseem et al., 2021a). By utilizing global knowledge, researchers can better identify patterns and relationships within the data, leading to more accurate and insightful findings. Thus, there is a pressing need to build representation comprising contextual and global information.

Pre-trained language models are mostly generic in nature and lack domain-specific knowledge (Liu et al., 2023). This can be a problem for tasks that require access to domain-specific knowledge, such as medical text classification and medical relation extraction (Naseem et al., 2021b). To mitigate this problem, Liu et al. (2020) proposed K-BERT, a knowledge-enabled language representation model that addresses this problem by incorporating knowledge from a knowledge graph into its language representation. This knowledge incorporation allows K-BERT to perform better on tasks that require access to domain-specific knowledge. K-BERT allows the triplets to be injected into sentences as knowledge, making it useful for domain-specific tasks. Thus, leveraging K-BERT for domain knowledge and kNN and GNN for global knowledge, we present methods for integrating (i) K-BERT with GNN and (ii) K-BERT with kNN. Our contributions are as follows:

- We present a method to improve the performance of knowledge graph-based language models by integrating semantic information from neighboring instances.
- We demonstrate significant performance improvements in relation extraction (RE) and classification tasks using our methods, showcasing their potential for improving NLP applications in the biomedical domain.

2 Methodology

As discussed above, our aim is to utilize the information of the neighbors in the dataset to help improve the performance of the language model in various downstream tasks. We use K-BERT as a language model because of its flexibility to adapt to domain knowledge.

2.1 Nearest Neighbor Enriched Language Model

Our nearest neighbor algorithm, kNN, extracts instances similar to the test samples from the feature library and uses the information of the instances to assist in prediction. kNN uses the classification label information of the neighbors. In order to combine K-BERT with kNN, the K-BERT model is first trained on the training data. The K-BERT model is then evaluated on the validation and test sets. During the prediction phase, kNN finds the k most similar instances in the training set to a given test sample. The classification labels of these instances are then used to predict the label of the test sample. The K-BERT model is also used to predict the given test sample.

The final prediction is obtained by combining the results of the kNN and K-Bert models using a weighted sum. This is done by applying a sigmoid function to the individual scores predicted by each model and taking their weighted average. The category with the highest score is chosen as the final prediction.

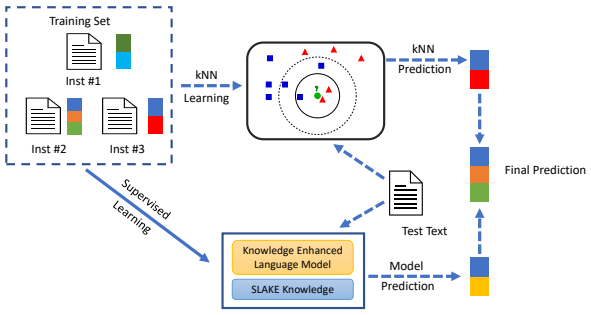

Figure 1: The overall flow of obtaining predictions through the information from nearest neighbors and language model.

2.2 Graph-NN Enriched Language Model

The integration of Graph Neural Networks (GNNs) with the feature level of the data has been shown to be effective in utilizing the feature information of the neighboring nodes. One of the widely used GNN networks is Graph Attention Network (GAT). In the early stages of data processing, a knowledge graph is added to the dataset and GAT is employed to construct a graph between nodes. The construction of this graph is achieved through the utilization of heuristic rules where edges are added to instances that belong to the same label category.

During the training process, the representation of each node obtained using K-BERT and GAT is subsequently employed to aggregate the representations of the neighboring nodes. The classification result of each instance is then predicted using GAT. In the prediction process, the graph is composed by obtaining the result through K-BERT and then using the result to connect nodes that belong to the same category. Finally, GAT is employed to aggregate the representations of the neighboring nodes to predict the classification result.

