E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT (Appendix)

Unsupervised QA (LAMA)

Data

We downloaded the LAMA dataset from https://dl.fbaipublicfiles.com/LAMA/data.zip. We use the LAMA-T-REx and LAMA-Google-RE relations, which are aimed at factual knowledge. Table 10 shows results on indiviual relations, as well as the number of questions per relation before and after applying the LAMA-UHN heuristics.

Preprocessing

As mentioned in Section 4.1, we do not use LAMA's oracle entity IDs. Instead, we map surface forms to entity IDs via the Wikidata query API (https://query.wikidata.org). For example, to look up *Jean Marais*:

```
SELECT ?id ?str WHERE {
  ?id rdfs:label ?str .
  VALUES ?str { 'Jean Marais'@en } .
  FILTER((LANG(?str)) = 'en') .
}
```

If more than one Wikidata ID is returned, we select the lowest one. We then map Wikidata IDs to the corresponding Wikipedia URLs:

```
SELECT ?id ?wikiurl WHERE {
  VALUES ?id { wd:Q168359 } .
  ?wikiurl schema:about ?id .
  ?wikiurl schema:inLanguage 'en' .
  FILTER REGEX(str(?wikiurl),
        '.*en.wikipedia.org.*') .
}
```

Relation classification

Data

The RC dataset, which is a subset of the FewRel corpus, was compiled by Zhang et al. (2019). We downloaded it from https://cloud.tsinghua.edu.cn/f/32668247e4fd4f9789f2/. Table 8 shows dataset statistics.

Preprocessing

The dataset contains sentences with annotated subject and object entity mentions, their oracle entity IDs and their relation (which must be predicted). We use the BERT wordpiece tokenizer to tokenize the sentence and insert special wordpieces to mark the entity mentions: *#* for subjects and *$* for objects. Then, we insert the entity IDs. For example, an input to E-BERT-concat would look like this:

> *[CLS] Taylor was later part of the ensemble cast in MGM 's classic $* ***World_War_II*** */ World War II $ drama “ #* ***Battleground_(film)*** */ Battle ##ground # ” (1949) . [SEP]*

We use the oracle entity IDs of the dataset, which are also used by ERNIE (Zhang et al., 2019).

Hyperparameters

We tune peak learning rate and number of epochs on the dev set (selection criterion: macro F1). We do a full search over the same hyperparameter space as Zhang et al. (2019):

Learning rate: $[2 \cdot 10^{-5}, 3 \cdot 10^{-5}, \mathbf{5 \cdot 10^{-5}}]$

Number of epochs: $[3, 4, 5, 6, 7, 8, 9, \mathbf{10}]$

The best configuration for E-BERT-concat is marked in bold. Figure 6 shows expected maximum performance as a function of the number of evaluated configurations (Dodge et al., 2019).

Entity linking (AIDA)

Data

We downloaded the AIDA dataset from:

- https://allennlp.s3-us-west-2.amazonaws.com/knowbert/wiki_entity_linking/aida_train.txt
- https://allennlp.s3-us-west-2.amazonaws.com/knowbert/wiki_entity_linking/aida_dev.txt
- https://allennlp.s3-us-west-2.amazonaws.com/knowbert/wiki_entity_linking/aida_test.txt

Preprocessing

Each AIDA file contains documents with annotated entity spans (which must be predicted). The documents are already whitespace tokenized, and we further tokenize words into wordpieces with the standard BERT tokenizer. If a document is too long (length > 512), we split it into smaller chunks by (a) finding the sentence boundary that is closest to the document midpoint, (b) splitting the document, and (c) repeating this process recursively until all chunks are short enough. Table 9 shows dataset statistics.

Hyperparameters

We tune batch size and peak learning rate on the AIDA dev set (selection criterion: strong match micro F1). We do a full search over the following hyperparameter space:

Batch size: $[16, 32, 64, \mathbf{128}]$

Learning rate: $[\mathbf{2 \cdot 10^{-5}}, 3 \cdot 10^{-5}, 5 \cdot 10^{-5}]$

The best configuration for E-BERT-MLM is marked in bold. Figure 7 shows expected maximum performance as a function of the number of evaluated configurations (Dodge et al., 2019).

