BERT is to NLP what AlexNet is to CV: Can Pre-Trained Language Models Identify Analogies?

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Asahi Ushio
Luis Espinosa-Anke
Steven Schockaert
Jose Camacho-Collados

https://arxiv.org/abs/2105.04949

https://github.com/asahi417/analogy-language-model

Language Model Understanding

Model Analysis

- Hewitt 2019, Tenney 2019 → The embeddings capture linguistics knowledge.
- Clark 2020 → The attention reflects dependency.

Factual Knowledge

- Petroni 2019 → LM can be used as a commonsense KB.

Generalization Capacity

- Warstadt 2020 → LMs need large data to achieve linguistic generalization.
- Min 2020 → LMs' poor performance on adversarial data can be improved by DA.

Can LMs identify analogies?

Why Analogies?

Query:		word:language
Candidates:	(1)	paint:portrait
	(2)	poetry:rhythm
	(3)	**note:music**
	(4)	tale:story
	(5)	week:year

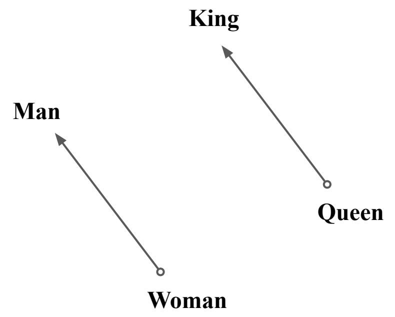
Sample from SAT analogy dataset.

Why Analogies?

Query:		word:language
Candidates:	(1)	paint:portrait
	(2)	poetry:rhythm
	(3)	**note:music**
	(4)	tale:story
	(5)	week:year

Sample from SAT analogy dataset.

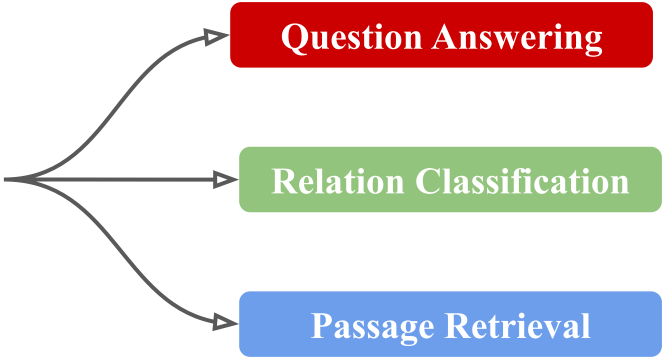
Analogy in word embedding



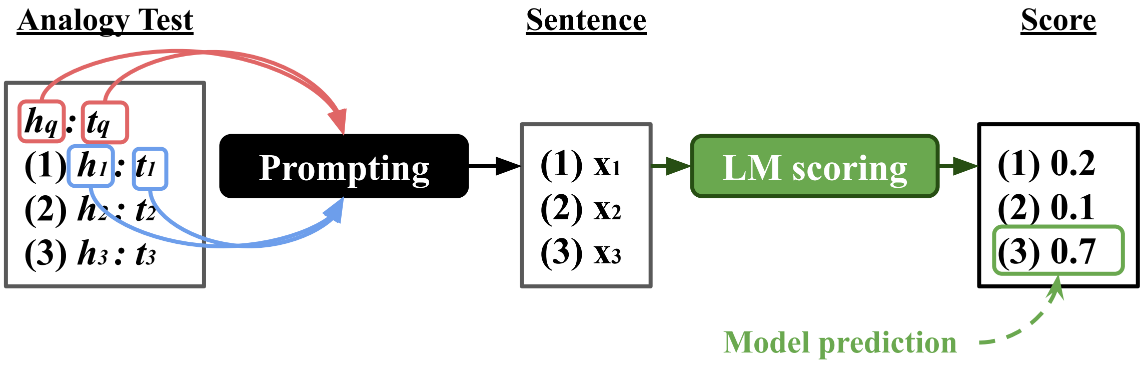
Why Analogies?

