REPORT

Fine-Tuning LLMs with Knowledge Graphs

1. Problem Statement:

Given any description about a disease such as its definition or some synonyms used commonly by people, it's crucial to classify it into its one of the common disease types.

2. Domain Description:

The domain selected for the task is Biomedical domain, specifically considering the Human Disease Ontology of which numerous variational knowledge graphs are available at URL.

3. Defining the Task and Knowledge Graph:

a) The NLP task our fine-tuned LLM will consider is the classification of a textual description of any human disease into its scientific disease name.

b) The HDO (Human Disease Ontology) knowledge graph has a single-hierarchy nature and contains information for any disease, such as its definition (the description), the synonyms (other frequently used names), and some unique scientific codes given to them, and the data is available in formats such as .owl, .xml, .json etc.

3. Preprocessing the Knowledge Graph Data

a) *Data Formation*: The information such as disease definition, synonyms and scientific codes are combined together to form an attribute "**text_sequence**" and take the disease names to form the attribute "**label**", and a dataframe is created.

b) *Maintaining Data Consistency*: The attribute "**label**" consisted of all unique rows, and for better training the number of classes had to be reduced to a good number. We selectively picked up the most commonly occurring words in the diseases and produced a class_list according to which many of the diseases were grouped under the same name (like various types of allergy called just allergy). Also, checked for irregular distribution of classes.

c) *Data Cleaning*: The attribute “text_sequence” is cleaned with the removal of stopwords, punctuations along with the lowercasing of the text.

The final prepared data is represented as follows:

	text_sequence	label
0	gallbladder leiomyosarcoma gallbladder sarcoma...	arcoma
1	autosomal recessive nonsyndromic deafness 1b a...	deafness
2	obsolete mucinous bronchioloalveolar lung carc...	obsolete
3	spinocerebellar ataxia type 8 autosomal domina...	ataxia
4	gallbladder small cell carcinoma definition av...	carcinoma
...	...	...
9872	otopalatodigital syndrome type 1 otopalatodigi...	syndrome
9873	fallopian tube germ cell cancer fallopian tube...	cancer
9874	obsolete recurrent pediatric cerebellar astroc...	obsolete
9875	congenital muscular dystrophy-dystroglycanopat...	dystrophy
9876	obsolete calculus bile duct acute cholecystiti...	obsolete

9877 rows × 2 columns

4. Model Selection and Training Preparations

a) Initializing two models BioBERT and *BERT (base-uncased)* tokenizer and model, pretrained on a large corpus of English Data.

b) The “**text_sequences**” and “**label**” attributes are tokenized and appropriate padding with attention masks are applied to ensure consistency in encodings.

c) The dataset is split into a 85:15 ratio of Train-Test, and appropriate data loaders are initialized for the fine-tuning process.

d) The parameters selected for the training are:

- Num_epochs = 10
- Learning Rate = 2e - 5
- Batch Size = 16
- Optimizer = AdamW

5. Fine-Tuning Results

The models are trained utilizing the GPU and the training losses seem to decrease for both the models.

The final model testing results are as follows:

BioBERT:

	Precision	Recall	f1-Score	Support
accuracy			0.98	1482
macro avg	0.95	0.95	0.95	1482
weighted avg	0.98	0.98	0.98	1482

BERT:

	Precision	Recall	f1-Score	Support
accuracy			0.97	1482
macro avg	0.94	0.94	0.94	1482
weighted avg	0.97	0.97	0.97	1482

6. Conclusion:

While the accuracy for both models are comparable and possess very similar results, BioBERT still outperforms BERT, the reason being that it was designed for the biomedical domain purposes, and pre trained on tasks specific to the biological terminology, rather than more general text based, which is the base for the BERT model. In both the cases, the fine-tuning of LLMs on knowledge graphs proved to be beneficial.

7. Resources:

a) https://github.com/dylanhogg/llmgraph?tab=readme-ov-file
b) https://github.com/RManLuo/Awesome-LLM-KG
c) https://github.com/JohannesJolkkonen/funktio-ai-samples/tree/main/knowledge-graph-demo
d) https://github.com/DiseaseOntology/HumanDiseaseOntology/tree/main/src/ontology