# relations	80		
# unique entities	54648		
	train	dev	test
# samples	8000	16000	16000
# samples per relation	100	200	200

Table 8: Relation classification dataset statistics.

# unique gold entities	5574		
# unique candidate entities	463663		
	train	dev	test
# documents	946	216	231
# documents (after chunking)	1111	276	271
# potential spans (candidate generator)	153103	38012	34936
# gold entities	18454	4778	4478

Table 9: Entity linking (AIDA) dataset statistics.

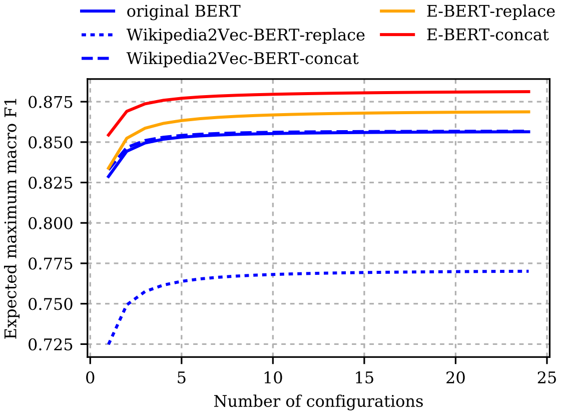

Figure 6: Relation classification: Expected maximum macro F1 (dev set) as a function of the number of hyperparameter configurations.

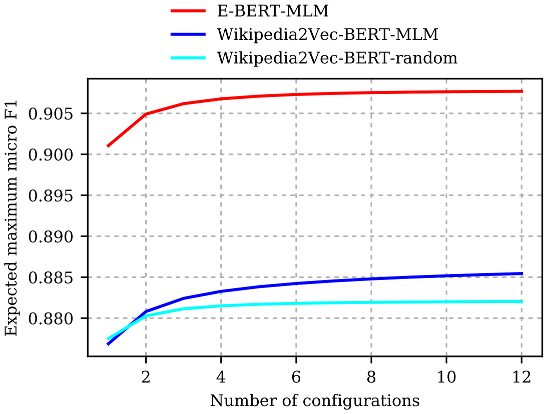

Figure 7: Entity linking: Expected maximum micro F1 (dev set) as a function of the number of hyperparameter configurations.