Query:		word:language
Candidates:	(1)	paint:portrait
	(2)	poetry:rhythm
	(3)	**note:music**
	(4)	tale:story
	(5)	week:year

Sample from SAT analogy dataset.



Solving Analogies with LMs



Eg) *word:language*

(1) *paint:portrait* → *word* is to *language* as *paint* is to *portrait* → Compute perplexity

(2) *note:music* → *word* is to *language* as *note* is to *music* → Compute perplexity

Prompt types

Type	Template
to-as	$[w_1]$ is to $[w_2]$ as $[w_3]$ is to $[w_4]$
to-what	$[w_1]$ is to $[w_2]$ What $[w_3]$ is to $[w_4]$
rel-same	The relation between $[w_1]$ and $[w_2]$ is the same as the relation between $[w_3]$ and $[w_4]$.
what-to	what $[w_1]$ is to $[w_2]$, $[w_3]$ is to $[w_4]$
she-as	She explained to him that $[w_1]$ is to $[w_2]$ as $[w_3]$ is to $[w_4]$
as-what	As I explained earlier, what $[w_1]$ is to $[w_2]$ is essentially the same as what $[w_3]$ is to $[w_4]$.

Scoring Functions

- Perplexity (PPL)
- Approximated point-wise mutual information (PMI)
- Marginal likelihood biased perplexity (mPPL)

Datasets

Dataset	Data size (val / test)	No. candidates	No. groups
SAT	37 / 337	5	2
UNIT 2	24 / 228	5,4,3	9
UNIT 4	48 / 432	5,4,3	5
Google	50 / 500	4	2
BATS	199 / 1799	4	3

Result (zeroshot)

	Model	Score	Tuned	SAT	U2	U4	Google	BATS	Avg
LM	BERT	s_{PPL}		32.9	32.9	34.0	80.8	61.5	48.4
			✓	39.8	41.7	41.0	86.8	67.9	55.4
		s_{PMI}		27.0	32.0	31.2	74.0	59.1	44.7
			✓	40.4	42.5	27.8	87.0	68.1	53.2
		s_{mPPL}	✓	41.8	44.7	41.2	88.8	67.9	56.9
	GPT-2	s_{PPL}		35.9	41.2	44.9	80.4	63.5	53.2
			✓	50.4	48.7	51.2	93.2	75.9	63.9
		s_{PMI}		34.4	44.7	43.3	62.8	62.8	49.6
			✓	51.0	37.7	50.5	91.0	79.8	62.0
		s_{mPPL}	✓	**56.7**	50.9	49.5	95.2	**81.2**	66.7
	RoBERTa	s_{PPL}		42.4	49.1	49.1	90.8	69.7	60.2
			✓	53.7	57.0	55.8	93.6	80.5	68.1
		s_{PMI}		35.9	42.5	44.0	60.8	60.8	48.8
			✓	51.3	49.1	38.7	92.4	77.2	61.7
		s_{mPPL}	✓	53.4	**58.3**	**57.4**	93.6	78.4	**68.2**
WE	FastText	-		47.8	43.0	40.7	**96.6**	72.0	60.0
	GloVe	-		47.8	46.5	39.8	96.0	68.7	59.8
	Word2vec	-		41.8	40.4	39.6	93.2	63.8	55.8
Base	PMI	-		23.3	32.9	39.1	57.4	42.7	39.1
	Random	-		20.0	23.6	24.2	25.0	25.0	23.6

RoBERTa is the best in **U2** & **U4** but otherwise **FastText** owns it 🤔

Result (tune on val)

