Toward a Better Understanding of Relational Knowledge in Language Models

NLPコロキウム
26th January 2022

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Asahi Ushio
Ph.D in Computer Science & Informatics, Cardiff University
HP: https://asahiushio.com

About Me

Ph.D at Cardiff University (2020~2023): Jose Camacho-Collados, Steven Schockaert

Internship at Amazon (2021 summer): Danushka Bollegala

Internship at snapchat (2021 winter): Leonardo Nerve, Francesco Barbieri

Projects:

- Relational Knowledge Probing of Language Model (Analogy LM, RelBERT)
- Question Generation (AutoQG)
- Named-Entity Recognition

Funs: Art, Whisky, Dance, Music

Social: Twitter, LinkedIn, GitHub

Outline

How much relational knowledge do pre-trained language models have?
- "BERT is to NLP what AlexNet is to CV: Can Pre-Trained Language Models Identify Analogies?", ACL 2021

If they have, what is the best way to purify the knowledge from the pre-trained language models?
- "Distilling Relation Embeddings from Pretrained Language Models", EMNLP 2021

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

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

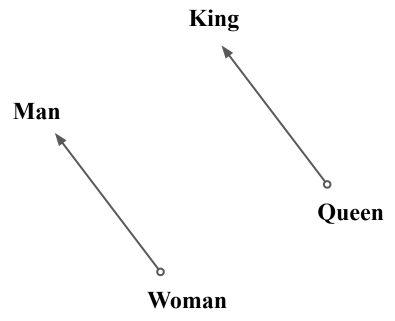
Sample from SAT analogy dataset.

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.

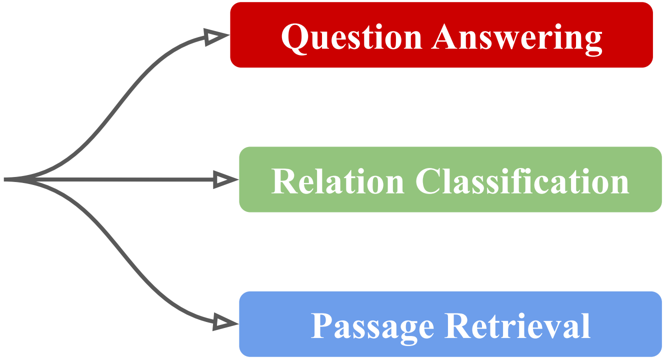
Analogy in word embedding



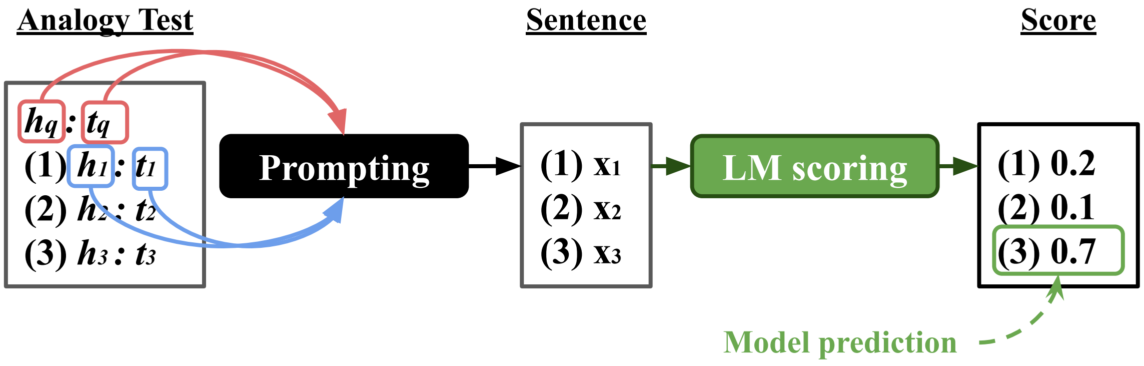
Why Analogies?

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



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 $[w_4]$.