	Model size:	BASE					LARGE			
Relation (dataset)	Model	original BERT	E-BERT replace	E-BERT-concat	ERNIE	KnowBert	original BERT	E-BERT-replace	E-BERT-concat	number of questions
T-REx:P17	(0, original LAMA)	31.3	53.7	52.4	55.3	23.7	36.5	43.3	42.8	930
T-REx:P17	(1)	31.0	55.0	53.3	55.5	23.2	36.2	44.5	43.3	885
T-REx:P17	(2, LAMA-UHN)	31.0	55.0	53.3	55.5	23.2	36.2	44.5	43.3	885
T-REx:P19	(0, original LAMA)	21.1	26.4	28.1	28.7	23.3	22.2	24.6	25.3	944
T-REx:P19	(1)	20.6	26.5	27.5	28.2	22.9	21.8	24.5	24.8	933
T-REx:P19	(2, LAMA-UHN)	9.8	20.3	18.7	19.4	12.2	11.7	18.1	15.5	728
T-REx:P20	(0, original LAMA)	27.9	29.7	35.8	16.6	31.1	31.7	37.1	33.5	953
T-REx:P20	(1)	28.2	29.9	36.0	16.5	31.0	32.0	37.2	33.8	944
T-REx:P20	(2, LAMA-UHN)	15.5	21.5	23.3	8.4	20.0	18.9	27.3	22.6	656
T-REx:P27	(0, original LAMA)	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.1	966
T-REx:P27	(1)	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.1	945
T-REx:P27	(2, LAMA-UHN)	0.0	0.0	0.2	0.0	0.1	0.0	0.0	0.2	423
T-REx:P30	(0, original LAMA)	25.4	69.9	69.8	66.8	24.0	28.0	75.0	60.4	975
T-REx:P30	(1)	25.1	70.3	69.9	66.6	23.9	27.5	75.0	60.3	963
T-REx:P30	(2, LAMA-UHN)	25.1	70.3	69.9	66.6	23.9	27.5	75.0	60.3	963
T-REx:P31	(0, original LAMA)	36.7	25.5	46.9	43.7	18.7	30.2	12.3	16.1	922
T-REx:P31	(1)	21.1	28.4	35.8	30.3	12.4	16.3	9.9	9.8	564
T-REx:P31	(2, LAMA-UHN)	21.1	28.4	35.8	30.3	12.4	16.3	9.9	9.8	564
T-REx:P36	(0, original LAMA)	62.2	42.1	61.6	57.3	62.2	67.0	44.7	66.0	703
T-REx:P36	(1)	51.5	41.9	53.9	45.9	51.7	57.5	43.8	58.8	534
T-REx:P36	(2, LAMA-UHN)	51.5	41.9	53.9	45.9	51.7	57.5	43.8	58.8	534
T-REx:P37	(0, original LAMA)	54.6	51.2	56.5	60.2	53.1	61.5	54.3	62.7	966
T-REx:P37	(1)	52.9	51.6	55.5	59.4	51.9	60.5	54.2	62.1	924
T-REx:P37	(2, LAMA-UHN)	52.9	51.6	55.5	59.4	51.9	60.5	54.2	62.1	924
T-REx:P39	(0, original LAMA)	8.0	22.9	22.5	17.0	17.2	4.7	8.1	8.6	892
T-REx:P39	(1)	7.5	23.0	22.3	17.1	16.5	4.6	8.1	8.5	878
T-REx:P39	(2, LAMA-UHN)	7.5	23.0	22.3	17.1	16.5	4.6	8.1	8.5	878
T-REx:P47	(0, original LAMA)	13.7	8.9	10.8	9.8	14.0	18.2	15.1	15.9	922
T-REx:P47	(1)	13.6	9.1	10.7	9.6	13.9	18.6	15.2	15.9	904
T-REx:P47	(2, LAMA-UHN)	13.6	9.1	10.7	9.6	13.9	18.6	15.2	15.9	904
T-REx:P101	(0, original LAMA)	9.9	37.8	40.8	16.7	12.2	11.5	37.8	36.1	696
T-REx:P101	(1)	9.5	38.2	40.9	16.1	11.4	10.8	38.0	35.8	685
T-REx:P101	(2, LAMA-UHN)	9.5	38.2	40.9	16.1	11.4	10.8	38.0	35.8	685
T-REx:P103	(0, original LAMA)	72.2	85.8	86.8	85.5	73.4	78.2	84.4	84.9	977
T-REx:P103	(1)	72.1	85.7	86.8	85.4	73.3	78.2	84.4	84.9	975
T-REx:P103	(2, LAMA-UHN)	45.8	81.9	74.7	83.6	72.2	58.6	81.2	71.1	415
T-REx:P106	(0, original LAMA)	0.6	6.5	5.4	8.4	1.6	0.6	4.3	2.1	958
T-REx:P106	(1)	0.6	6.5	5.4	8.4	1.6	0.6	4.3	2.1	958
T-REx:P106	(2, LAMA-UHN)	0.6	6.5	5.4	8.4	1.6	0.6	4.3	2.1	958
T-REx:P108	(0, original LAMA)	6.8	9.9	23.2	14.1	10.7	1.6	11.7	15.9	383
T-REx:P108	(1)	6.5	9.9	23.0	13.9	10.5	1.3	11.8	16.0	382
T-REx:P108	(2, LAMA-UHN)	6.5	9.9	23.0	13.9	10.5	1.3	11.8	16.0	382
T-REx:P127	(0, original LAMA)	34.8	24.0	34.9	36.2	31.4	34.8	25.3	35.8	687
T-REx:P127	(1)	14.2	19.7	23.5	17.1	15.5	14.6	21.1	24.6	451
T-REx:P127	(2, LAMA-UHN)	14.2	19.7	23.5	17.1	15.5	14.6	21.1	24.6	451

Table 10: Mean Hits@1 and number of questions per LAMA relation. 0: original LAMA dataset, 1: after applying heuristic 1 (string match filter), 2: after applying both heuristics (LAMA-UHN).

