Toward a Better Understanding of Relational Knowledge in Language Models

UCL NLP Meetup
3rd October 2022

Asahi Ushio
Ph.D in Computer Science & Informatics, Cardiff University
Cardiff NLP
https://asahiushio.com, @asahiushio , https://github.com/asahi417

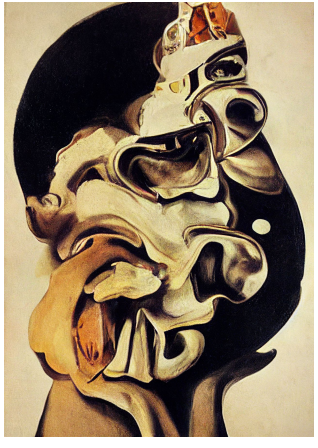
About Me

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 Neves, Francesco Barbieri
Research Interests:

- Relational Knowledge Representation: Analogy LM [ACL 2021], RelBERT [EMNLP 2021]
- Question Generation: AutoQG
- Twitter NLP: TweetNLP [EMNLP 2022], TweetTopic [COLING 2022], TweetNER7 [ACL 2022]

Funs: Art, Whisky, Dance, Jazz

Social: Twitter, LinkedIn, GitHub



Outline

What is relational knowledge and how we evaluate it?

- "BERT is to NLP what AlexNet is to CV: Can Pre-Trained Language Models Identify Analogies?", ACL 2021

Can we extract relational knowledge by fine-tuning?

- "Distilling Relation Embeddings from Pretrained Language Models", EMNLP 2021

Resources:

- HuggingFace: https://huggingface.co/relbert
- GitHub: https://github.com/asahi417/relbert, https://github.com/asahi417/analogy-language-model

Relational Knowledge and Word Analogies

Relational Knowledge

Understand the relation in between terms

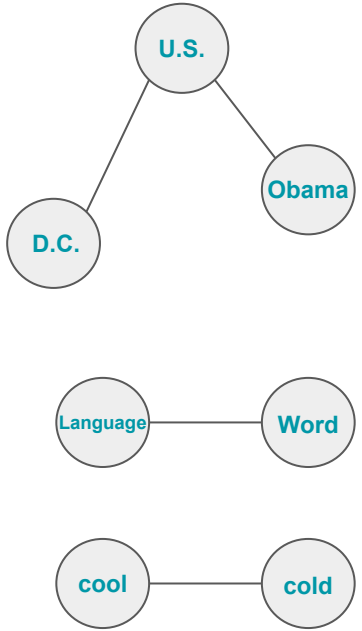
→ **(Washington D.C., U.S.) = a capital in the country**

