Word and Phrase Features in Graph Convolutional Network for Automatic Question Classification

Junyoung Lee, Ninad Dixit*, Kaustav Chakrabarti*, and S. Supraja
*Equal contribution

Paper on arXiv

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS
IJCNN2025
30 JUNE - 5 JULY 2025 | ROME, ITALY
INTERNATIONAL NEURAL NETWORK SOCIETY

Background

Why Automatic Question Classification?

- Categorize questions for learning diagnosis and analytics
- Provide foundation for more complex tasks such as information retrieval and question answering [1]

Why Graph Convolutional Network (GCN)?

- Graphs can capture intra-question relationships, such as syntactic dependencies, semantic similarities, and proximity measures
- GCNs provide effective localized aggregation of node features into embeddings for downstream tasks

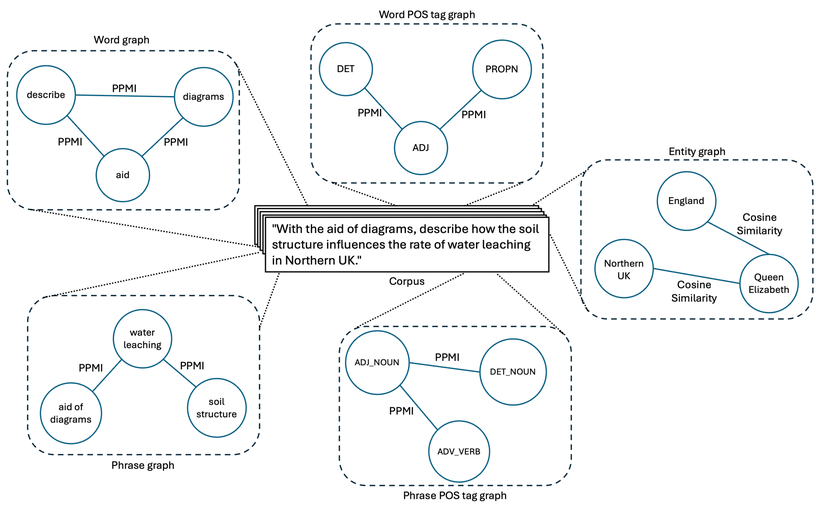

Figure 1: Feature extraction and graph construction

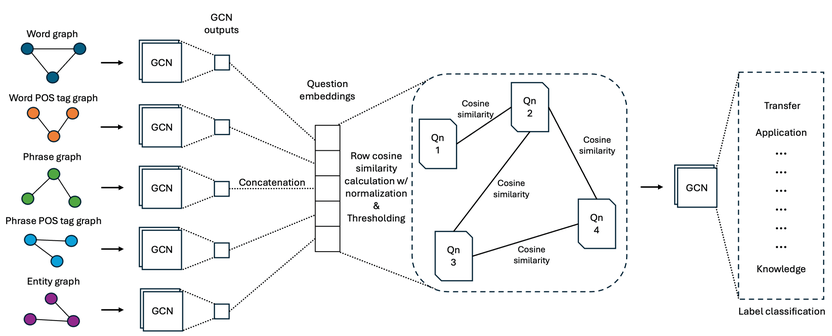

Figure 2: Graph propagation

Challenges

- Limited information in text compared to document-level counterparts
- Lack of question text-specific feature extraction and graph construction methods - current text graph construction methods work with word-level nodes only

Approach

PQ-GCN: adapting text graphs for question classification

- Capturing meaningful chunks (phrases) as nodes
- Disambiguation and contextualisation of word-level information

Feature Extraction and Graph Construction

Node set construction:

- Words: whitespace tokenization
- Phrases: noun-/verb-phrase regex matching
- Word-/phrase-level POS tags: default set from NLTK
- Named entities: NELL knowledge base

Edge set construction:

- Words/phrases: pointwise mutual information (PMI)

$$\text{PMI}(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

- Named entities: cosine similarity

Node features from pre-trained models; one-hot encoding otherwise

	Node Type	Node Feature	Edge Weight
Word PPMI	Word	Word2Vec	PMI value
Word POS	POS tag	-	PMI value
Phrase PPMI	Phrase	PhraseBERT	PMI value
Phrase POS	Phrase POS tag	-	PMI value
Named Entities	Named entity	TransE	Cosine similarity