Model size:	BASE					LARGE			
Model Relation (dataset)	original BERT	E-BERT replace	E-BERT-concat	ERNIE	KnowBert	original BERT	E-BERT-replace	E-BERT-concat	number of questions
T-REx:P131 (0, original LAMA)	23.3	33.4	36.4	37.3	27.7	26.3	31.4	37.2	881
T-REx:P131 (1)	16.7	32.0	33.9	32.7	21.5	20.1	31.0	33.4	706
T-REx:P131 (2, LAMA-UHN)	16.7	32.0	33.9	32.7	21.5	20.1	31.0	33.4	706
T-REx:P136 (0, original LAMA)	0.8	5.2	9.1	0.6	0.6	1.3	6.9	13.1	931
T-REx:P136 (1)	0.2	5.1	8.7	0.2	0.1	0.2	6.9	12.2	913
T-REx:P136 (2, LAMA-UHN)	0.2	5.1	8.7	0.2	0.1	0.2	6.9	12.2	913
T-REx:P138 (0, original LAMA)	61.6	8.8	26.5	0.2	63.7	45.1	2.6	24.0	645
T-REx:P138 (1)	5.0	10.0	8.8	0.0	6.9	4.4	4.4	6.2	160
T-REx:P138 (2, LAMA-UHN)	5.0	10.0	8.8	0.0	6.9	4.4	4.4	6.2	160
T-REx:P140 (0, original LAMA)	0.6	0.6	1.1	0.0	0.8	0.6	1.1	0.6	473
T-REx:P140 (1)	0.4	0.6	0.9	0.0	0.6	0.4	0.9	0.4	467
T-REx:P140 (2, LAMA-UHN)	0.4	0.6	0.9	0.0	0.6	0.4	0.9	0.4	467
T-REx:P159 (0, original LAMA)	32.4	30.3	48.3	41.8	36.8	34.7	22.3	45.2	967
T-REx:P159 (1)	23.1	31.6	41.9	34.4	28.7	25.6	20.9	37.8	843
T-REx:P159 (2, LAMA-UHN)	23.1	31.6	41.9	34.4	28.7	25.6	20.9	37.8	843
T-REx:P176 (0, original LAMA)	85.6	41.6	74.6	81.8	90.0	87.5	36.6	81.3	982
T-REx:P176 (1)	31.4	42.9	51.8	26.2	51.3	40.8	44.5	57.1	191
T-REx:P176 (2, LAMA-UHN)	31.4	42.9	51.8	26.2	51.3	40.8	44.5	57.1	191
T-REx:P178 (0, original LAMA)	62.8	49.8	66.6	60.1	70.3	70.8	51.2	69.4	592
T-REx:P178 (1)	40.7	42.6	51.6	36.9	52.2	53.6	51.1	57.7	366
T-REx:P178 (2, LAMA-UHN)	40.7	42.6	51.6	36.9	52.2	53.6	51.1	57.7	366
T-REx:P190 (0, original LAMA)	2.4	2.9	2.5	2.6	2.8	2.3	2.3	2.8	995
T-REx:P190 (1)	1.5	2.4	1.6	1.6	2.0	1.7	1.9	2.3	981
T-REx:P190 (2, LAMA-UHN)	1.5	2.4	1.6	1.6	2.0	1.7	1.9	2.3	981
T-REx:P264 (0, original LAMA)	9.6	30.5	33.6	13.3	21.2	8.2	23.1	15.6	429
T-REx:P264 (1)	9.6	30.6	33.4	13.3	21.3	8.2	23.1	15.7	428
T-REx:P264 (2, LAMA-UHN)	9.6	30.6	33.4	13.3	21.3	8.2	23.1	15.7	428
T-REx:P276 (0, original LAMA)	41.5	23.8	47.7	48.4	43.3	43.8	23.1	51.8	959
T-REx:P276 (1)	19.8	26.1	31.7	27.0	20.6	23.4	25.0	36.0	625
T-REx:P276 (2, LAMA-UHN)	19.8	26.1	31.7	27.0	20.6	23.4	25.0	36.0	625
T-REx:P279 (0, original LAMA)	30.7	14.7	30.7	29.4	31.6	33.5	15.5	29.8	963
T-REx:P279 (1)	3.8	8.6	8.0	4.6	5.3	6.8	8.6	10.1	474
T-REx:P279 (2, LAMA-UHN)	3.8	8.6	8.0	4.6	5.3	6.8	8.6	10.1	474
T-REx:P361 (0, original LAMA)	23.6	19.6	23.0	25.8	26.6	27.4	22.3	25.4	932
T-REx:P361 (1)	12.6	17.9	17.7	13.7	15.3	18.5	20.2	22.0	633
T-REx:P361 (2, LAMA-UHN)	12.6	17.9	17.7	13.7	15.3	18.5	20.2	22.0	633
T-REx:P364 (0, original LAMA)	44.5	61.7	64.0	48.0	40.9	51.1	60.6	61.3	856
T-REx:P364 (1)	43.5	61.7	63.5	47.4	40.0	50.7	60.5	61.2	841
T-REx:P364 (2, LAMA-UHN)	43.5	61.7	63.5	47.4	40.0	50.7	60.5	61.2	841
T-REx:P407 (0, original LAMA)	59.2	68.0	68.8	53.8	60.1	62.1	57.9	56.3	877
T-REx:P407 (1)	57.6	69.5	67.9	53.1	58.6	61.0	59.0	55.2	834
T-REx:P407 (2, LAMA-UHN)	57.6	69.5	67.9	53.1	58.6	61.0	59.0	55.2	834
T-REx:P413 (0, original LAMA)	0.5	0.1	0.0	0.0	41.7	4.1	14.0	7.0	952
T-REx:P413 (1)	0.5	0.1	0.0	0.0	41.7	4.1	14.0	7.0	952
T-REx:P413 (2, LAMA-UHN)	0.5	0.1	0.0	0.0	41.7	4.1	14.0	7.0	952

