Text Classification Using a Word-Reduced Graph

Authors: Hiromu Nakajima, Minoru Sasaki

Presenter: Hiromu Nakajima

Graduate School of Science and Engineering, Ibaraki University(Japan)

Presenter's email: 22nm738g@vc.Ibaraki.ac.jp

茨城大学
Ibaraki University

IARIA

About presenter

Hiromu Nakajima received the bachelor's degree from the Ibaraki University in 2022. He is currently a master's student majoring in computer and information sciences at the Graduate School of Science and Engineering, Ibaraki University.

His research interest lies in artificial intelligence and NLP(particularly, text classification).

Outline

- Introduction
- RoBERTaGCN
- Method
- Experiment
- Result
- Discussion
- Conclusion and Future Work

Introduction

What's RoBERTaGCN ?

・Text classification method was proposed by Yuxiao Lin et al. in July 2021.

・This method Learns by inputting heterogeneous graphs of words and documents into GCN.

・This method achieves state-of-the-art in text classification tasks

Introduction

Problem of RoBERTaGCN

・RoBERTaGCN doesn’t consider relationships between documents in graph.

・Topic drift, the problem of documents on different topics being associated through words with multiple meanings, can occur.

We solved these problems by adding a cosine similarity value to the weights of edges between document nodes.

Introduction

Newly arising problem

・Increasing the number of weights increases the size of the graph and requires a lot of memory.

⬇

Purposes of this study

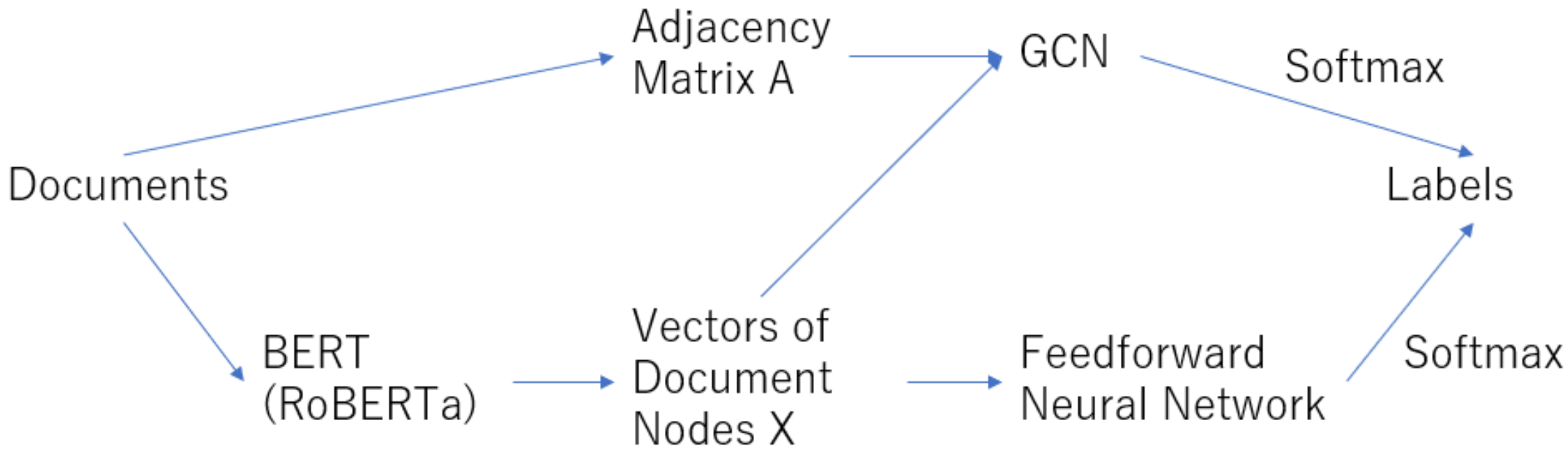
・Maintain classification performance while reducing memory size.

・Improve classification accuracy over conventional methods by utilizing the reduced memory and using larger trained models.