Model Architecture

Considerations

- Incorporating phrase-based features as modular add-on
- Consolidating 5 homogenous graphs into single question embedding

Taking inspiration from SHINE [2],

Graph propagation

- Each graph is passed through their own 2-layer GCN
- Each GCN output is then concatenated to form question embeddings
- Dynamic question graph is created by taking questions as nodes and calculating cosine similarity between question embeddings as edges
- Dynamic question graph is passed through a final 2-layer GCN and a linear layer for label classification

Evaluation

Datasets

- NU: cognitive complexities - 596 questions
- ARC: reasoning capabilities - 279 questions
- LREC: expected answer types - 344 questions
- Bloom: educational objectives - 2522 questions
- TREC: question topics - 5952 questions

Baselines

- Non-graph models: CNN, Bi-LSTM
- Graph models: TextGCN, Text-Level_GNN, HyperGAT, TensorGCN, SHINE, ME-GCN, InducT-GCN

BASELINE COMPARISONS FOR MACRO-AVERAGED F1, PRECISION, AND RECALL SCORES ACROSS DATASETS
BEST F1, PRECISION, AND RECALL SCORES ARE IN **BOLD** AND THE SECOND BEST SCORES ARE UNDERLINED FOR EACH DATASET.

Model	Metric	NU	ARC	LREC	Bloom	TREC
CNN	F1	0.085	0.226	0.198	**0.752**	0.782
	Precision	0.048	0.171	0.141	**0.763**	0.783
	Recall	0.333	0.333	0.333	**0.750**	**0.790**
Bi-LSTM	F1	0.607	0.564	0.480	0.425	0.653
	Precision	0.606	0.554	0.511	0.465	0.692
	Recall	0.620	0.598	0.495	0.422	0.642
TextGCN	F1	0.722	0.694	0.671	0.663	0.730
	Precision	0.715	0.687	0.677	0.652	0.675
	Recall	**0.735**	0.735	0.679	0.680	0.780
Text-Level-GNN	F1	0.185	0.200	0.404	0.287	0.623
	Precision	0.694	0.362	0.424	0.250	0.651
	Recall	0.183	0.275	0.424	0.344	0.606
HyperGAT	F1	0.715	0.372	0.745	0.173	0.678
	Precision	0.733	0.386	**0.766**	0.176	0.673
	Recall	0.703	0.374	0.737	0.178	0.691
TensorGCN	F1	0.412	0.499	0.566	0.107	**0.805**
	Precision	0.421	0.557	0.692	0.147	0.851
	Recall	0.450	0.604	0.575	0.157	0.775
SHINE	F1	0.560	0.610	0.620	0.459	0.560
	Precision	0.553	0.600	0.620	0.461	0.568
	Recall	0.583	0.623	0.627	0.473	0.613
ME-GCN	F1	0.632	0.603	0.601	0.607	0.659
	Precision	**0.738**	0.601	0.607	0.629	0.718
	Recall	0.619	0.620	0.608	0.592	0.634
InducT-GCN	F1	0.661	0.667	0.649	0.533	0.688
	Precision	**0.738**	0.649	0.648	0.754	0.753
	Recall	0.639	0.711	0.654	0.478	0.671
PQ-GCN	F1	**0.724**	**0.712**	**0.751**	0.672	0.801
	Precision	0.723	**0.695**	0.754	0.692	**0.882**
	Recall	0.727	**0.750**	**0.749**	0.662	0.777

Analysis and Conclusion

- Best macro-averaged F1 scores: NU (0.724), LREC (0.751), TREC (0.801) with consistent high precision and recall
- At least 0.1 point improvement from base SHINE model

Note: CNN's unusual performance on Bloom - due to distinct and mutually exclusive label class-specific keywords in corpus

Why PQ-GCN?

- Promising classification performance compared to baselines across datasets
- Demonstrates effectiveness of alternative feature extraction methods to enhance existing text graph-based tasks

[1] V. A. Silva, I. I. Bittencourt, and J. C. Maldonado, "Automatic question classifiers: A systematic review," IEEE Transactions on Learning Technologies, vol. 12, pp. 485–502, Oct 2019.
[2] Y. Wang, S. Wang, Q. Yao, and D. Dou, "Hierarchical heterogeneous graph representation learning for short text classification," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Online and Punta Cana, Dominican Republic), pp. 3091–3101, Association for Computational Linguistics, Nov. 2021.