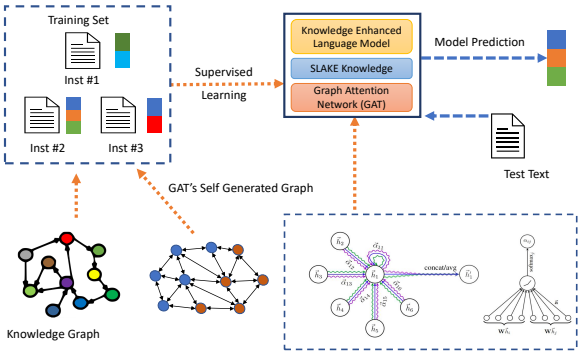

Figure 2: The overall methodology of obtaining prediction using an aggregate representation of language model and neighboring nodes.

2.3 Domain-specific Medical Knowledge

In order to add medical domain-specific knowledge to K-BERT, the triples are injected into sentences as domain knowledge. We leverage the SLAKE (Liu et al., 2021) to get triplets to be injected in K-BERT. SLAKE, a large multilingual dataset, contains rich ensemble semantic labels and a new structural medical knowledge base. For the relation extraction task, we chose SLAKE in the English language, whereas for the classification tasks, we used SLAKE in the Chinese language.

2.4 Datasets

In the experimentation, we have tested the performance of K-BERT + kNN and K-BERT + GNN in two downstream tasks–relation extraction and classification tasks.

Relation Extraction Datasets: For relation extraction task, we use GAD (Bravo et al., 2015), EU-ADR (Van Mulligen et al., 2012), and i2b2 (Uzuner et al., 2011) datasets. The Genetic Association Database (GAD) is a collection of studies that examine the association between genetic variations and the risk of developing complex diseases and disorders in humans (Bravo et al., 2015). Similarly, the EU-ADR corpus has been annotated for drugs, disorders, genes, and their interrelationships (Van Mulligen et al., 2012). i2b2 dataset is a dataset of patient medical problems, treatments, and tests used in the 2010 i2b2/VA Workshop on Natural Language Processing Challenges for Clinical Records (Uzuner et al., 2011).

Classification Tasks Datasets: For the classification task, we use cMedIC and cMedTC datasets (Zhang et al., 2020) in the Chinese language. cMedTC dataset consists of biomedical texts with multiple labels. Similarly, the cMedIC dataset consists of queries with three intent labels (no intention, weak intention, and firm intention).

3 Results and Discussion

We use accuracy as a measure to evaluate the performance of our algorithms for both downstream tasks (relation extraction and text classification). From Table 1, it can be seen that the integration neighbor information has improved performance in relation extraction. Integration of KNN into K-BERT has significantly improved the model performance significantly, with an 8% increase in performance with both the GAD and EU-ADR datasets. Similarly, there is an increase of 6.4% in performance for the i2b2 dataset when KNN is integrated into K-BERT. Integration of GNN to create the aggregated representations has helped K-BERT to improve its performance significantly. There is an increased performance across all the datasets using the aggregated representations from GNN and K-BERT.

	GAD	EU-ADR	i2b2
K-BERT	0.634	0.807	0.814
K-BERT + KNN	0.687	**0.871**	0.866
K-BERT + GNN	**0.696**	0.860	**0.875**

Table 1: Results of different models with various datasets in relation extraction domain. The results show that the addition of KNN and GNN to K-BERT improves the performances in relation extraction significantly.

Similar to the relationship extraction task, the performance of K-BERT with our approach to integrating additional information has shown better performance than the base K-BERT model. From Table 2, it can be observed that adding information on neighboring nodes has improved the accuracy by 1% to 2%. The performance improvement of 1% to 2% is very important in the medical domain, where the decisions impact the lives of people.

	cMedIC	cMedTC	Δ w.r.t. K-BERT CMedIC	Δ w.r.t. K-BERT CMedTC
K-BERT	0.927	0.609	-	-
K-BERT + KNN	0.939	0.615	**+1.3%**	**+0.99%**
K-BERT + GNN	**0.941**	**0.621**	**+1.51%**	**+1.97%**

Table 2: Results of different models in datasets related to classification tasks. The results show that there is around a 1% to 2% increment in performance by integrating GNN and KNN.

