Life Sciences and Facility Management

Master‘s Thesis 2023
Applied Computational Life Sciences

Ontology-Aware Biomedical Text Classification

Author: Jasmin Saxer
Supervisors: Dr. Ahmad Aghaebrahimian, Dr. Manuel Gil
Research Centre of Bioinformatics, Institute of Computational Life Sciences (ICLS)

Introduction

Text Classification has become an exciting field in Machine Learning and Natural Language Processing. In the biomedical field, Text Classification is used to tackle the ever-increasing number of scientific publications. The biomedical domain is rich in structured knowledge, such as ontologies. Ontologies are collections of semantic knowledge on a specific domain and are proved to be useful in many Natural Language Processing tasks, including Text Classification. The contribution of this work is to incorporate biomedical ontologies into Text Classification.

Materials and Methods

The TextGCN model from L. Yao, Mao, and Luo, 2018 and a simple Convolutional Neural Network model were used as the baselines. Ontological information was incorporated into the TextGCN model, by adding word-to-ontology and document-to-ontology connections into the text document graph. For the Convolutional Neural Network model pretrained UMLS embeddings were used as input to incorporate ontology information. Further, a new architecture combining the TextGCN with the Convolutional Neural Network model was proposed.

The datasets Ohsumed, CRAFT and MedOBO were utilized to evaluate the new models

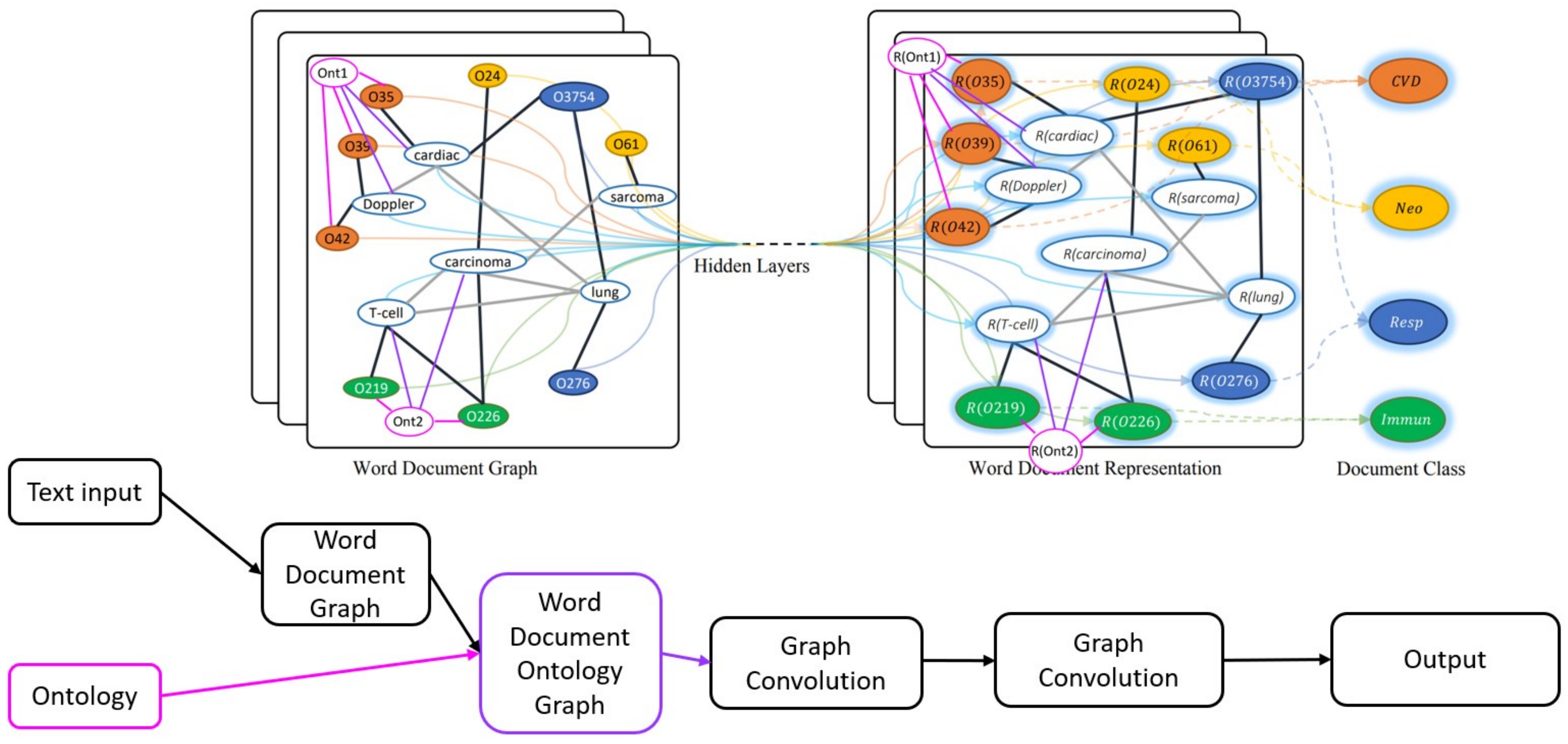

Figure 1: The TextGCN model architecture with adding ontology connections as a flow chart and as a graph. The word document graph figure was adopted from (L. Yao, Mao, and Luo, 2018).