RoBERTaGCN

Linear Interpolation

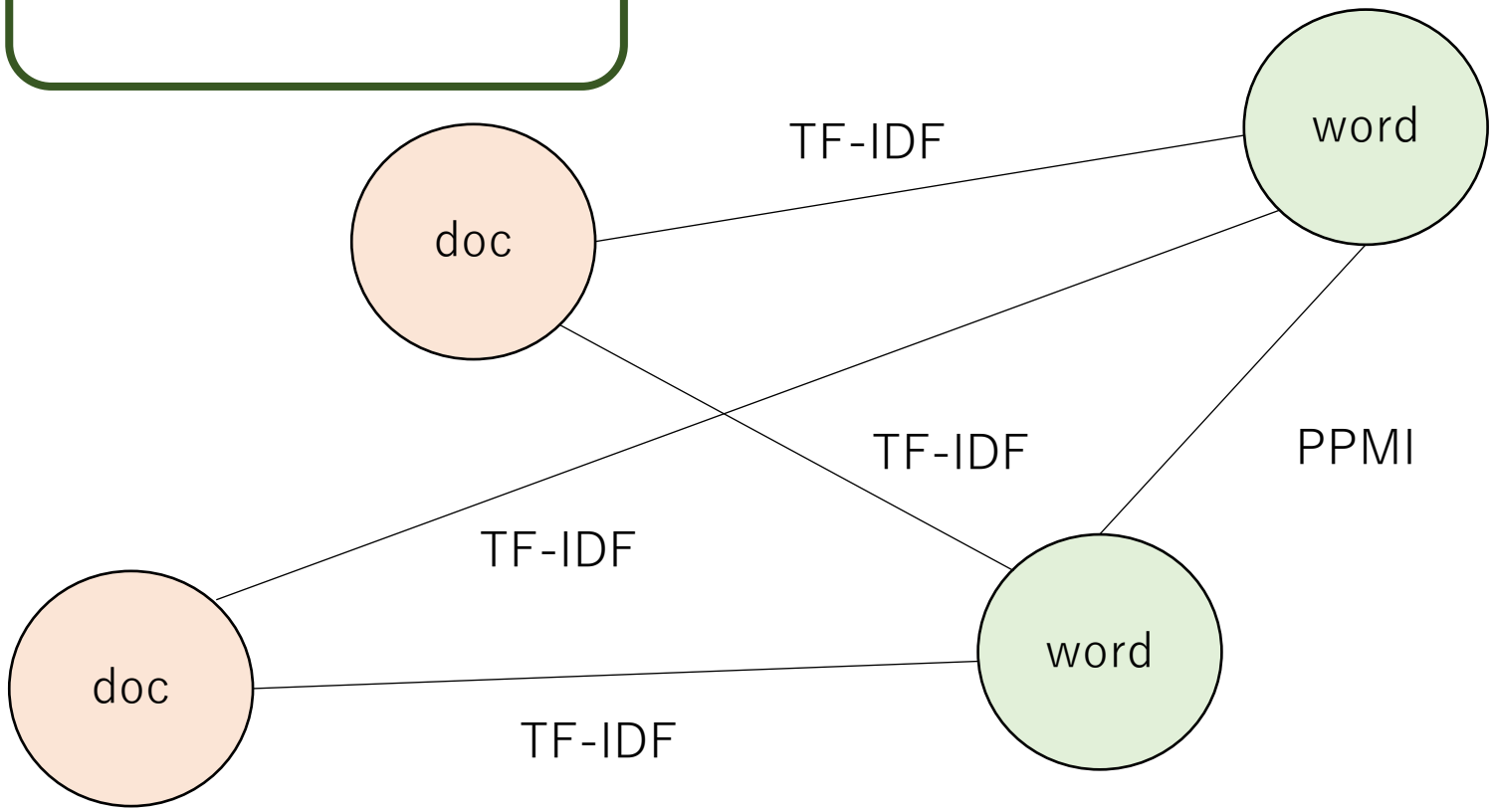
$$Z = \lambda Z_{GCN} + (1 - \lambda) Z_{BERT}$$



・A heterogeneous graph of words and documents is input to GCN and the document vector is input to Feedforward Neural Network to obtain the prediction results for both.

・Then, a linear interpolation of the two prediction results is computed and the result is adopted as the final prediction.