The results show that the base model K-BERT improved performance by integrating supplementary information. The classification and relation extraction tasks are important in the medical domain. The ability of our proposed methodology to get improved performance can also be adapted to other tasks in the medical domain.

4 Conclusion

In this paper, we propose a methodology for integrating additional information, such as neighboring nodes and instances, to improve the performance of a language model in the medical domain. The additional information also provides models with the ability to learn beyond contextual information. The methodology proposed in our work is generic and can be adapted to multiple tasks and domains. The work can be extended to solving other critical problems in the medical domain like report generation, medical dialogue generation, etc. It would also be interesting to integrate both nearest neighbor and graph information into the language model and evaluate the model performance. Another important future direction can be exploring how adding information in our framework contributes to the explainability of the model. Overall, our proposed methodology shows promising results and has the potential to enhance the performance and interpretability of language models in various domains.

Limitations

While the proposed methods offer promising results in improving the performance of knowledge graph-based language representation models, there are some limitations to this work that should be noted. Firstly, the experiments were conducted on a limited number of datasets, and the results may not be generalized to other datasets or domains. Therefore, further experiments are needed to validate the effectiveness of the proposed methods on a broader range of datasets and NLP tasks. Secondly, the proposed methods require additional computation and may increase the complexity of the models. Therefore, it is important to consider the trade-off between performance improvement and computational cost when applying these methods in real-world applications. Lastly, while the proposed methods address some of the limitations of existing knowledge graph-based language representation models, they still may not capture all the contextual relationships and nuances of natural language, leading to potential semantic errors and reduced accuracy. Therefore, it is essential to continue exploring new approaches and techniques to further improve the performance of NLP models.

References

Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Hai Ya Lu, Gnana Bharathy, and Mukesh Prasad. 2023. Explainable hybrid word representations for sentiment analysis of financial news. *Neural Networks*, 164:115–123.

Surabhi Adhikari, Surendrabikram Thapa, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2021. A comparative study of machine learning and nlp techniques for uses of stop words by patients in diagnosis of alzheimer's disease. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16:1–17.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Nora Kassner and Hinrich Schütze. 2020. Bert-knn: Adding a knn search component to pretrained language models for better qa. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3424–3430.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Linyang Li, Demin Song, Ruotian Ma, Xipeng Qiu, and Xuanjing Huang. 2021. Knn-bert: fine-tuning pretrained models with knn classifier. *arXiv preprint arXiv:2110.02523*.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13362–13370.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Jiachang Liu, Qi Zhang, Chongyang Shi, Usman Naseem, Shoujin Wang, and Ivor Tsang. 2023. Causal intervention for abstractive related work generation. *CoRR*, abs/2305.13685.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, pages 507–516. PMLR.

Zhibin Lu, Pan Du, and Jian-Yun Nie. 2020. Vgcn-bert: augmenting bert with graph embedding for text classification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*, pages 369–382. Springer.

Usman Naseem, Ajay Bandi, Shaina Raza, Junaid Rashid, and Bharathi Raja Chakravarthi. 2022a. Incorporating medical knowledge to transformer-based language models for medical dialogue generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 110–115.

Usman Naseem, Adam G Dunn, Matloob Khushi, and Jinman Kim. 2022b. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *BMC bioinformatics*, 23(1):1–15.

Usman Naseem, Matloob Khushi, Shah Khalid Khan, Kamran Shaukat, and Mohammad Ali Moni. 2021a. A comparative analysis of active learning for biomedical text mining. *Applied System Innovation*, 4(1):23.

Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. 2021b. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Nazneen Fatema Rajani, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher, and Caiming Xiong. 2020. Explaining and improving model behavior with k nearest neighbor representations. *arXiv preprint arXiv:2010.09030*.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Aditya Shah, Surendrabikram Thapa, Aneesh Jain, and Lifu Huang. 2023. Adept: Adapter-based efficient prompt tuning approach for language models. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. 2012. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884.

Eric Wallace, Shi Feng, and Jordan Boyd-Graber. 2018. Interpreting neural networks with nearest neighbors. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 136–144.

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, Bo Long, et al. 2023. Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning*, 16(2):119–328.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*.