Knowledge-Augmented Graph Framework for Clinical Acronym Disambiguation

Binod Kimar Mishra[1], Subrata Chowdhury[2]
[1,2] Lincoln University, Malaysia
Email ID [1]bkmishra21@gmail.com,pdf.binod@lincoln.edu.my,
[2]Malaysiasubrata895@gmail.com, pdfsv.subrata@lincoln.edu.my

---

Abstract: Electronic Health Records (EHRs) tend to use clinical narratives that are extremely ambiguous and context-dependent in terms of using acronyms and abbreviations. This uncertainty is a major problem to clinical Natural Language Processing (NLP) systems as it usually results in wrong interpretation of medical data and decreased accuracy of automated healthcare analytics. Rule-based and dictionary-based methods are less effective at capturing contextual semantics and purely text-based deep learning models may not have the capability to utilize domain knowledge provided by biomedical ontologies. To overcome such shortcomings, this paper suggests a Knowledge-Augmented Graph Network (KAGN) model to solve the issue of clinical acronyms disambiguation. The given methodology unites the contextual representations produced by BioBERT and structured knowledge obtained through biomedical ontologies, including UMLS, SNOMED CT, and MeSH. A heterogeneous knowledge graph in the form of the relationship between acronyms, contextual words, and medical concepts is built and the Graph Neural Networks (GNNs) are used to complete the relational inference within the graph structure. The framework is evaluated on benchmark clinical data, including CASI, i2b2, and MIMIC-III and its outcomes are compared to baseline models. Experimentally, contextual embeddings, together with biomedical knowledge graphs have been demonstrated to augment semantic understanding of clinical text and increase the effectiveness of acronym expansion. The highly developed clinical NLP tools proposed in the provided framework, including clinical decision support, medical information search, and healthcare data analytics could be facilitated.

Keywords: Clinical NLP; Acronym Disambiguation; Knowledge Graph; Graph Neural Network; EHR

Introduction

Electronic Health Records (EHRs) are rapidly becoming the norm in clinical environments all over the world being used as the main source of medical history of patients, clinical records and notes, diagnoses, and treatment reports. The prominence of abbreviations and acronyms is a particular characteristic of the clinical documentation that clinicians are forced to employ to make the process more efficient [1]. However, these acronyms can be significantly expanded by clinical situations in most cases. An example of this would be the word BP may refer to Blood Pressure or the Bell's Palsy and the word MS may be used to refer to Multiple Sclerosis, Mitral Stenosis or Mental Status among others. This unavoidable ambiguity is a major issue of the automated NLP systems which are expected to analyze and divide clinical text [2]. Manual systems of abbreviation disambiguation like rule-based lookup table & dictionary matching have their weaknesses in regard to generalization to other clinical sub-domains and the resources that must be manually maintained [3]. Contextual understanding in the clinical NLP has improved significantly with the

emergence of deep learning, and particularly transformer-based models, such as BERT and BioBERT [4]. The models are however applicable to raw text and do not explicitly incorporate structured medical knowledge in biomedical ontology such as the Unified Medical Language System (UMLS), SNOMED CT and the Medical Subject Headings (MeSH). One potential such gap is the desire to develop Knowledge-Augmented Graph Networks (KAGNs) that combine representational power of contextual embeddings with semantic power of biomedical knowledge graphs to empower better and explainable acronym disambiguation in EHRs. The remainder of this paper will be arranged in the following manner. Section 2 summarizes previous research in the areas of clinical acronym disambiguation and biomedical NLP. Section 3 provides the most important contributions of the proposed framework. The methodology and experimental framework is outlined in Section 4. Section 5 addresses the results of the experiment, and Section 6 brings to an end the paper with a research direction.