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

Scoring Functions

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

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”

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)

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

	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 is to NLP what AlexNet is to CV: Can Pre-Trained Language Models Identify Analogies?
Asahi Ushio, Luis Espinosa-Anke, Steven Schockaert, and Jose Camacho-Collados

Result (tune on val)

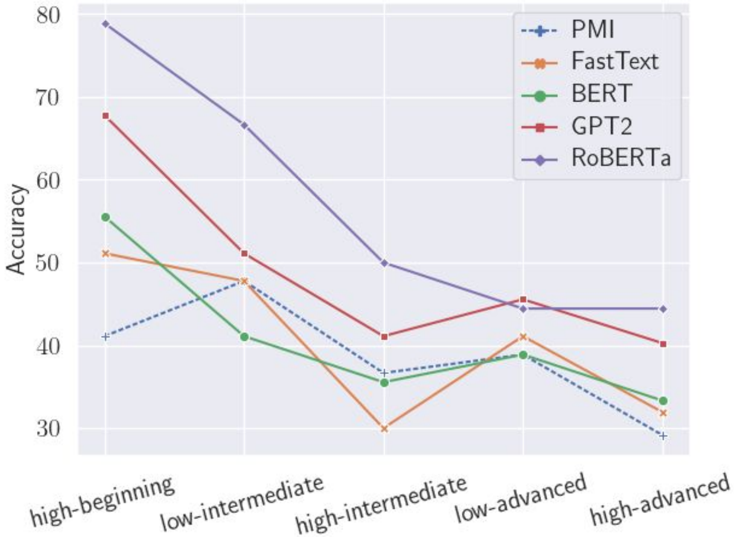
BERT still worse 🧐
but
RoBERTa & **GPT2**
achieve the best 🤗

	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

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

Difficulty Level Breakdown (U2 & U4)



UNIT 4

UNIT 2

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.

2. Distilling Relation Embeddings from Pre-trained Language Models

Distilling Relation Embeddings from Pre-trained Language Models

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Asahi Ushio
Jose Camacho-Collados
Steven Schockaert

https://github.com/asahi417/relbert

Language Model Understanding

Syntactic Knowledge

- Probing embedding: Hewitt 2019, Tenney 2019
- Probing attention weight: Clark 2020

Factual Knowledge a.k.a Language Model as a Commonsense KB

- Petroni 2019, Kassner 2020, Jiang 2020, etc

Relational Knowledge a.k.a Language Model as a Lexical Relation Reasoner

- LM fine-tuning on relation classification: Bouraoui 2019
- Vanilla LM evaluation: Ushio 2021

Language Model Understanding

Syntactic Knowledge

- Probing embedding: Hewitt 2019, Tenney 2019
- Probing attention weight: Clark 2020

Factual Knowledge a.k.a Language Model as a Commonsense KB

- Petroni 2019, Kassner 2020, Jiang 2020, etc

Relational Knowledge a.k.a Language Model as a Lexical Relation Reasoner

- LM fine-tuning on relation classification: Bouraoui 2019
- Vanilla LM evaluation: Ushio 2021

Language Model Understanding

Syntactic Knowledge

- Probing embedding: Hewitt 2019, Tenney 2019
- Probing attention weight: Clark 2020

Factual Knowledge a.k.a Language Model as a Commonsense KB

- Petroni 2019, Kassner 2020, Jiang 2020, etc

Relational Knowledge a.k.a Language Model as a Lexical Relation Reasoner

- LM fine-tuning on relation classification: Bouraoui 2019
- Vanilla LM evaluation: Ushio 2021

Can we distil relational knowledge as relation embedding?

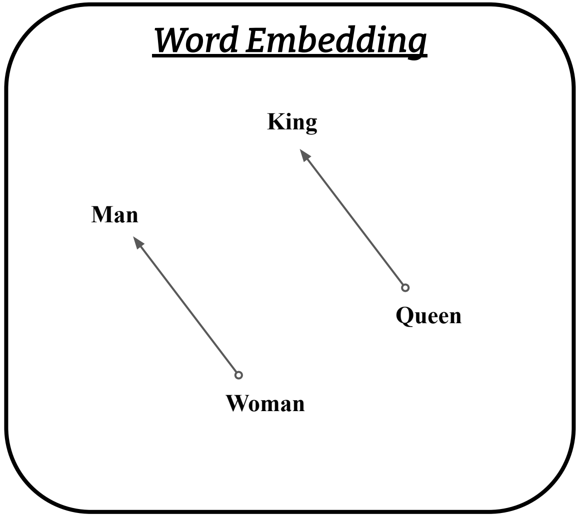
Relation Embedding

Word Embedding [Mikolov (2013)]

Pair2Vec [Joshi (2019)]

Relative [Camacho-Collados (2019)]

X	Y	Contexts
hot	*cold*	with **X** and **Y** baths too **X** or too **Y** neither **X** nor **Y**
Portland	*Oregon*	in **X**, **Y** the **X** metropolitan area in **Y** **X** International Airport in **Y**
crop	*wheat*	food **X** are maize, **Y**, etc dry **X**, such as **Y**, more **X** circles appeared in **Y** fields
Android	*Google*	**X** OS comes with **Y** play the **X** team at **Y** **X** is developed by **Y**



