Language Models as Knowledge Bases?

Fabio Petroni, Tim Rocktäschel, Patrick Lewis,
Anton Bakhtin, Yuxiang Wu, Alexander H. Miller,
Sebastian Riedel @EMNLP2019

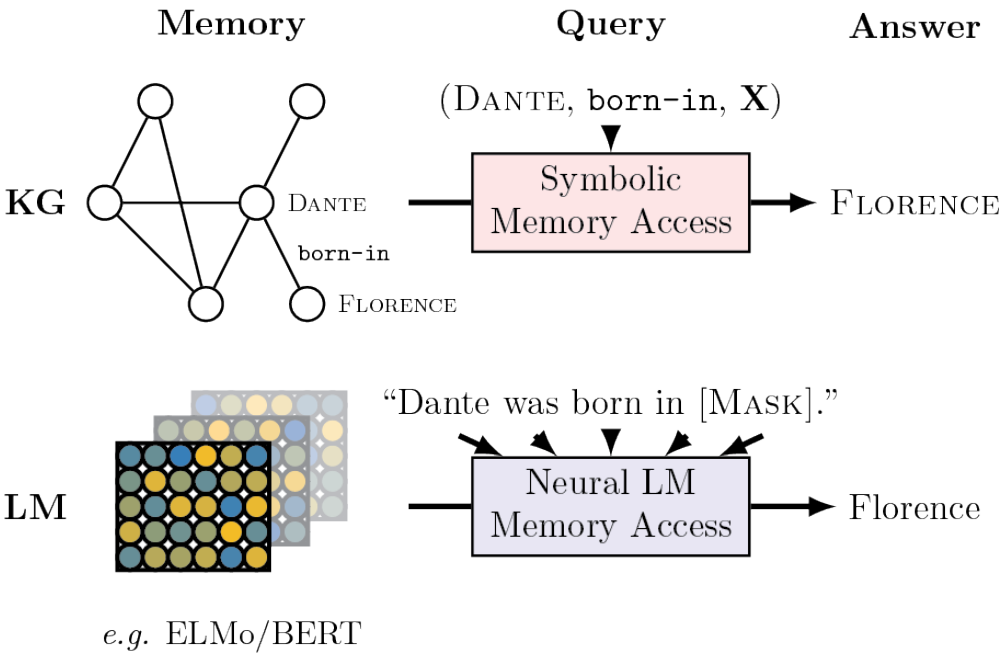
理研AIP
栗田修平
shuhei.kurita [at] riken.jp

Language Models as Knowledge Bases?

Abstract

- 事前学習されたLMには学習コーパスからの関係知識が入っているはずであり、穴埋め形式のクエリに解答できるはずである
- LMにはKBに対しいくつかの利点がある：schemeが不要、オープンクラスのクエリが可能、データの拡張が用意、学習にhuman supervisionが不要
- fine-tuningなしで公開済みLMに含まれている関係知識を評価
  （本研究ではBERTをfine-tuneする話は出てきません！→参考論文）
  - (i) 伝統的な手法とcompetitiveな関係知識をBERTは含んでいる
  - (ii) BERTはオープンドメインQAで教師ありベースラインより明確に優れている
  - (iii) LMの事前学習では、ある種の事実知識が他よりよく学習されている
- LAMA probeを公開

Intro



- LMには事前学習にて言語的な知識が入っている
  "Dante was born in Florence in 1265."
  "Dante was born in [MASK] in 1265."
- KBでは、(Dante, born-in, ?)のような形式で関係知識をクエリし、増やそうとする (KB population)
- ELMoやBERTには関係知識が入っているのでは？
- 関係知識を評価するためにLAMA probe を作成
- 結果
  - BERT-largeがoff the shelve rel. extractorとcomparable
  - 事実的な知識はLMから取れる
    N対M関係はダメ
  - BERT-largeが他手法より一貫して良い
  - Open domain QAで BERT-largeが57.1%
    supervisedなシステムからのKBは
    63.5% (precision@10)