RoBERTaGCN



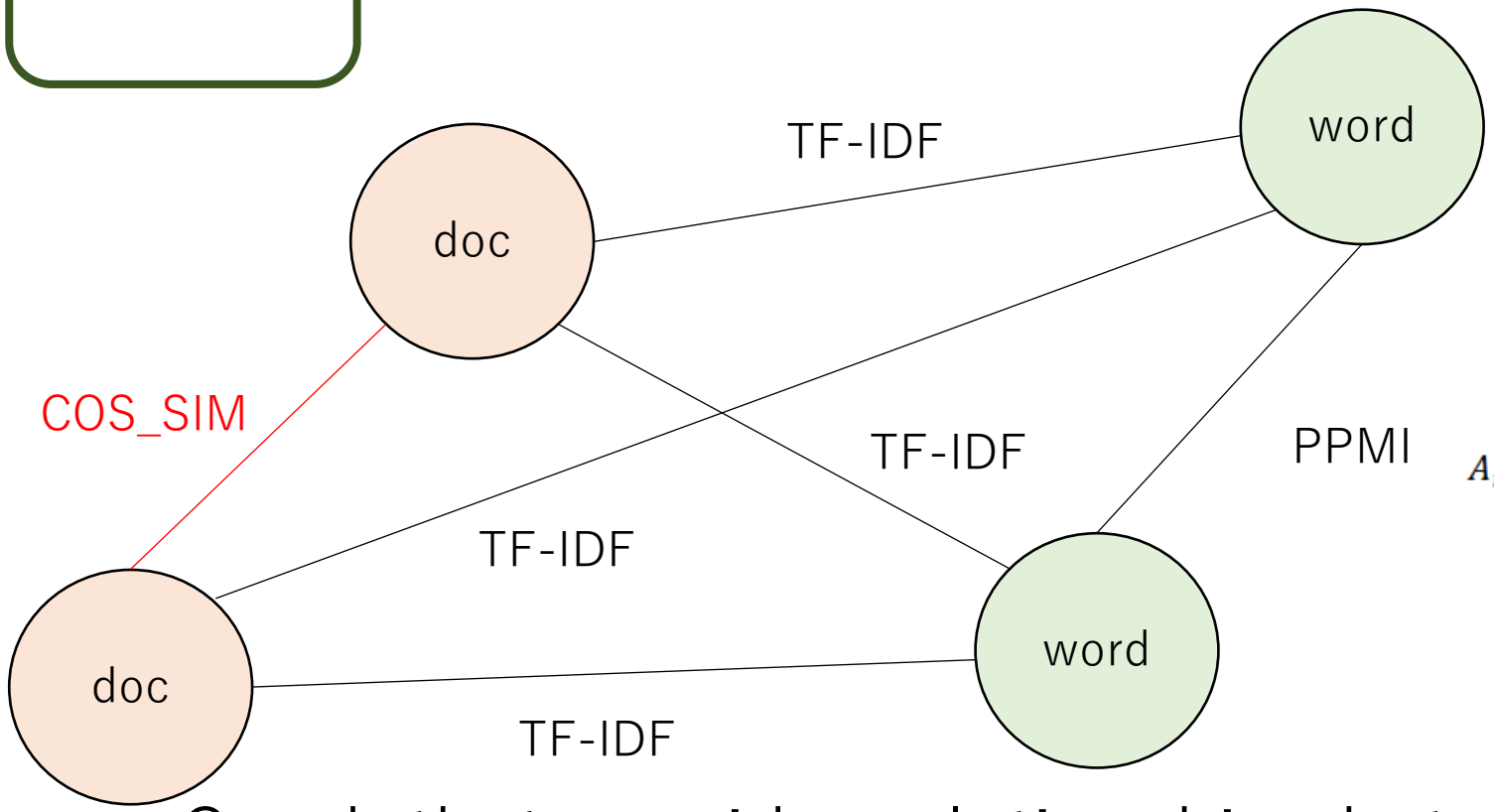
$$A_{i,j} = \begin{cases} PPMI(i,j), & i,j \text{ are words and } i \neq j \\ TF-IDF(i,j), & i \text{ is document}, j \text{ is word} \\ 1, & i = j \\ 0, & otherwise \end{cases}$$

This graph doesn't consider relationships between documents in graph.

PPMI：Co-occurrence of infrequent word pairs　ex.(beer,wine)>(he,has)

TF-IDF：Importance level of words in the documents

Method



$$A_{i,j} = \begin{cases} COS\_SIM(i,j), & i,j \text{ are documents and } i \neq j \\ PPMI(i,j), & i,j \text{ are words and } i \neq j \\ & df(i) > 1, df(j) > 1 \\ TF - IDF(i,j), & i \text{ is document}, j \text{ is word} \\ & df(j) > 1 \\ 1, & i = j \\ 0, & otherwise \end{cases}$$

df :Word Frequency

Graph that consider relationships between documents

We propose the graph structure that is more compact than conventional methods by removing words that appear in only one document.

Method

1） Tokenize documents

If the document has more than 510 words,
use 510 words from the beginning of the document.

2） Obtain CLS vectors

Input each tokenized document into BERT
and obtain a CLS vector at its final layer.

