KG-BERT: BERT for Knowledge Graph Completion

Liang Yao, Chengsheng Mao, Yuan Luo
Northwestern University, 2019

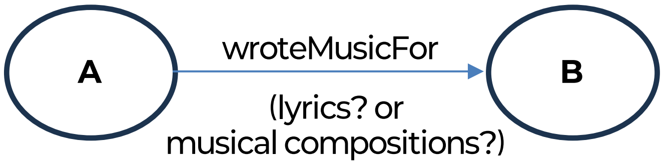
2025-02-11
HoonUi Lee

Contents

- **Previous Work**
- **KG-BERT**
  - Predicting plausibility
  - Predicting relation
- **Experiment**
  - Dataset
  - Triple classification
  - Link prediction
  - Relation prediction
- **Conclusion**

Previous Work

- ❖ **Attempt to incorporate text data**
  - ❑ But learn unique text embedding for the same entity/relation in different triples
    - ignore contextual information



Previous Work

- ❖ **Existing methods employ 3 concepts**
  - ❑ Entity descriptions
    - entity's simple text data
  - ❑ Relation mentions
    - Is relation mentioned in the description of the entity
  - ❑ Word co-occurrence with entities
    - Words that frequently appear together in the description of the entity

➔ **difficult to make accurate inferences due to the inability to learn context**

Entity descriptions

Lionel Andrés Messi is an Argentine professional footballer who plays for Inter Miami and the Argentina national team