LAMA probe

- **LAMA (LAnguage Model Analysis)**
  関係タプル(またはQAペア)をcloze形式に変換
  - **Google-RE**
    - 5つの関係のうち、“Place of birth”, “date-of-birth” and “place-of-death” のみを抽出
      (“institution”と”education-degree”を除外)
    - 3つの関係に人手で作成したテンプレートを適用しcloze作成
  - **T-Rex**
    - Wikidata subset
    - 41 の関係を考慮し、それぞれ関係に1000のfactをサンプル
    - 人手で作成したテンプレートを適用しcloze作成
  - **ConceptNet**
    - English part of ConceptNet, single token objects covering 16 relations
    - Open Mind Common Sense (OMCS) sentencesからsub, objを含む文を抽出してobjをMASK
  - **SQuAD**
    - Squadのdev setから、コンテクストに依存せず、single tokenがanswerのQAを抽出
    - 305例。すべて人手でcloze形式に:
      “Who developed the theory of relativity?”→ “The theory of relativity was developed by ____”

Baselines

- Frequency (Freq)
- Relation Extracter (RE)

  Sorolon and Gurevych (2017). LSTM-based encoder + attention
  文から(sub`, rel, obj`)を抽出した後に
  obj`をentity linkingで処理

  RE_o : entity linkingにoracle使用

  RE_n : entityの完全一致

- Open domain QA model (DrQA)

  Chen+ (2017)
  IF/IDF retrieval + reading comprehension
  only for SQuAD probe

LAMA probe

Considerations

- **Manually Defined Templates**
  1つの関係ラベルに付き人手で1つのクエリテンプレートを作成
  “we are measuring *lower* bound for what language models know”
- **Single token (!)**
  クエリの答えに現れるのはsingle tokenなエンティティのみ！
  (BERTのようなモデルには、エンティティのtoken数がヒントになる)
- **Object Slots**
  クエリの答えに現れるのは(sub, rel ,obj)のobjのみ！
- **Intersection of Vocabularies**
  LMはvocab.からtokenを選択
  → vocab.が多いほどgold tokenを選びづらい
  → common vocab.から21Kのcasedなtokenを考慮

例

	Relation	Query
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.
	P20	Adolphe Adam died in ____.
	P279	English bulldog is a subclass of ____.
	P37	The official language of Mauritius is ____.
	P413	Patrick Oboya plays in ____ position.
	P138	Hamburg Airport is named after ____.
	P364	The original language of Mon oncle Benjamin is ____.
	P54	Dani Alves plays with ____ .
	P106	Paul Toungui is a ____ by profession .
	P527	Sodium sulfide consists of ____.
	P102	Gordon Scholes is a member of the ____ political party.
	P530	Kenya maintains diplomatic relations with ____.
	P176	iPod Touch is produced by ____.
	P30	Bailey Peninsula is located in ____.
	P178	JDK is developed by ____.
	P1412	Carl III used to communicate in ____.
	P17	Sunshine Coast, British Columbia is located in ____.
	P39	Pope Clement VII has the position of ____ .
	P264	Joe Cocker is represented by music label ____.
	P276	London Jazz Festival is located in ____.
	P127	Border TV is owned by ____.
	P103	The native language of Mammootty is ____.
	P495	The Sharon Cuneta Show was created in ____.
ConceptNet	AtLocation	You are likely to find a overflow in a ____.
	CapableOf	Ravens can ____.
	CausesDesire	Joke would make you want to ____.
	Causes	Sometimes virus causes ____.
	HasA	Birds have ____.
	HasPrerequisite	Typing requires ____.
	HasProperty	Time is ____.
	MotivatedByGoal	You would celebrate because you are ____.
	ReceivesAction	Skills can be ____.
	UsedFor	A pond is for ____.

Table 3: Examples of generation for BERT-large.
with the associated log probability (in square brack

Model

Model	Base Model	#Parameters	Training Corpus	Corpus Size
fairseq-fconv (Dauphin et al., 2017)	ConvNet	324M	WikiText-103	103M Words
Transformer-XL (large) (Dai et al., 2019)	Transformer	257M	WikiText-103	103M Words
ELMo (original) (Peters et al., 2018a)	BiLSTM	93.6M	Google Billion Word	800M Words
ELMo 5.5B (Peters et al., 2018a)	BiLSTM	93.6M	Wikipedia (en) & WMT 2008-2012	5.5B Words
BERT (base) (Devlin et al., 2018a)	Transformer	110M	Wikipedia (en) & BookCorpus	3.3B Words
BERT (large) (Devlin et al., 2018a)	Transformer	340M	Wikipedia (en) & BookCorpus	3.3B Words

Table 1: Language models considered in this study.