→ **(Obama,U.S.) = a former president**

→ **(Word, Language) = a component**

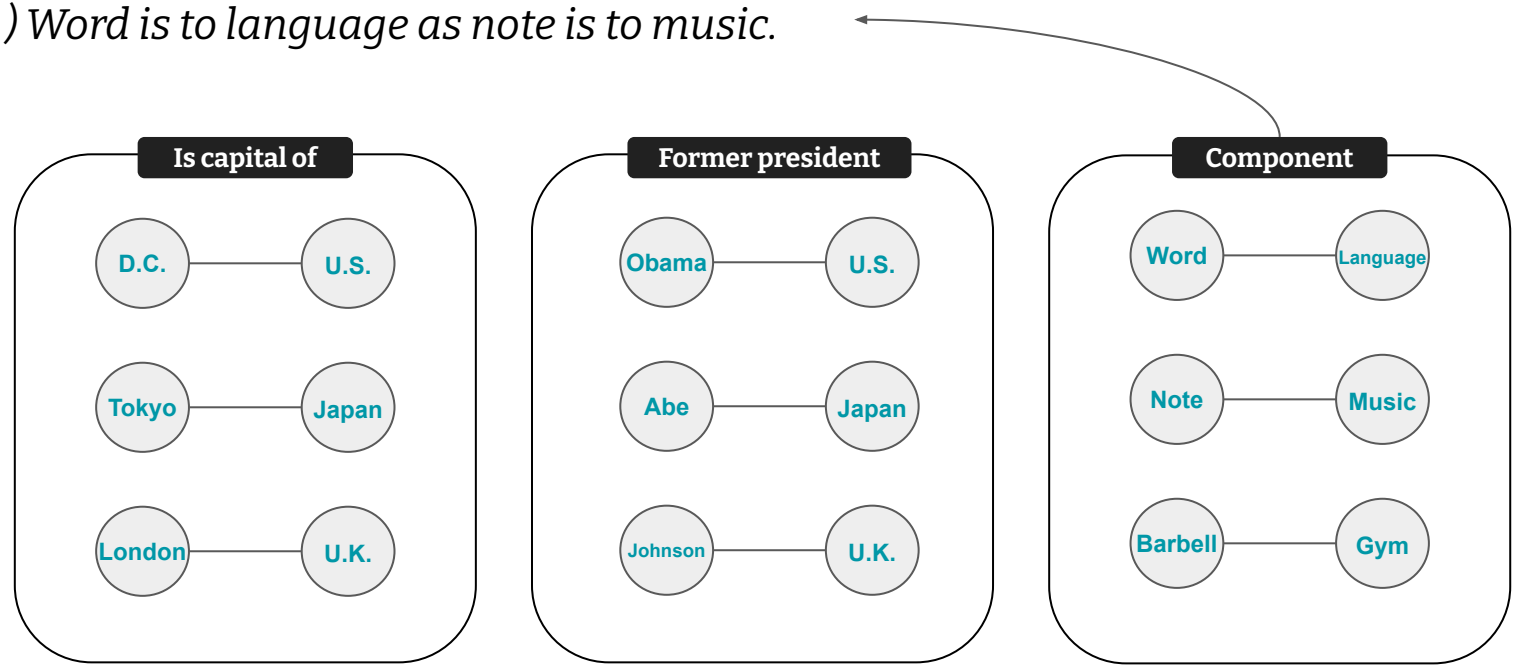
→ **(cool, cold) = synonym**

→ **(warm, warmer) = comparative**



Word Analogies

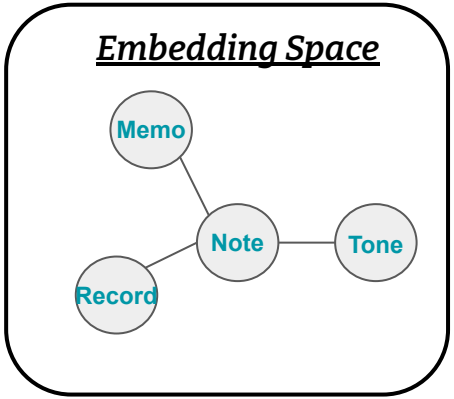
eg.) Word is to language as note is to music.



Analogical Reasoning

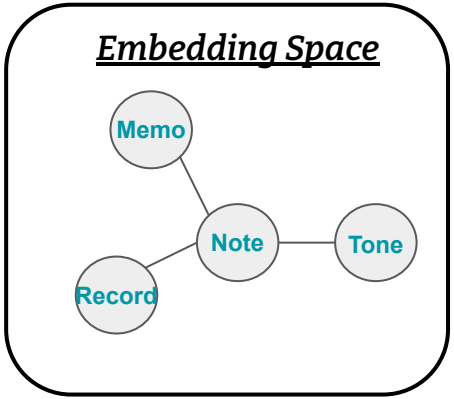
Question: What is the note?

- *Case-based Reasoning*
  - Nearest Neighbour → Note is “memo”.

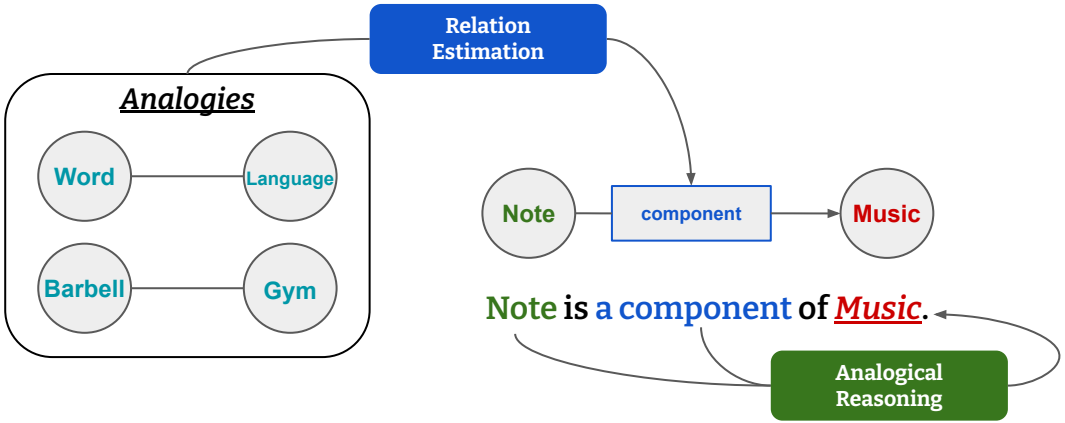


Analogical Reasoning

Question: What is the note?

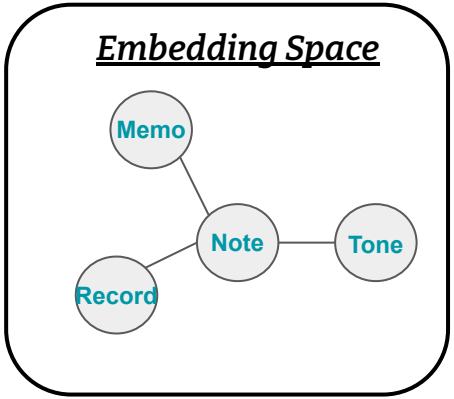


- *Case-based Reasoning*
  - Nearest Neighbour → Note is “memo”.
- *Analogical Reasoning* [Prade 2017]
  - Transfer knowledge from analogies → Note is “a component of music”.