Relation mentions

Messi chose to stay Barcelona in 2013

Messi stayed in PSG before Miami

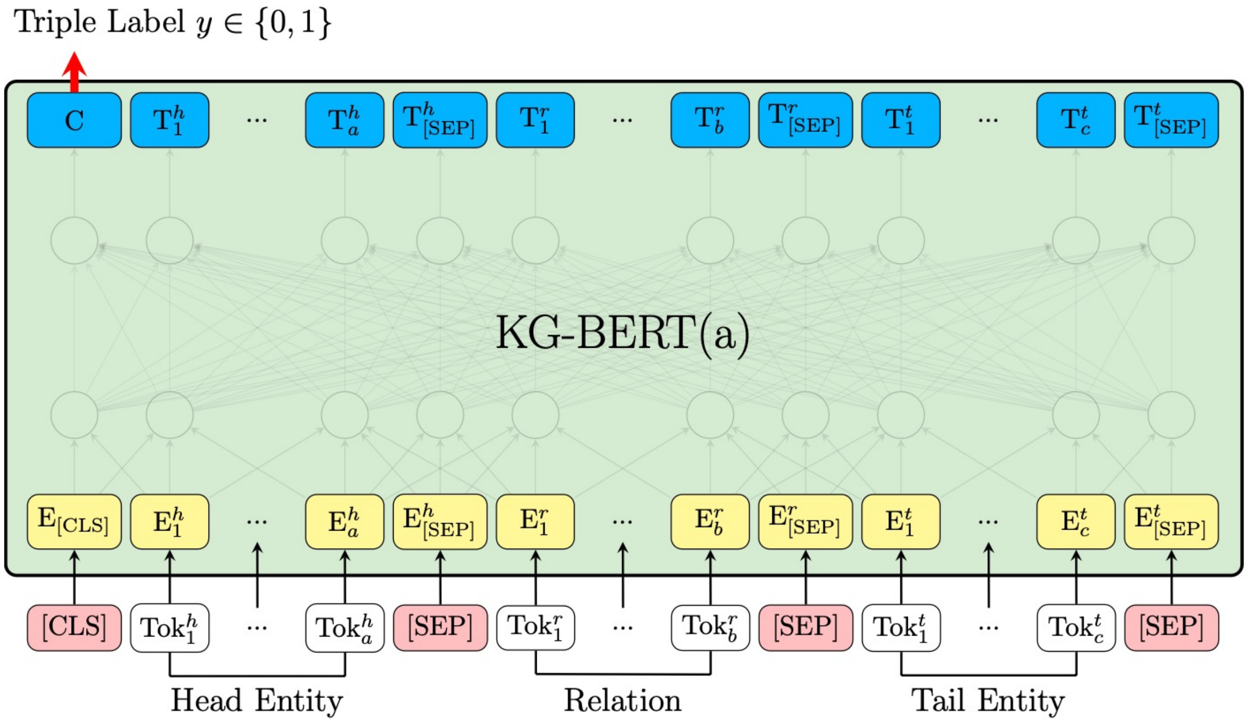
Word co-occurrence

Messi and Ronaldo played in the same match

Messi and Neymar played for the same club

KG-BERT

- Predicting plausibility



KG-BERT

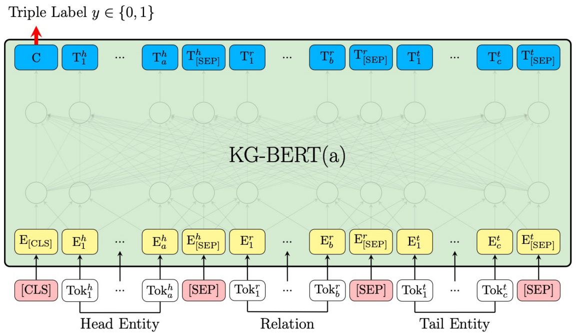
❖ BERT-based sentence representation learning

❑ Possible to dynamically learn contextual information
- the meaning of relations
- connectivity between entities in context

❑ Perform tasks on triples using a pre-trained BERT
- masked language modeling
- next sentence prediction

KG-BERT

- Predicting plausibility



❖ Convert the triple into a sentence

- ❑ Using entity and relation's text description
  - ‣ [CLS] + {head text} + [SEP] + {relation text} + [SEP] + {tail text} + [SEP]
- ❑ 3 embedding layer
  - ‣ token embedding
  - ‣ segment embedding
  - ‣ position embedding
- ❑ The final hidden state C corresponding to [CLS]
  - ‣ [CLS] is used as the aggregate sequence representation for computing triple scores

KG-BERT

- Predicting plausibility

❖ Scoring function

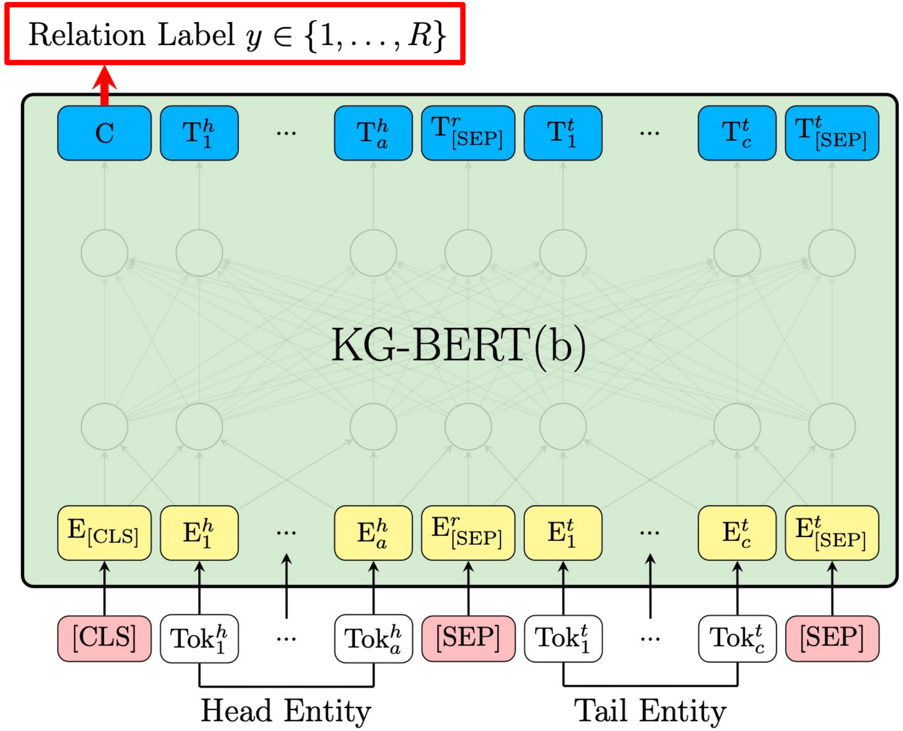
- ❑ $\mathbf{s}_\tau = f(h, r, t) = \mathrm{sigmoid}(CW^T)$
  - classification layer weights W
  - probability of being a valid
  - $C \in \mathbb{R}^H$, $W \in \mathbb{R}^{2\times H}$
  - $\mathbf{s}_\tau \in \mathbb{R}^2$ is a 2-dimensional real vector with $s_{\tau 0}, s_{\tau 1} \in [0, 1]$ ($s_{\tau 0} + s_{\tau 1} = 1$)