Result

Fairseq-fconv　ELMo 5B
Transformer-XL　BERT-base
ELMo　BERT-large

Corpus	Relation	Statistics #Facts	Statistics #Rel	Baselines Freq	Baselines DrQA	KB RE_n	KB RE_o	LM Fs	LM Txl	LM Eb	LM E5B	LM Bb	LM Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Table 2: Mean precision at one (P@1) for a frequency baseline (Freq), DrQA, a relation extraction with naïve entity linking (RE_n), oracle entity linking (RE_o), fairseq-fconv (Fs), Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl) across the set of evaluation corpora.

※本文中に記載
SQuADにてP@10で比較したところ、
BERT-Large　57.1
DrQA　63.5

Result

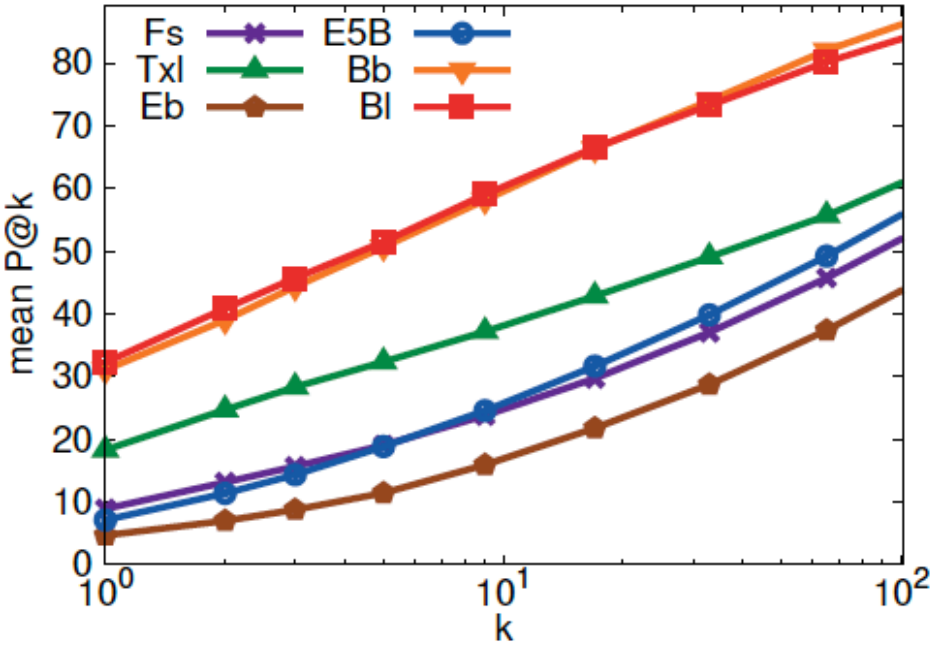


Figure 2: Mean P@k curve for T-REx varying k. Base-10 log scale for X axis.

Analysis

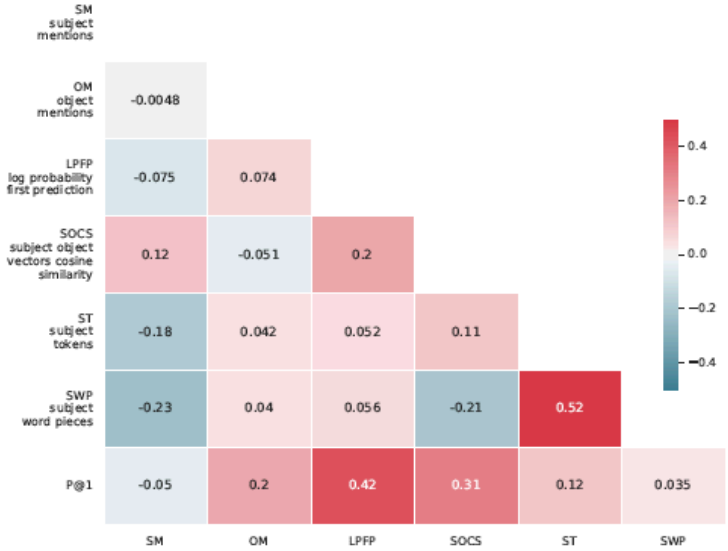


Figure 3: Pearson correlation coefficient for the P@1 of the BERT-large model on T-REx and a set of metrics: SM and OM refer to the number of times a subject and an object are mentioned in the BERT training corpus[4]respectively; LPFP is the log probability score associated with the first prediction; SOCS is the cosine similarity between subject and object vectors (we use spaCy[5]); ST and SWP are the number of tokens in the subject with a standard tokenization and the BERT WordPiece tokenization respectively.