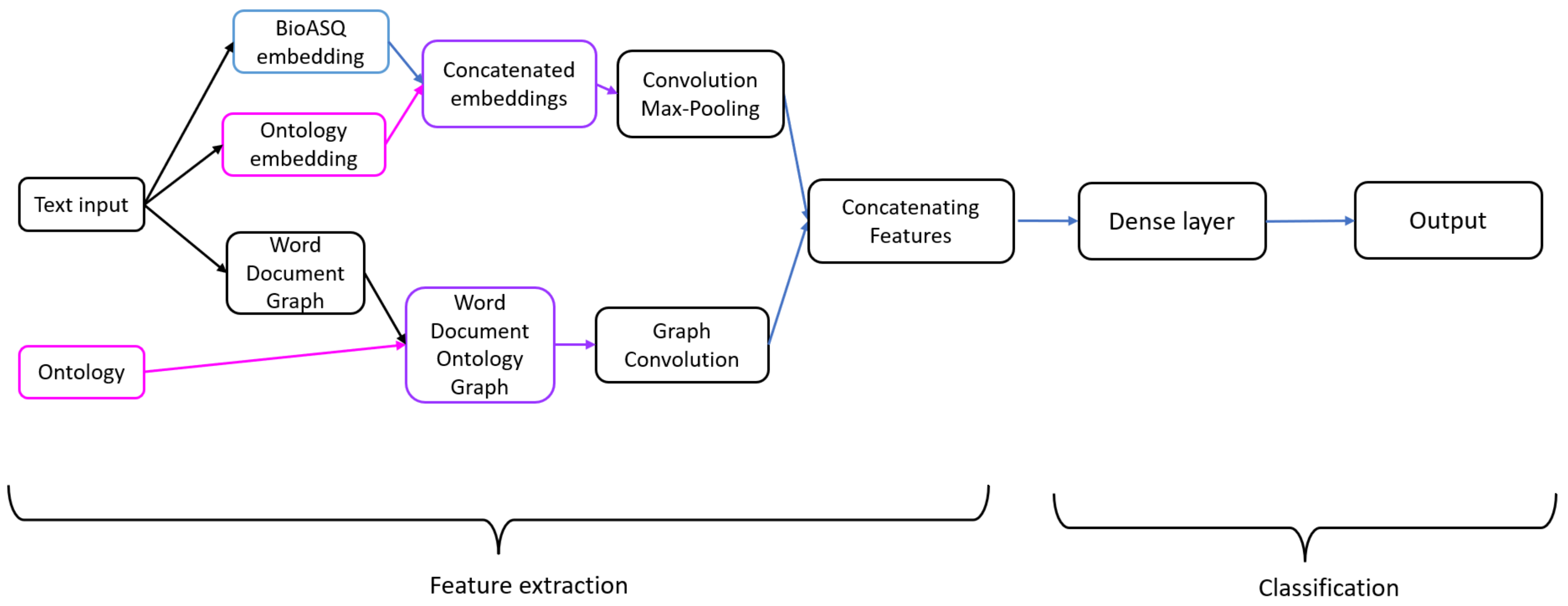

Figure 2: The new TextGCN-CNN model architecture with adding ontology connections as a flow chart.

Results and Discussions

In some experiments, classification was improved by incorporating existing biomedical ontologies.

Incorporating biomedical ontology information into Text Graph Convolutional network (TextGCN) increased the classification of the MedOBO. However, it did not improve the classification of the Ohsumed and CRAFT dataset. Incorporating pretrained biomedical ontology embeddings from UMLS into Convolutional neural network (CNN) did not improve the classification. Using the Ohsumed, CRAFT, and MedOBO datasets with the new model architecture called TextGCN-CNN did not improve the classification.

	Macro			Micro		
Description	Precision	Recall	F1 Score	Precision	Recall	F1 Score
TextGCN Baseline (BS)	0.2152	0.1614	0.1546	0.6426	0.7059	0.6727
BS + Ontology (all)	0.2092	0.1603	0.1543	0.6468	0.7031	0.6738
BS + Ontology ($\pm 1std$)	0.2215	0.1589	0.1539	0.6490	0.6998	0.6734

Table 1: Results of TextGCN and additional information from ontologies on MedOBO dataset. Ontology refers to the TextGCN with ontology connections. All refers to including all ontologies from the MedOBO development dataset. ±1std refers to including only the ontologies in the range of ±1std.

Conclusion and Outlook

The Conclusion is that incorporating biomedical ontologies with Text Classification has the potential to improve the system performance and that more research on this topic is needed.

There are many different aspects that could be further explored, especially in terms of datasets, ontology integration, and the model architecture.