	Model	Score	Tuned	SAT	U2	U4	Google	BATS	Avg
LM	BERT	s_{PPL}		32.9	32.9	34.0	80.8	61.5	48.4
			✓	39.8	41.7	41.0	86.8	67.9	55.4
		s_{PMI}		27.0	32.0	31.2	74.0	59.1	44.7
			✓	40.4	42.5	27.8	87.0	68.1	53.2
		s_{mPPL}	✓	41.8	44.7	41.2	88.8	67.9	56.9
	GPT-2	s_{PPL}		35.9	41.2	44.9	80.4	63.5	53.2
			✓	50.4	48.7	51.2	93.2	75.9	63.9
		s_{PMI}		34.4	44.7	43.3	62.8	62.8	49.6
			✓	51.0	37.7	50.5	91.0	79.8	62.0
		s_{mPPL}	✓	**56.7**	50.9	49.5	95.2	**81.2**	66.7
	RoBERTa	s_{PPL}		42.4	49.1	49.1	90.8	69.7	60.2
			✓	53.7	57.0	55.8	93.6	80.5	68.1
		s_{PMI}		35.9	42.5	44.0	60.8	60.8	48.8
			✓	51.3	49.1	38.7	92.4	77.2	61.7
		s_{mPPL}	✓	53.4	**58.3**	**57.4**	93.6	78.4	**68.2**
WE	FastText	-		47.8	43.0	40.7	**96.6**	72.0	60.0
	GloVe	-		47.8	46.5	39.8	96.0	68.7	59.8
	Word2vec	-		41.8	40.4	39.6	93.2	63.8	55.8
Base	PMI	-		23.3	32.9	39.1	57.4	42.7	39.1
	Random	-		20.0	23.6	24.2	25.0	25.0	23.6

BERT still worse 🧐
but
RoBERTa & **GPT2**
achieve the best 🤗

Results (SAT full)

	Model	Score	Tuned	Accuracy
LM	BERT	s_{PPL}		32.6
			✓	40.4*
		s_{PMI}		26.8
			✓	41.2*
		s_{mPPL}	✓	42.8*
	GPT-2	s_{PPL}		41.4
			✓	56.2*
		s_{PMI}		34.7
			✓	56.8*
		s_{mPPL}	✓	57.8*
	RoBERTa	s_{PPL}		49.6
			✓	55.8*
		s_{PMI}		42.5
			✓	54.0*
		s_{mPPL}	✓	55.8*
	GPT-3	*Zero-shot*		*53.7*
		Few-shot	✓	*65.2**
-	LRA	-		*56.4*
WE	FastText	-		49.7
	GloVe	-		48.9
	Word2vec	-		42.8
Base	PMI	-		23.3
	Random	-		20.0

Conclusion

- Some LMs can solve analogies in a true zero-shot setting to some extent.
- Language models are better than word embeddings at understanding abstract relations, but have ample room for improvement.
- Language models are very sensitive to hyperparameter tuning in this task, and careful tuning leads to competitive results.

🌳 Thank you! 🌳

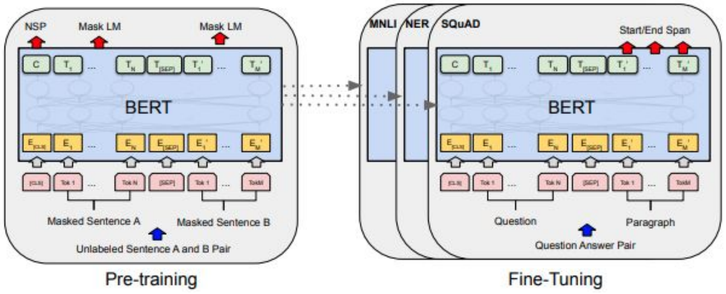
Story

LM is very good at all downstream tasks
→ Recent studies have further confirmed the linguistic semantics encoded in LM in a various way.
→ Also factual knowledge probing shows the capacity of LM
→ what about relational knowledge? Like w2v?
→ we did research on it! The result?
→ Very bad
→ With validation set, some LMs outperforms baseline
→ CONCLUSION

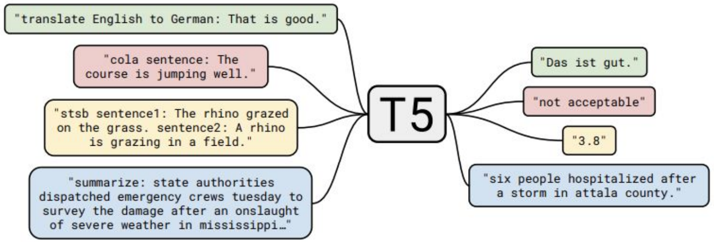
- Some language model represents relation knowledge
- With carefully tuned method, some LM can achieve very high accuracy (SoTa)