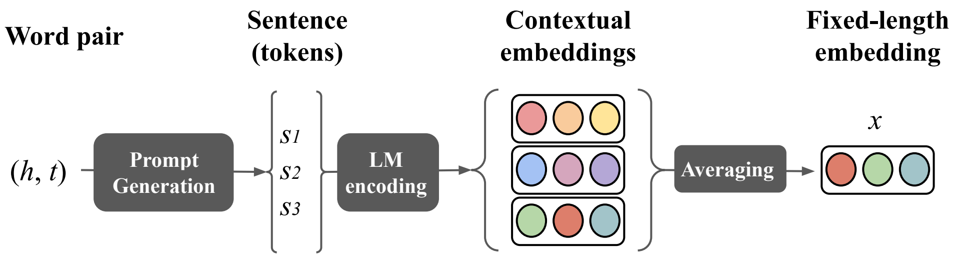
RelBERT

Relation Embedding from LM

Prompt Generation
Custom Template, AutoPrompt (Shin 2020), P-tuning (Liu 2021)

LM embedding

Averaging over the context



Relation Embedding from LM

Prompt Generation

Custom Template, AutoPrompt (Shin 2020), P-tu

LM embedding

Averaging over the context

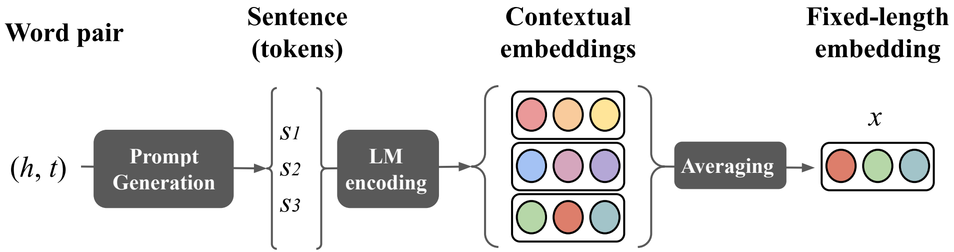
1. Today, I finally discovered the relation between **camera** and **photographer** : **camera** is the <mask> of **photographer**

2. Today, I finally discovered the relation between **camera** and **photographer** : **photographer** is **camera**'s <mask>

3. Today, I finally discovered the relation between **camera** and **photographer** : <mask>

4. I wasn't aware of this relationship, but I just read in the encyclopedia that **camera** is the <mask> of **photographer**

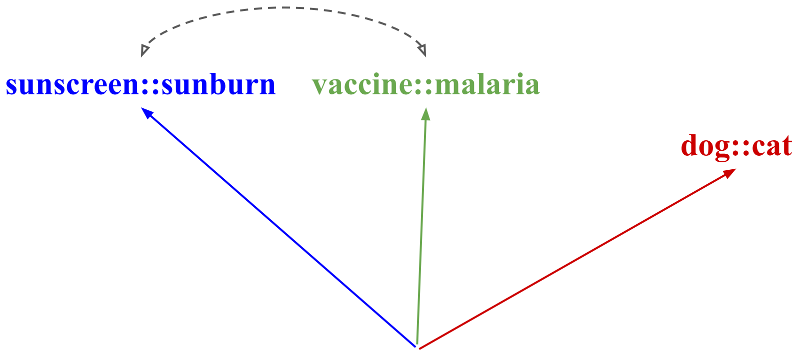
5. I wasn't aware of this relationship, but I just read in the encyclopedia that **photographer** is **camera**'s <mask>



Fine-tuning on Triples

Given a triple: anchor "`sunscreen::sunburn`", positive "`vaccine::malaria`", and negative "`dog::cat`", we want the embeddings of the anchor and the positive close but far from the negative.

Loss function: *Triplet loss* and *classification loss* following SBERT (Reimers 2019).