❖ Cross-entropy loss

- ❑ $\mathcal{L} = -\sum_{\tau \in \mathbb{D}^+ \cup \mathbb{D}^-} (y_\tau \log(s_{\tau 0}) + (1 - y_\tau) \log(s_{\tau 1}))$
  - Maximize $s_{\tau 0}$ when the triple is positive
  - Maximize $s_{\tau 1}$ when the triple is negative

KG-BERT

- Predicting relation



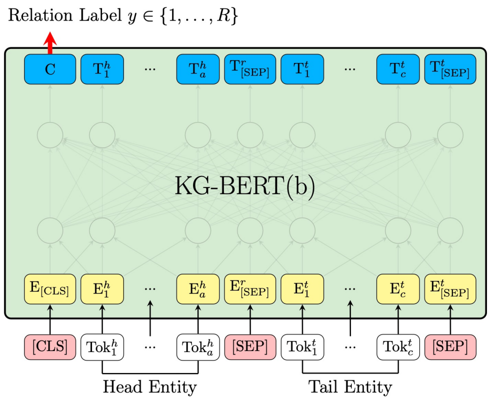
KG-BERT

- Predicting relation

❖ Multi-class Classification of relations

❑ Only using sentences of head & tail entities

- [CLS] + {head text} + [SEP] + {tail text} + [SEP]



❑ Better performance in predicting relations

- than using KG-BERT with **relation corruption**
- generating negative triples by replacing relation r with a random relation r'

❑ The final hidden state C corresponding to [CLS]

- [CLS] is used as the representation of the two entities

KG-BERT

- Predicting relation

Scoring function

- $\mathbf{s}'_\tau = f(h, r, t) = \mathrm{softmax}(CW'^T)$
  - classification layer weights $W' \in \mathbb{R}^{R \times H}$
  - R is number of relations in KG
  - $\mathbf{s}'_\tau \in \mathbb{R}^R$ is a R-dimensional real vector with $s'_{\tau i} \in [0, 1]$ ($\sum_i^R s'_{\tau i} = 1$)

Cross-entropy loss

- $\mathcal{L}' = -\sum_{\tau \in \mathbb{D}^+} \sum_{i=1}^{R} y'_{\tau i} \log(s'_{\tau i})$
  - $y'_{\tau i}$ = 1 if r = 1 else $y'_{\tau i}$ = 0
  - Maximize the predicted probability $s_{\tau i}$ for the correct relation r

Experiment

❖ Dataset

Dataset	# Ent	# Rel	# Train	# Dev	# Test
WN11	38,696	11	112,581	2,609	10,544
FB13	75,043	13	316,232	5,908	23,733
WN18RR	40,943	11	86,835	3,034	3,134
FB15K	14,951	1,345	483,142	50,000	59,071
FB15k-237	14,541	237	272,115	17,535	20,466
UMLS	135	46	5,216	652	661

Table 1: Summary statistics of datasets.