3） Calculate cosine similarity and add weights

If the cosine similarity exceeds a predetermined threshold,
it is added to the weights between document nodes.

Experiment

Experiment 1: Experiment to confirm the effectiveness of the graphs of the proposed method.

In Experiment 1, the classification performance of the proposed method using compact graphs was compared with other methods. The trained model used was roberta-base. Accuracy was used to evaluate the experiment. Positive is the label of the correct answer, negative is the label of the incorrect answer, and negative is all the remaining labels except the correct label.

Experiment 2: Experiment to check classification accuracy when changing to a larger trained model.

In Experiment 2, we take advantage of the memory savings and check the accuracy of the proposed method by applying a larger trained model. Specifically, the learned model is changed from roberta-base to roberta-large. λ and cosine similarity values are set to the same values as in Experiment 1.

Experiment

Dataset: 20NG(20-news-groups), R8, R52, Ohsumed,
MR(movie-review)

Stopwords and symbols were removed as preprocessing.

Dataset	Number of Documents	Average of Words	Training Data	Test Data
20NG	18846	206.4	11314	7532
R8	7674	65.7	5485	2189
R52	9100	69.8	6532	2568
Ohsumed	7400	129.1	3357	4043
MR	10662	20.3	7108	3554

Experiment

- Threshold values are shown in the table.
- Batch_size = 64
- Epochs = 70
- Dropout = 0.5
- GCN_lr = 0.001
- BERT_lr = 0.00001
- Parameter $\lambda = 0.7$

Dataset	Optimal Threshold Value
20NG	0.99
R8	0.975
R52	0.96
Ohsumed	0.965
MR	0.97

Result

	20NG	R8	R52	Ohsumed	MR
Text GCN	86.34	97.07	93.56	68.36	76.74
Simplified GCN	88.50	-	-	68.50	-
LEAM	81.91	93.31	91.84	58.58	76.95
SWEM	85.16	95.32	92.94	63.12	76.65
TF-IDF +LR	83.19	93.74	86.95	54.66	74.59
LSTM	65.71	93.68	85.54	41.13	75.06
fastText	79.38	96.13	92.81	57.70	75.14
BERT	85.30	97.80	96.40	70.50	85.70
RoBERTa	83.80	97.80	96.20	70.70	89.40
RoBERTa GCN	89.15	98.58	94.08	72.94	88.66
[5]	89.82	**98.81**	94.16	74.13	89.00
Experiment 1	**90.02**	98.58	**96.88**	73.53	89.65
Experiment 2	89.95	98.58	96.81	**76.08**	**91.50**

◎Ohsumed : 76.08%(1.95%↑)

◎MR : 91.50%(1.85%↑)

[5] shows the classification performance when using the graph structure consider relationships between documents.

Discussion

1)About the effectiveness of the proposed graph structure

・From these three tables, it can be seen that the graph of the proposed method reduces the number of edges by 1 to 20%.

・We believe that the reason why the accuracy was maintained even with a compact graph is because the words to be removed were limited to words that appear only in a single document.

・Words that appear in only one document do not propagate document topic information through the word node, and thus text classification performance is maintained even if those words are removed.

Dataset	Number of Words	Number of Words Removed
20NG	42757	755
R8	7688	225
R52	8892	245
Ohsumed	14157	851
MR	18764	8687

Dataset	Number of PPMI Edges	Number of Edges Removed
20NG	22413246	127662
R8	2841760	32954
R52	3574162	36138
Ohsumed	6867490	129938
MR	1504598	314950

Dataset	Number of TF-IDF Edges	Number of Edges Removed
20NG	2276720	755
R8	323670	225
R52	407084	245
Ohsumed	588958	851
MR	196826	8687

Discussion

2) Why did accuracy improve when the trained model was changed to a larger one?

・When the learned model was changed from roberta-base to roberta-large, the accuracy improved significantly.

・It is thought that the change to roberta-large improved the accuracy because it was able to acquire embedded representations that better reflect the characteristics of the documents.

Conclusion and Future Work

・This study proposes the text classification method using compact graphs in which words that appear only in one document are removed to solve the memory-consuming problem.

・Experiments confirmed that the proposed method can maintain the accuracy of the conventional method while saving a lot of memory.

・Experiments also showed that the acuracy of text classification improves when the learned model is changed to a larger one.

Future Work

・Comparing the accuracy with the proposed method when other features are used instead of cosine similarity

・Optimizing the parameter λ for each data

・Devise a method to extract sentences of documents exceeding 510 words