Fine-tuning on Triples

Given a triple of the anchor x_a (eg. "`sunscreen`"), the positive x_p (eg. "`sunburn`"), and the negative x_n (eg. "`evil`"), the **triplet loss** is defined as

$$L_t = \max\left(0, \|x_a - x_p\| - \|x_a - x_n\| + \varepsilon\right)$$

and the **classification loss** is defined as

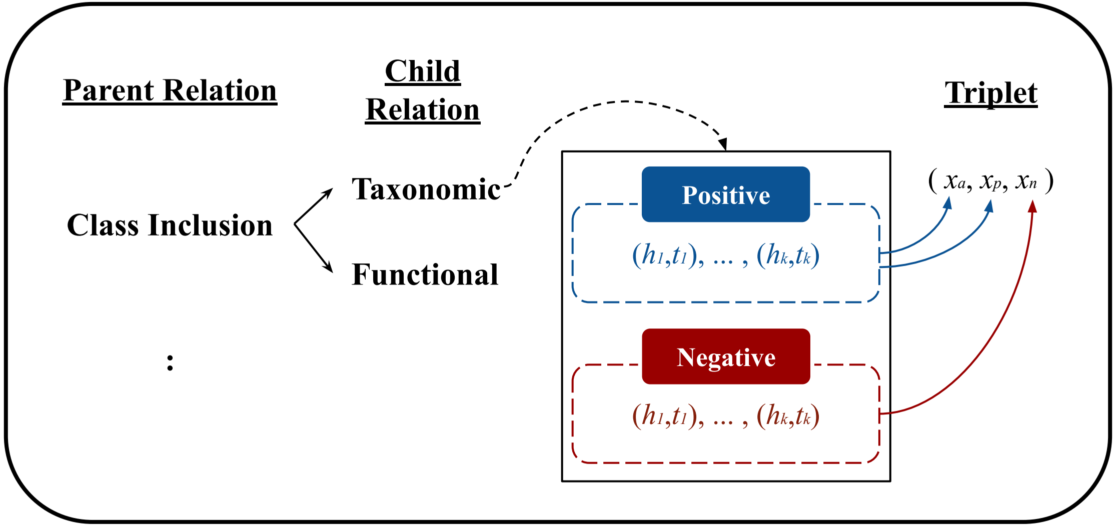
$$L_c = -\log(g(x_a, x_p)) - \log(1 - g(x_a, x_n))$$

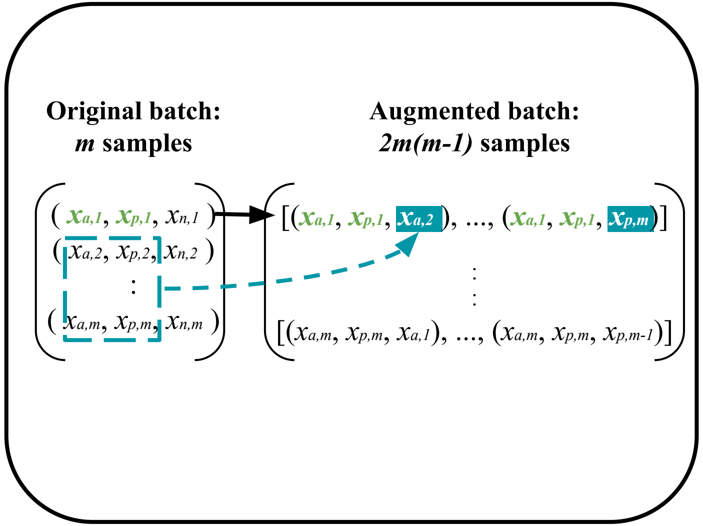
$$g(u, v) = \mathrm{sigmoid}(W \cdot [u \oplus v \oplus |v - u|]^T)$$

where W is a learnable weight. The loss functions are inspired by SBERT (Reimers 2019).

Dataset

We create the dataset from **SemEval 2012 Task 2**.





EXPERIMENTS

Experiment: Analogy

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

Sample from SAT analogy dataset.

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

Data statistics.

Setup

- Cosine similarity in between embeddings.
- No training.
- Accuracy as the metric.
- No validation.

Experiment: Analogy

SotA in 4 / 5 datasets 🎉

Better than tuned methods on dev set 🤗

Model	SAT†	SAT	U2	U4	Google	BATS
GPT-3 (zero)	*53.7*	-	-	-	-	-
GPT-3 (few)	*65.2**	-	-	-	-	-
RELATIVE	24.9	24.6	32.5	27.1	62.0	39.0
pair2vec	33.7	34.1	25.4	28.2	66.6	53.8
FastText	49.7	47.8	43.0	40.7	**96.6**	72.0
Analogical Proportion Score (tuned)						
· GPT-2	*57.8**	*56.7**	*50.9**	*49.5**	*95.2**	*81.2**
· BERT	*42.8**	*41.8**	*44.7**	*41.2**	*88.8**	*67.9**
· RoBERTa	*55.8**	*53.4**	*58.3**	*57.4**	*93.6**	*78.4**
RelBERT						
· Manual	**69.5**	**70.6**	**66.2**	**65.3**	92.4	**78.8**
· AutoPrompt	61.0	62.3	61.4	63.0	88.2	74.6
· P-tuning	54.0	55.5	58.3	55.8	83.4	72.1