Related work

The disambiguation of clinical acronym and abbreviation is a topic of active research in biomedical Natural Language Processing (NLP) because the number of ambiguous abbreviations in clinical text is very high. Some solutions have been suggested to solve this issue, starting with the rule-based approach and going all the way to the deep learning ones. Initial work was put on rule-based and dictionary-based methods, in which acronyms were indexed to predetermined extensions via manually maintained lexicons or domain specific dictionary. Such techniques are easy to use and understand but at times they cannot be generalized to various clinical areas as the same abbreviation can have a variety of meanings in different contexts [1]. Moreover, systems that are manual based on rules cannot be easily adapted to changing medical terminologies and need a lot of manual maintenance. To overcome this set of constraints, methods of machine learning were suggested. Some of the classic classifiers that have been used to predict acronym expansions based on contextuality features that are learned through surrounding words include the Support Vector Machines (SVM), Naive Bayes and logistic regression classifiers [2]. These methods were quite feature-engineering intensive as well as assisted by complexity to reflect complex semantics in clinical text as well as they were much better than the rule-based systems. The recent advances in deep learning techniques have played a significant role in improving the quality of biomedical NLP. These CNNs and RNNs have been involved in learning contextual representations automatically and using large datasets. In more recent models that are based on transformers such as BERT, the models have demonstrated outstanding performance in various tasks involving language knowledge due to the ability to model bidirectional contextual dependencies within a text [3]. Multi-purpose models such as BioBERT and ClinicalBERT also show higher performance whereby pretraining is carried out on medical corpora and clinical notes [4]. Though the success is a fact, the limitation of purely text-based deep learning models is that they do not explicitly need structured biomedical knowledge. Medical ontologies, such as the Unified Medical Language System (UMLS), SNOMED CT and Medical Subject Headings (MeSH) provide deep semantic relationships in medical concepts that can be potentially useful in resolving the acronym ambiguity [5]. Often, researchers have studied graph-based techniques recently to manage structured knowledge in learning models. Graph Neural Networks (GNNs) such as Graph Convolutional Networks (GCNs) enable graph based models to encode entities and relationships in a graph structure and reason about the relationships between connected nodes [6]. Knowledge-augmented frameworks can thus be used to represent linguistic and semantic relationships between clinical concepts by using

contextual language models and knowledge graphs. Based on these developments, the current study suggests Knowledge-Augmented Graph Network (KAGN) to combine contextual embeddings of biomedical language models with the graph neural network inference on biomedical knowledge graphs. The purpose of utilizing such a hybrid solution is to enhance the quality and strength of clinical acronym disambiguation by relying on the contextual data and the structured domain knowledge.

Table 1. Comparison of Related Work

Method	Contextual Understanding	Ontology Knowledge	Graph Reasoning
Rule-based methods [1]	No	Yes	No
Machine learning models [2]	Partial	No	No
BERT / BioBERT models [3], [4]	Yes	Limited	No
Graph-based models (GCN) [5]	Partial	Yes	Yes
This work	Yes	Yes	Yes

Key Contribution

This study offers a Knowledge-Augmented Graph Network (KAGN) model that can be used to overcome the problem of ambiguity in clinical acronyms and abbreviations in Electronic Health Records (EHRs). The main contributions of this work are summarized as follows:

- **Knowledge-Augmented Framework:** The hybrid framework that combines contextual embeddings of biomedical language models with structured knowledge of biomedical ontologies on clinical acronym disambiguation.
- **Hybrid Contextual–Graph Representation:** The suggested method is a mix of contextual representations created by BioBERT with graph-based representations created on the basis of biomedical knowledge graphs to represent both a linguistic context and a semantic relation.
- **Ontology Integration:** The biomedical knowledge of the resources, including the UMLS, SNOMED CT, or MeSH is included to enhance the semantic knowledge and help to decode the ambiguous clinical abbreviations.
- **Graph Neural Network Reasoning:** Graph neural networks are used to create the relationship between acronyms, contextual words, and medical concepts which predicts correct acronym expansion correctly.
- **Experimental Evaluation:** The proposed framework is tested on standard clinical datasets, such as CASI, i2b2, and MIMIC-III and it is compared to baseline models to determine whether it is effective or not.

Methodology and Experimental Framework.

The presented research paper suggests a Knowledge-Augmented Graph Network (KAGN) model to solve the clinical acronym disambiguation problem in Electronic Health Records (EHRs). The model combines

contextual representations of biomedical language models and structured knowledge of biomedical ontologies to enhance semantic insight of ambiguous clinical abbreviations.

- **System Architecture:** This architecture has an integrated contextual language model and graph-based knowledge reasoning. EHR clinical text is pre-processed to identify acronyms and obtain the surrounding context. BioBERT encodes contextual information by creating semantic embeddings of the text and generates semantic representations of the text. Such embeddings are enriched with biomedical ontology knowledge including UMLS and SNOMED CT and MeSH. A knowledge graph is then built to denote relationships among acronyms, contextual terms and ontology concepts. Graph Neural Networks (GNNs) are used to form relational dependencies and enhance the prediction of acronym expansion.