- **WN11, FB13** ➔ Triple Classification
- **WN18RR, FB15K, FB15k-237, UMLS** ➔ Entity Prediction & Relation Prediction

Experiment

- Triple Classification

❖ Result

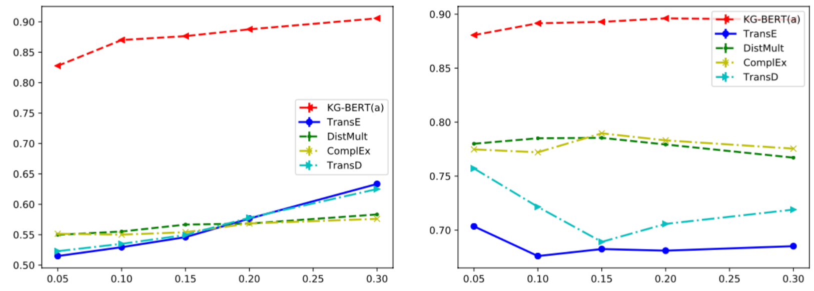
Method	WN11	FB13	Avg.
NTN (Socher et al. 2013)	86.2	90.0	88.1
TransE (Wang et al. 2014b)	75.9	81.5	78.7
TransH (Wang et al. 2014b)	78.8	83.3	81.1
TransR (Lin et al. 2015b)	85.9	82.5	84.2
TransD (Ji et al. 2015)	86.4	89.1	87.8
TEKE (Wang and Li 2016)	86.1	84.2	85.2
TransG (Xiao, Huang, and Zhu 2016)	87.4	87.3	87.4
TranSparse-S (Ji et al. 2016)	86.4	88.2	87.3
DistMult (Zhang et al. 2018)	87.1	86.2	86.7
DistMult-HRS (Zhang et al. 2018)	88.9	89.0	89.0
AATE (An et al. 2018)	88.0	87.2	87.6
ConvKB (Nguyen et al. 2018a)	87.6	88.8	88.2
DOLORES (Wang, Kulkarni, and Wang 2018)	87.5	89.3	88.4
KG-BERT(a)	**93.5**	**90.4**	**91.9**

- ❑ Judge whether a given triple (h, r, t) is correct or not
- ❑ Average accuracy with 10 times
- ❑ WN11 has strong linguistic characteristics, making it well-suited for BERT

Experiment

- Triple Classification

❖ Result



(a) WN11 (b) FB13

Figure 3: Test accuracy of triple classification by varying training data proportions.

- ❑ Limitations on the use of training data
- ❑ KG-BERT(a) can overcome the sparseness of knowledge graphs
  - by using linguistic patterns in large external text data

Experiment

- Link prediction

❖ Result

Method	WN18RR		FB15k-237		UMLS	
	MR	Hits@10	MR	Hits@10	MR	Hits@10
TransE (our results)	2365	50.5	223	47.4	1.84	98.9
TransH (our results)	2524	50.3	255	48.6	1.80	**99.5**
TransR (our results)	3166	50.7	237	51.1	1.81	99.4
TransD (our results)	2768	50.7	246	48.4	1.71	99.3
DistMult (our results)	3704	47.7	411	41.9	5.52	84.6
ComplEx (our results)	3921	48.3	508	43.4	2.59	96.7
ConvE (Dettmers et al. 2018)	5277	48	246	49.1	–	–
ConvKB (Nguyen et al. 2018a)	2554	52.5	257	51.7	–	–
R-GCN (Schlichtkrull et al. 2018)	–	–	–	41.7	–	–
KBGAN (Cai and Wang 2018)	–	48.1	–	45.8	–	–
RotatE (Sun et al. 2019)	3340	**57.1**	177	**53.3**	–	–
KG-BERT(a)	**97**	52.4	**153**	42.0	**1.47**	99.0

❑ Predict missing head or tail

❑ Lower Hits@10 score than existing methods

- inability to utilize structural information such as neighboring entities in the KG