Experiment: Classification

Setup

- Supervised Task
- LMs are frozen
- macro/micro F1
- Tuned on dev

	BLESS	CogALex	EVALution	K&H+N	ROOT09
Random	8,529/609/3,008	2,228/3,059	-	18,319/1,313/6,746	4,479/327/1,566
Meronym	2,051/146/746	163/224	218/13/86	755/48/240	-
Event	2,657/212/955	-	-	-	-
Hypernym	924/63/350	255/382	1,327/94/459	3,048/202/1,042	2,232/149/809
Co-hyponym	2,529/154/882	-	-	18,134/1,313/6,349	2,222/162/816
Attribute	1,892/143/696	-	903/72/322	-	-
Possession	-	-	377/25/142	-	-
Antonym	-	241/360	1,095/90/415	-	-
Synonym	-	167/235	759/50/277	-	-

Data statistics.

Experiment: Classification

SotA in 4 / 5 datasets in macro F1 score 🎉

SotA in 3 / 5 datasets in micro F1 score 🎉

Model		BLESS macro	BLESS micro	CogALexV macro	CogALexV micro	EVALution macro	EVALution micro	K&H+N macro	K&H+N micro	ROOT09 macro	ROOT09 micro
GloVe	*cat*	92.9	93.3	42.8	73.5	56.9	58.3	88.8	94.9	86.3	86.5
	cat+dot	**93.1**	93.7	51.9	79.2	55.9	57.3	89.6	95.1	88.8	89.0
	cat+dot+pair	91.8	92.6	56.4	81.1	58.1	59.6	89.4	95.7	89.2	89.4
	cat+dot+rel	91.1	92.0	53.2	79.2	58.4	58.6	89.3	94.9	89.3	89.4
	diff	91.0	91.5	39.2	70.8	55.6	56.9	87.0	94.4	85.9	86.3
	diff+dot	92.3	92.9	50.6	78.5	56.5	57.9	88.3	94.8	88.6	88.9
	diff+dot+pair	91.3	92.2	55.5	80.2	56.0	57.4	88.0	95.5	89.1	89.4
	diff+dot+rel	91.1	91.8	52.8	78.6	56.9	57.9	87.4	94.6	87.7	88.1
FastText	*cat*	92.4	92.9	40.7	72.4	56.4	57.9	88.1	93.8	85.7	85.5
	cat+dot	92.7	93.2	48.5	77.4	56.7	57.8	89.1	94.0	88.2	88.5
	cat+dot+pair	90.9	91.5	53.0	79.3	56.1	58.2	88.3	94.3	87.7	87.8
	cat+dot+rel	91.4	91.9	50.6	76.8	57.9	59.1	86.9	93.5	87.1	87.4
	diff	90.7	91.2	39.7	70.2	53.2	55.5	85.8	93.3	85.5	86.0
	diff+dot	92.3	92.9	49.1	77.8	55.2	57.4	86.5	93.6	88.5	88.9
	diff+dot+pair	90.0	90.8	53.9	79.0	55.8	57.8	86.6	94.2	87.7	88.1
	diff+dot+rel	90.6	91.3	53.6	78.2	57.1	58.0	86.3	93.4	86.9	87.4
RelBERT	Manual	91.7	92.1	**71.2**	**87.0**	68.4	69.6	88.0	96.2	**90.9**	**91.0**
	AutoPrompt	91.9	92.4	68.5	85.1	**69.5**	**70.5**	**91.3**	97.1	90.0	90.3
	P-tuning	91.3	91.8	67.8	84.9	69.1	70.2	88.5	96.3	89.8	89.9
SotA	LexNET	-	89.3	-	-	-	60.0	-	98.5	-	81.3
	SphereRE	-	**93.8**	-	-	-	62.0	-	**99.0**	-	86.1

ANALYSIS

Relation Memorarization

Does RelBERT just memorize the relations in the training set... ?

Experiment: Train RelBERT without hypernymy.

Result: No significant decrease in hyperbyny prediction.

→ RelBERT **does not** rely on the memorization!