Table 11: Mean Hits@1 and number of questions per LAMA relation (cont'd). 0: original LAMA dataset, 1: after applying heuristic 1 (string match filter), 2: after applying both heuristics (LAMA-UHN).

Model size:		BASE					LARGE			
Model Relation (dataset		original BERT	E-BERT replace	E-BERT-concat	ERNIE	Know-Bert	original BERT	E-BERT-replace	E-BERT-concat	number of questions
T-REx:P449	(0, original LAMA)	20.9	30.9	34.7	33.8	57.0	24.0	32.5	28.6	881
T-REx:P449	(1)	18.8	31.1	33.4	32.0	56.0	21.8	32.9	27.5	848
T-REx:P449	(2, LAMA-UHN)	18.8	31.1	33.4	32.0	56.0	21.8	32.9	27.5	848
T-REx:P463	(0, original LAMA)	67.1	61.8	68.9	43.1	35.6	61.3	52.0	66.7	225
T-REx:P463	(1)	67.1	61.8	68.9	43.1	35.6	61.3	52.0	66.7	225
T-REx:P463	(2, LAMA-UHN)	67.1	61.8	68.9	43.1	35.6	61.3	52.0	66.7	225
T-REx:P495	(0, original LAMA)	16.5	46.3	48.3	1.0	30.8	29.7	56.7	46.9	909
T-REx:P495	(1)	15.0	46.0	47.5	0.9	29.6	28.5	56.6	46.2	892
T-REx:P495	(2, LAMA-UHN)	15.0	46.0	47.5	0.9	29.6	28.5	56.6	46.2	892
T-REx:P527	(0, original LAMA)	11.1	7.4	11.9	5.4	12.9	10.5	8.9	12.9	976
T-REx:P527	(1)	5.7	7.6	8.7	0.5	3.0	4.2	8.7	6.3	804
T-REx:P527	(2, LAMA-UHN)	5.7	7.6	8.7	0.5	3.0	4.2	8.7	6.3	804
T-REx:P530	(0, original LAMA)	2.8	1.8	2.0	2.3	2.8	2.7	2.3	2.8	996
T-REx:P530	(1)	2.8	1.8	2.0	2.3	2.8	2.7	2.3	2.8	996
T-REx:P530	(2, LAMA-UHN)	2.8	1.8	2.0	2.3	2.8	2.7	2.3	2.8	996
T-REx:P740	(0, original LAMA)	7.6	10.5	14.7	0.0	10.4	6.0	13.1	10.4	936
T-REx:P740	(1)	5.9	10.3	13.5	0.0	9.0	5.2	12.7	9.5	910
T-REx:P740	(2, LAMA-UHN)	5.9	10.3	13.5	0.0	9.0	5.2	12.7	9.5	910
T-REx:P937	(0, original LAMA)	29.8	33.0	38.8	40.0	32.3	24.9	28.3	34.5	954
T-REx:P937	(1)	29.9	32.9	38.7	39.9	32.2	24.8	28.2	34.4	950