Experiment

- Relation prediction

❖ Result

Method	Mean Rank	Hits@1
TransE (Lin et al. 2015a)	2.5	84.3
TransR (Xie, Liu, and Sun 2016)	2.1	91.6
DKRL (CNN) (Xie et al. 2016)	2.5	89.0
DKRL (CNN) + TransE (Xie et al. 2016)	2.0	90.8
DKRL (CBOW) (Xie et al. 2016)	2.5	82.7
TKRL (RHE) (Xie, Liu, and Sun 2016)	1.7	92.8
TKRL (RHE) (Xie, Liu, and Sun 2016)	1.8	92.5
PTransE (ADD, len-2 path) (Lin et al. 2015a)	**1.2**	93.6
PTransE (RNN, len-2 path) (Lin et al. 2015a)	1.4	93.2
PTransE (ADD, len-3 path) (Lin et al. 2015a)	1.4	94.0
SSP (Xiao et al. 2017)	**1.2**	–
ProjE (pointwise) (Shi and Weninger 2017)	1.3	95.6
ProjE (listwise) (Shi and Weninger 2017)	**1.2**	95.7
ProjE (wlistwise) (Shi and Weninger 2017)	**1.2**	95.6
KG-BERT (b)	**1.2**	**96.0**

❑ Predict missing relation

❑ KG-BERT(b) leverage the advantages of pre-trained BERT

- by using BERT trained with Sentence Pair Classification

Experiment

- Attention Visualization

❖ Result (KG-BERT(a))

❑ About triple (twenty dollar bill, hypernym, note)

❑ Certain attention heads focus on structural tokens

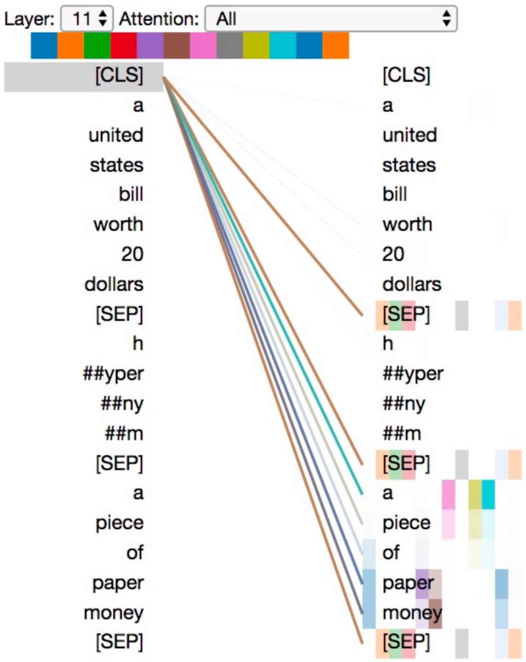
- [SEP]

❑ Other attention heads focus on common words

- 'a' and 'piece’

❑ Important concept words are highlighted from specific attention heads

- 'paper', 'money'



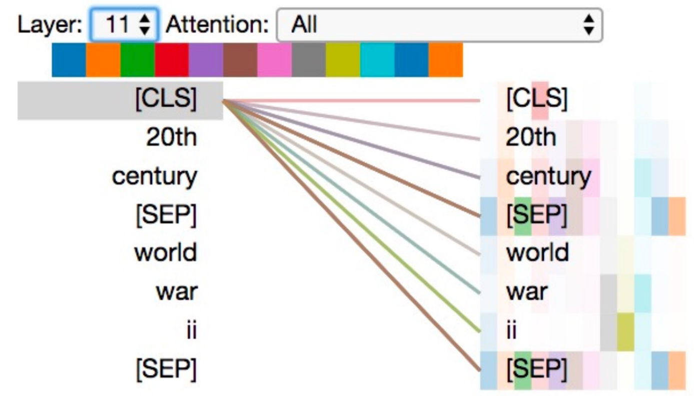
Experiment

- Attention Visualization

❖ Result (KG-BERT(b))

❑ Two entities 20th century and World War II as input

❑ Relation label is /time/event/includes event



❑ Six attention heads focus on '**century**',

❑ While three other attention heads focus on '**war**' and '**ii**'

❑ Multi-head attention can attend to different aspects of two entities in a triple

Conclusion

- Previous work
  - Learn unique text embedding for the same entity/relation in different triples
  - Syntactic and semantic information in large-scale text data is not fully utilized

- KG-BERT
  - Use a pre-trained BERT
  - Convert the descriptions (or the entities and relations themselves) into sentence and use it as input

- Experiment
  - Well suited for datasets with linguistic characteristics
  - Overcome the sparseness of knowledge graphs