	BLESS	CogALex	EVAL	K&H+N	ROOT09
rand	93.7 (+0.3)	94.3 (-0.2)	-	97.9 (+0.2)	91.2 (-0.1)
mero	89.8 (+1.4)	72.9 (+2.7)	69.2 (+1.9)	74.5 (+5.4)	-
event	86.5 (-0.3)	-	-	-	-
hyp	94.1 (+0.8)	60.9 (-0.7)	61.7 (-1.5)	93.5 (+5.0)	83.0 (-0.4)
cohyp	96.4 (+0.3)	-	-	97.8 (+1.2)	97.4 (-0.5)
attr	92.6 (+0.3)	-	84.7 (+1.6)	-	-
poss	-	-	67.1 (-0.2)	-	-
ant	-	76.8 (-2.6)	81.3 (-0.9)	-	-
syn	-	49.9 (-0.6)	53.6 (+2.7)	-	-
macro	92.2 (+0.5)	71.0 (-0.2)	69.3 (+0.9)	90.9 (+2.9)	90.5 (-0.4)
micro	92.5 (+0.4)	86.9 (-0.1)	70.2 (+0.6)	97.2 (+1.0)	90.7 (-0.3)

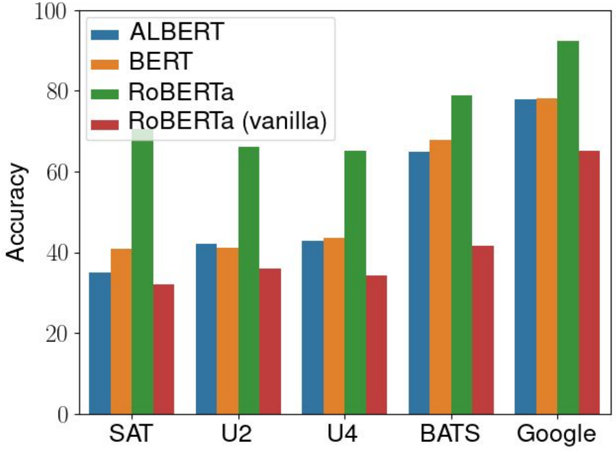
Fine-tuning? Other LMs?

Train RelBERT on BERT, ALBERT in addition to RoBERTa.

→ **RoBERTa is the best.**

Vanilla RoBERTa (no fine-tuning).

→ **Fine-tuning (distillation) is necessary.**



Conclusion

- We propose **RelBERT**, a framework to achieve relation embedding model based on pretrained LM.
- RelBERT **distil the LM's relational knowledge** and realize a **high quality relation embedding**.
- Experimental results show that **RelBERT embedding outperform existing baselines**, establishing **SotA in analogy and relation classification.**

Release of RelBERT Library

We release python package **relbert** (install via `pip install relbert`) along with model checkpoints on the huggingface modelhub.

Please check our project page https://github.com/asahi417/relbert !!

```
from relbert import RelBERT
model = RelBERT('asahi417/relbert-roberta-large')
# the vector has (1024,)
v_tokyo_japan = model.get_embedding(['Tokyo', 'Japan'])
```

Nearest Neighbours

Target	Nearest Neighbors
barista:coffee	baker:bread, brewer:beer, bartender:cocktail, winemaker:wine, bartender:drink, baker:cake
bag:plastic	bottle:plastic, bag:leather, container:plastic, box:plastic, jug:glass, bottle:glass
duck:duckling	chicken:chick, pig:piglet, cat:kitten, ox:calf, butterfly:larvae, bear:cub
cooked:raw	raw:cooked, regulated:unregulated, sober:drunk, loaded:unloaded, armed:unarmed, published:unpublished
chihuahua:dog	dachshund:dog, poodle:dog, terrier:dog, chinchilla:rodent, macaque:monkey, dalmatian:dog
dog:dogs	cat:cats, horse:horses, pig:pigs, rat:rats, wolf:wolves, monkey:monkeys
spy:espionage	pirate:piracy, robber:robbery, lobbyist:lobbying, scout:scouting, terrorist:terrorism, witch:witchcraft

Comparing to Word Embeddings

FastText is still better than RelBERT in Google Analogy Question.

Breakdown per relation types shows that FastText is better in the morphological relation, while very poor in the lexical relation.

Model	Google		BATS		
	Mor	Sem	Mor	Sem	Lex
FastText	95.4	98.1	90.4	71.1	33.8
RelBERT	89.8	95.8	87.0	66.2	75.1

Thank You!