

## 報告書

### 1 今週の進捗

- JSAI タイトル, 概要の提出
- ナレッジグラフ推論チャレンジ
- 学習時のマスク位置固定

### 2 ナレッジグラフ推論チャレンジ

表 1, 2 に本実験で用いるデータセットである Devils Foot におけるデータ数, パラメータを示す.

表 1: データセット

Dataset	Entity	Relation	Train	Test
Devils Foot	2,373	32	4,684	562

表 2: パラメータ

parameters	値
learning rate	5e-5
batch size	32
max seq length	128
epoch	5

表 3: MLM を用いた実験結果

モデル	WN18RR				
	MR	MRR	Hits@1	Hits@3	Hits@10
KG-BERT (文献値)	97	-	-	-	52.4
KG-BERT (再現実験)	117.77	0.25	12.41	29.44	51.85
ホームズデータセット	59.35	0.3268	16.19	43.95	61.56

データセット WN18RR より良い結果が得られた. これはデータセット数やデータの形式が異なることが影響していると考えられる. 犯人推定として, 犯行シーンを架空のシーン「000」と設定し, それを head とした triple を作成して triple 「000-subject-?」の tail を推定することを検討する.

表 4, 5 にホームズデータセット (Devils Foot) の triple の例, 予測した tail におけるランクと予測結果を示す.

表 4: triple の例

Head	Relation	Tail
001	source	"In the early spring of 1897, Holmes and Watson were in a cottage in Pordeux Bay"
001	hasPredicate	stay
001	subject	Holmes
001	subject	Watson
001	where	cottage of Pordeux Bay
001	when	Early spring of 1897
001	when	1897-03-01
059	source	"Roundhay said the following"
059	hasPredicate	say
059	what	060
059	what	061
090	source	"Owen and George were fearful"
090	infoSource	Mortimer
090	hasPredicate	become
090	what	Crazy
wakeUp	type	Action
wakeUp	label	"wakeUp"

### 3 学習時のマスク位置固定

コードの理解が足りず, まだ学習時のマスク位置を固定するコードは書けていない. 以下に現在使用している MLM のコードの一部を示す. `DataCollatorForLanguageModeling` でマスクかける位置 (15%) を設定しているため, `DataCollatorForLanguageModeling` を継承したクラスに引数として `train` データを渡してマスク位置の固定をすることを検討している.

Listing 1: code

```

1 dataset = LineByLineTextDataset(
2     tokenizer = tokenizer,
3     file_path = args.data_dir + '/test_triples.txt',
4     block_size = args.max_seq_length, # の tokenizer_max_length

```

表 5: 予測した tail の例

rank	Head	Relation	Tail	正解 tail
1	002	subject	Roundhay	Roundhay
1	071	what	078	078
1	106	infoSource	Mortimer	Mortimer
2	news	label	news "news"	"news"
32	notAgree	type	Agree notAgree notAction "Poison label"	Action

```

5         )
6
7         data_collator = DataCollatorForLanguageModeling(
8             tokenizer=tokenizer,
9             mlm=True,
10            mlm_probability= args.mlm_probability
11        )
12
13        training_args = TrainingArguments(
14            output_dir= args.output_dir, # output_dir = 'mlm_result/output_1'
15            overwrite_output_dir=True,
16            num_train_epochs=args.num_train_epochs,
17            per_device_train_batch_size=args.train_batch_size,
18            save_steps=10000,
19            save_total_limit=2,
20            prediction_loss_only=True,
21            learning_rate=args.learning_rate,
22        )
23
24        trainer = Trainer(
25            model=model,
26            args=training_args,
27            data_collator=data_collator,
28            train_dataset=dataset
29        )
30
31        trainer.train()
32        trainer.save_model(model_save_path)

```

---

## 4 今後したいこと

- MLM を用いた実験の改良

## 参考文献

- [1] Liang Yao, Chengsheng Mao, and Yuan Luo. KG-BERT: BERT for knowledge graph completion. *CoRR*, Vol. abs/1909.03193, , 2019.

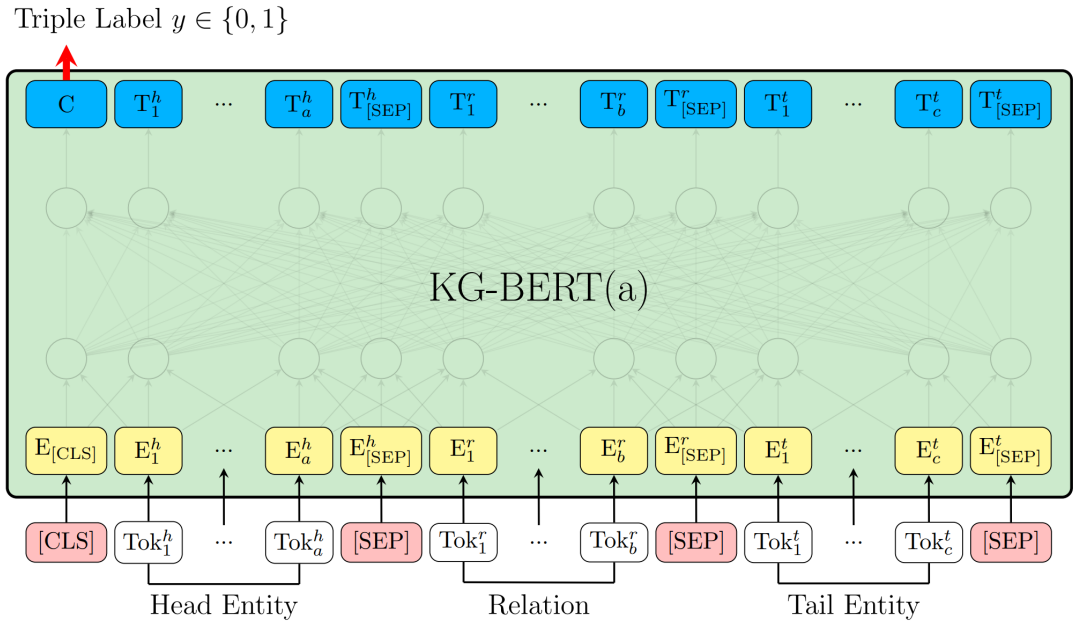


图 1: KG-BERT model [1]