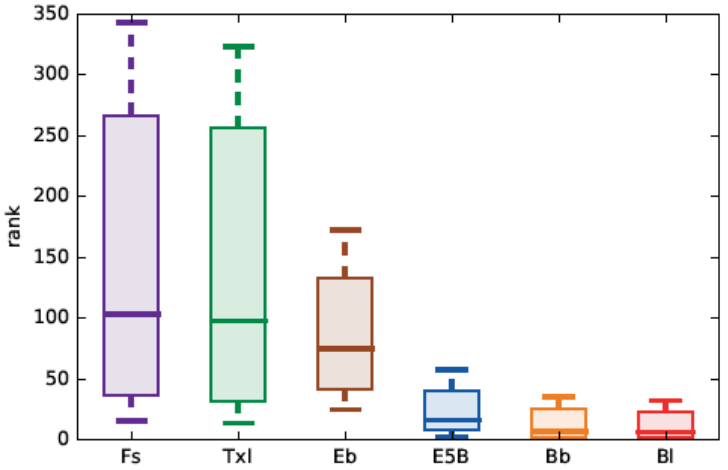


Figure 4: Average rank distribution for 10 different mentions of 100 random facts per relation in T-REx. ELMo 5.5B and both variants of BERT are least sensitive to the framing of the query but also are the most likely to have seen the query sentence during training.

生成例

	Relation	Query	Answer	Generation
T-Rex	P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8], **Florence [-1.8]**, Naples [-1.9], Milan [-2.4], Bologna [-2.5]
	P20	Adolphe Adam died in ____.	Paris	**Paris [-0.5]**, London [-3.5], Vienna [-3.6], Berlin [-3.8], Brussels [-4.0]
	P279	English bulldog is a subclass of ____.	dog	dogs [-0.3], breeds [-2.2], **dog [-2.4]**, cattle [-4.3], sheep [-4.5]
	P37	The official language of Mauritius is ____.	English	**English [-0.6]**, French [-0.9], Arabic [-6.2], Tamil [-6.7], Malayalam [-7.0]
	P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0], center [-2.2], **midfielder [-2.4]**, forward [-2.4], midfield [-2.7]
	P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], **Hamburg [-7.5]**, Ludwig [-7.5]
	P364	The original language of Mon oncle Benjamin is ____.	French	**French [-0.2]**, Breton [-3.3], English [-3.8], Dutch [-4.2], German [-4.9]
	P54	Dani Alves plays with ____.	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]
	P106	Paul Toungui is a ____ by profession .	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]
	P527	Sodium sulfide consists of ____.	sodium	water [-1.2], sulfur [-1.7], **sodium [-2.5]**, zinc [-2.8], salt [-2.9]
	P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], **Labor [-2.9]**
	P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0], **Uganda [-3.2]**, Tanzania [-3.5], China [-3.6], Pakistan [-3.6]
	P176	iPod Touch is produced by ____.	Apple	**Apple [-1.6]**, Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]
	P30	Bailey Peninsula is located in ____.	Antarctica	**Antarctica [-1.4]**, Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]
	P178	JDK is developed by ____.	Oracle	IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]
	P1412	Carl III used to communicate in ____.	Swedish	German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]
	P17	Sunshine Coast, British Columbia is located in ____.	Canada	**Canada [-1.2]**, Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]
	P39	Pope Clement VII has the position of ____.	pope	cardinal [-2.4], Pope [-2.5], **pope [-2.6]**, President [-3.1], Chancellor [-3.2]
	P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6], BMG [-2.6], Universal [-2.8], **Capitol [-3.2]**, Columbia [-3.3]
	P276	London Jazz Festival is located in ____.	London	**London [-0.3]**, Greenwich [-3.2], Chelsea [-4.0], Camden [-4.6], Stratford [-4.8]
	P127	Border TV is owned by ____.	ITV	Sky [-3.1], **ITV [-3.3]**, Global [-3.4], Frontier [-4.1], Disney [-4.3]
	P103	The native language of Mammootty is ____.	Malayalam	**Malayalam [-0.2]**, Tamil [-2.1], Telugu [-4.8], English [-5.2], Hindi [-5.6]
	P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2], **Philippines [-3.6]**, February [-3.7], December [-3.8], Argentina [-4.0]
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], **drain [-3.6]**
	CapableOf	Ravens can ____.	fly	**fly [-1.5]**, fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7], die [-1.7], **laugh [-2.0]**, vomit [-2.6], scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2], cancer [-2.0], **infection [-2.6]**, plague [-3.3], fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8], nests [-3.1], **feathers [-3.2]**, died [-3.7], eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], **speed [-4.1]**
	HasProperty	Time is ____.	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4], human [-3.3], **alive [-3.3]**, young [-3.6], free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], **fish [-2.8]**, recreation [-3.1]

Table 3: Examples of generation for BERT-large. The last column reports the top five tokens generated together with the associated log probability (in square brackets).