→ Future work: prompt mining, supervision

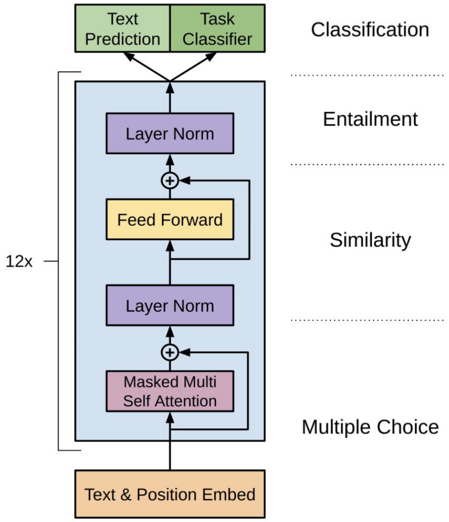
Language Model Pretraining



[BERT (Devlin, 2018)]



[T5 (Raffel, 2020)]



[GPT (Radford, 2018)]

Permutation Invariance

**Analogical Proportion Score*

$$AP(h_q, t_q, h_i, t_i) = \mathcal{A}_{g_{\text{pos}}}(\boldsymbol{p}) - \beta \cdot \mathcal{A}_{g_{\text{neg}}}(\boldsymbol{n})$$

$$\boldsymbol{p} = [s(a, b|c, d)]_{(a:b,c:d) \in \boldsymbol{P}}$$

$$\boldsymbol{n} = [s(a, b|c, d)]_{(a:b,c:d) \in \boldsymbol{N}}$$

Permutations of (a:b) and (c:d)

Positive

1. a : b :: c : d
2. a : c :: b : d
3. b : a :: d : c
4. b : d :: a : c
5. c : d :: a : b
6. c : a :: d : b
7. d : c :: b : a
8. d : b :: c : a

Negative

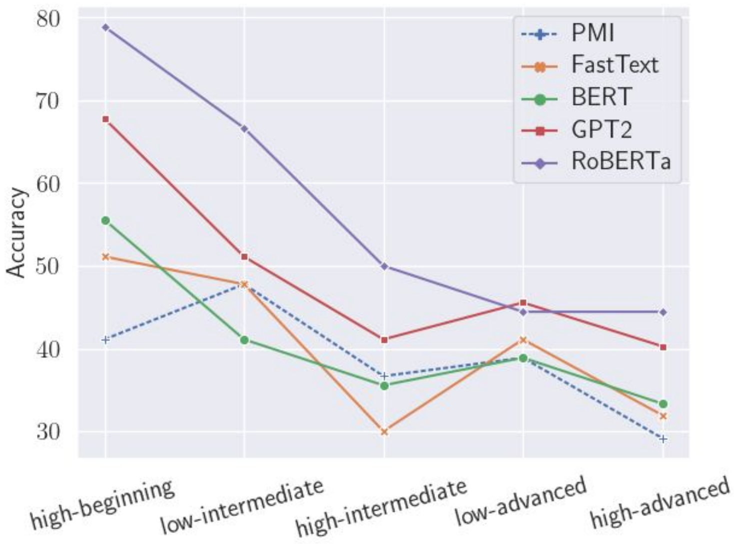
1. a : b :: d : c
2. a : c :: d : b
3. a : d :: b : c
4. a : d :: c : b
5. b : a :: c : d
6. b : c :: a : d
7. b : c :: d : a
8. b : d :: c : a
9. c : a :: b : d
10. c : b :: a : d
11. c : b :: d : a
12. c : d :: b : a
13. d : a :: b : c
14. d : a :: c : b
15. d : b :: a : c
16. d : c :: a : b

eg)

"word is to language as note is to music" = "language is to word as music is to note"

"word is to language as note is to music" ≠ "language is to word as note is to music"

Difficulty Level Breakdown (U2 & U4)



UNIT 4

UNIT 2