T-REx:P937	(2, LAMA-UHN)	29.9	32.9	38.7	39.9	32.2	24.8	28.2	34.4	950
T-REx:P1001	(0, original LAMA)	70.5	56.9	76.0	75.7	73.0	73.3	49.5	78.0	701
T-REx:P1001	(1)	38.1	67.7	66.7	65.6	43.4	40.7	60.3	66.7	189
T-REx:P1001	(2, LAMA-UHN)	38.1	67.7	66.7	65.6	43.4	40.7	60.3	66.7	189
T-REx:P1303	(0, original LAMA)	7.6	20.3	26.6	5.3	9.1	12.5	29.7	33.2	949
T-REx:P1303	(1)	7.6	20.3	26.6	5.3	9.1	12.5	29.7	33.2	949
T-REx:P1303	(2, LAMA-UHN)	7.6	20.3	26.6	5.3	9.1	12.5	29.7	33.2	949
T-REx:P1376	(0, original LAMA)	73.9	41.5	62.0	71.8	75.2	82.1	47.4	70.1	234
T-REx:P1376	(1)	74.8	42.2	62.8	73.4	75.2	83.5	48.6	72.0	218
T-REx:P1376	(2, LAMA-UHN)	74.8	42.2	62.8	73.4	75.2	83.5	48.6	72.0	218
T-REx:P1412	(0, original LAMA)	65.0	54.0	67.8	73.1	69.2	63.6	49.3	61.2	969
T-REx:P1412	(1)	65.0	54.0	67.8	73.1	69.2	63.6	49.3	61.2	969
T-REx:P1412	(2, LAMA-UHN)	37.7	42.9	47.4	69.2	65.7	51.5	43.5	54.8	361
Google-RE:date_of_birth	(0)	1.6	1.5	1.9	1.9	2.4	1.5	1.5	1.3	1825
Google-RE:date_of_birth	(1)	1.6	1.5	1.9	1.9	2.4	1.5	1.5	1.3	1825
Google-RE:date_of_birth	(2)	1.6	1.5	1.9	1.9	2.4	1.5	1.5	1.3	1825
Google-RE:place_of_birth	(0)	14.9	16.2	16.9	17.7	17.4	16.1	14.8	16.6	2937
Google-RE:place_of_birth	(1)	14.9	16.2	16.8	17.7	17.4	16.0	14.8	16.6	2934
Google-RE:place_of_birth	(2)	5.9	9.4	8.2	10.3	9.4	7.2	8.5	7.9	2451
Google-RE:place_of_death	(0)	13.1	12.8	14.9	6.4	13.4	14.0	17.0	14.9	766
Google-RE:place_of_death	(1)	13.1	12.8	14.9	6.4	13.4	14.0	17.0	14.9	766
Google-RE:place_of_death	(2)	6.6	7.5	7.8	2.0	7.5	7.6	11.8	8.9	655

Table 12: Mean Hits@1 and number of questions per LAMA relation (cont'd). 0: original LAMA dataset, 1: after applying heuristic 1 (string match filter), 2: after applying both heuristics (LAMA-UHN).