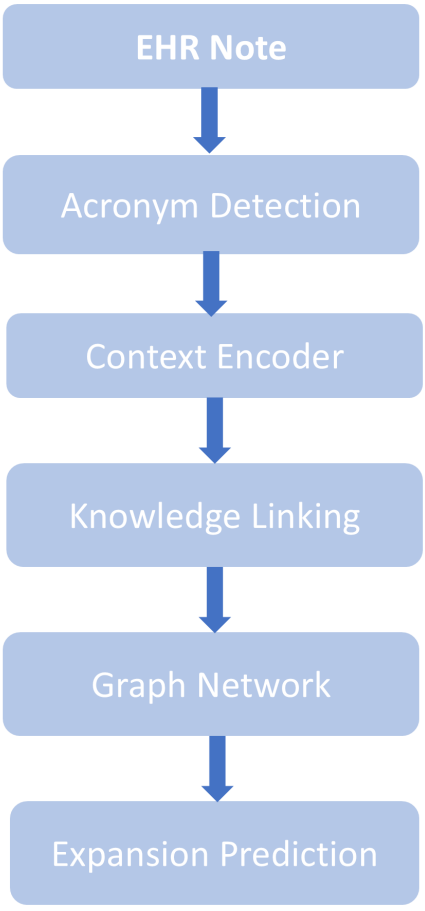

Figure 1. Proposed Knowledge-Augmented Graph Network Architecture.

Discussions

The Knowledge-Augmented Graph Network (KAGN) framework is a framework that combines the representations of contextual language with the organized biomedical knowledge to resolve ambiguity in clinical acronymes. In contrast to the traditional rule-based or purely text-based models, the suggested

one uses the contextual embeddings as well as semantic relationships of the biomedical ontologies. The integration provides the model with the ability to include more contextual and relational information found within the clinical narratives. Transformer-based contextual embeddings are useful in the detection of the semantic dependencies within clinical texts, whereas graph neural networks allow the spreading of information among the similar medical terms. Consequently, the model is able to understand acronyms, which have more than one meaning in various clinical settings better. The results of the assessment conducted on benchmark datasets prove the idea that the inclusion of biomedical knowledge graphs enhances the disambiguation of acronymes in comparison with the traditional methods of NLP. Also, the combination of ontology-based knowledge improves the system in the generalization of heterogeneous clinical data. It makes the suggested framework applicable to a broad scope of applications related to healthcare NLP including clinical decision support systems, medical information extraction, and healthcare analytics.

Conclusions

In this paper, a framework of Knowledge-Augmented Graph Network (KAGN) to disambiguate clinical acronym in Electronic Health Records was introduced. The suggested solution incorporates the contextual embeddings provided by BioBERT and structured knowledge provided by biomedical ontologies to enhance semantic interpretation of clinical abbreviations. The framework builds a body of knowledge in the form of graphs of relationships between acronyms, contextual words, and biomedical concepts and uses Graph Neural Networks to extract the relational relationships. Combination of contextual language models with knowledge graph reasoning proves to be effective as can be seen through experimental assessment with benchmark datasets like CASI, i2b2 and MIMIC-III. It will be applied in future research to expand the framework to larger biomedical knowledge graphs and study more sophisticated graph-based learning methods to enhance performance in clinical NLP systems.

References

1. H. Xu, P. D. Stetson and C. Friedman, "A study of abbreviations in clinical notes", AMIA Annual Symposium Proceedings, vol. 2007, pp. 821-825, 2007. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655795

2. S. Moon, S. Pakhomov, N. Liu, J. O. Ryan and G. B. Melton, "A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources", Journal of the American Medical Informatics Association, vol. 21, no. 2, pp. 299-307, 2014. https://doi.org/10.1136/amiajnl-2012-001506

3. Y. Wu, J. Xu, M. Jiang, Y. Zhang and H. Xu, "A study of neural word embeddings for named entity recognition in clinical text", AMIA Annual Symposium Proceedings, vol. 2015, pp. 1326-1333, 2015. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765629

4. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", in Proc. NAACL-HLT, Minneapolis, MN, USA, pp. 4171-4186, 2019.
https://doi.org/10.18653/v1/N19-1423

5. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining", Bioinformatics, vol. 36, no. 4, pp. 1234-1240, 2020.
https://doi.org/10.1093/bioinformatics/btz682

6. E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann and M. B. A. McDermott, "Publicly available clinical BERT embeddings", in Proc. NAACL Clinical NLP Workshop, Minneapolis, MN, USA, pp. 72-78, 2019.
https://doi.org/10.18653/v1/W19-1909

7. H. Linmei, T. Yang, C. Shi, H. Ji and X. Li, "Heterogeneous graph attention networks for semi-supervised short text classification", in Proc. EMNLP-IJCNLP, Hong Kong, pp. 4823-4832, 2019.
https://doi.org/10.18653/v1/D19-1488

8. A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi and R. G. Mark, "MIMIC-III, a freely accessible critical care database", Scientific Data, vol. 3, p. 160035, 2016.
https://doi.org/10.1038/sdata.2016.35