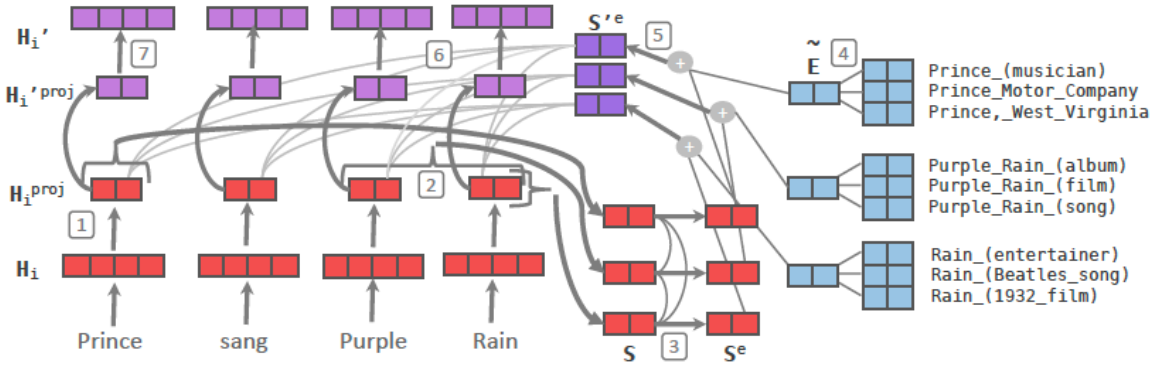
Pros and Cons

- Pros
  - LMに学習されている知識の精度を定量的に評価
  - 幅広いLMで統一的に評価
  - LAMA probeの提案
- Cons
  - LAMA probeはやや人工的に見える (single tokenの制約, relationの制約)
  - Multi tokenで構成されるentityについては未評価
  - Fine-tuneの影響は未評価

直接には比較不可能だったLMに学習されている関係知識を、
LAMA probeを作成することで統一的に比較可能にした

モデルを統一的に比較するために、
比較する関係知識の種類には制約が加わった

参考: Knowledge Enhanced Contextual Word Representations @EMNLP2019



System	F_1
WN-first sense baseline	65.2
ELMo	69.2
$BERT_{BASE}$	73.1
$BERT_{LARGE}$	73.9
KnowBert-WordNet	74.9
KnowBert-W+W	**75.1**

Table 2: Fine-grained WSD F_1.

System	AIDA-A	AIDA-B
Daiber et al. (2013)	49.9	52.0
Hoffart et al. (2011)	68.8	71.9
Kolitsas et al. (2018)	**86.6**	**82.6**
KnowBert-Wiki	80.2	74.4
KnowBert-W+W	82.1	73.7

Table 3: End-to-end entity linking strong match, micro averaged F_1.

System	LM	P	R	F_1
Zhang et al. (2018)	—	69.9	63.3	66.4
Alt et al. (2019)	GPT	70.1	65.0	67.4
Shi and Lin (2019)	$BERT_{BASE}$	73.3	63.1	67.8
Zhang et al. (2019)	$BERT_{BASE}$	70.0	66.1	68.0
Soares et al. (2019)	$BERT_{LARGE}$	—	—	70.1
Soares et al. (2019)	$BERT_{LARGE}$†	—	—	**71.5**
KnowBert-W+W	$BERT_{BASE}$	71.6	71.4	**71.5**

Table 4: Single model test set results on the TACRED relationship extraction dataset. † with MTB pretraining.

System	LM	F_1
Wang et al. (2016)	—	88.0
Wang et al. (2019b)	$BERT_{BASE}$	89.0
Soares et al. (2019)	$BERT_{LARGE}$	89.2
Soares et al. (2019)	$BERT_{LARGE}$†	**89.5**
KnowBert-W+W	$BERT_{BASE}$	89.1

Table 5: Test set F_1 for SemEval 2010 Task 8 relationship extraction. † with MTB pretraining.