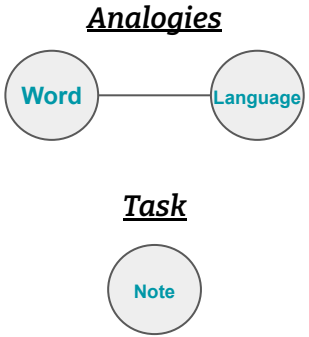


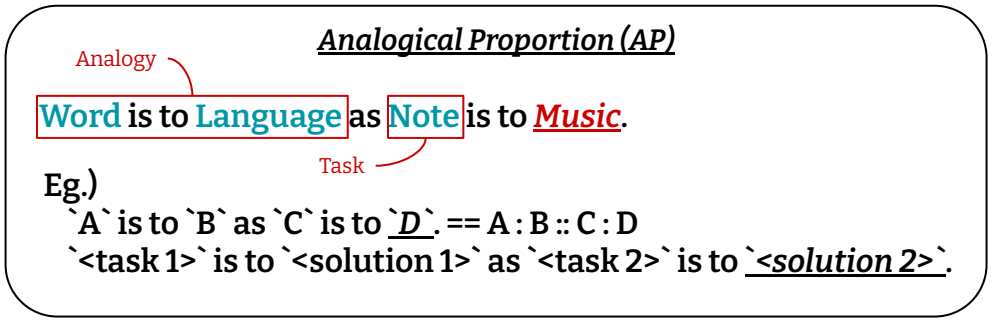
Analogical Reasoning

Question: What is the note?



- *Case-based Reasoning*
  - Nearest Neighbour → Note is “memo”.
- *Analogical Reasoning* [Prade 2017]
  - Transfer knowledge from analogies → Note is “a component of music”.





Related Tasks in NLP

Tasks that require an explicit understanding of relational knowledge in NLP.

- Analogy Question
- Relation Mapping Problem
- Lexical Relation Classification
- etc

Source	→ Target
solar system	→ atom
sun	→ nucleus
planet	→ electron
mass	→ charge
attracts	→ attracts
revolves	→ revolves
gravity	→ electromagnetism

Relation mapping problem [Turney 2008].

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

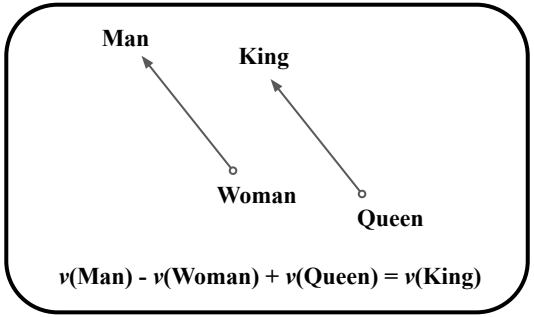
SAT analogy question [Turney 2003].

Solving Analogy with LMs!

Analogy Question & Solution with Word Embedding

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

SAT analogy question [Turney 2003].



Word Embedding [Mikolov 2013]

Solving Analogy with Word Embedding

$h_q : t_q$

(1) $h_1 : t_1$

(2) $h_2 : t_2$

(3) $h_3 : t_3$

$\rightarrow$

$r_q = v(h_q) - v(t_q)$

$r_1 = v(h_1) - v(t_1)$

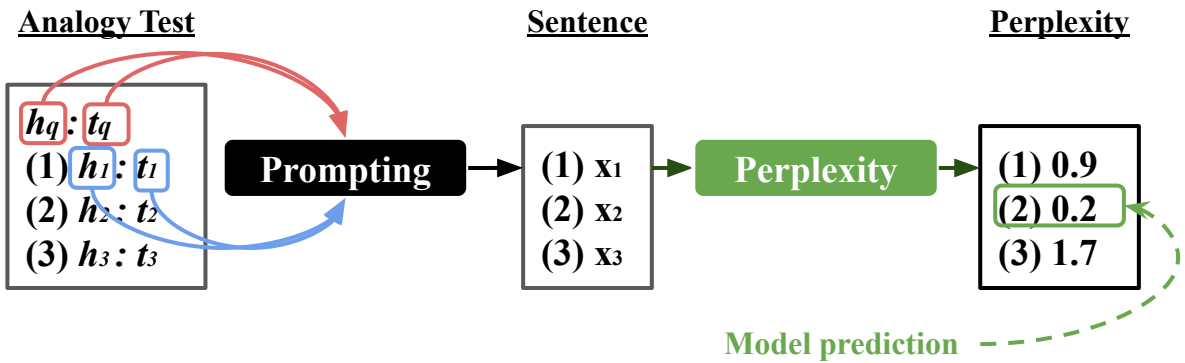
$r_2 = v(h_2) - v(t_2)$

$r_3 = v(h_3) - v(t_3)$

$\rightarrow$ $y = argmax\{sim(r_q, r_i)\}$

GPT3 to Solve Analogy

GPT3 [Brown 2020] uses template of "A" is to "B" as "C" is to "D" (analogical proportion).



Eg) *word:language*

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

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

Result on SAT Analogy Question

Remark

- Human performance is not high.
- LRA (SVD on a PMI matrix over large word-pair corpus) is competitive with recent methods.
- None of zero-shot outperforms human baselines including GPT3.

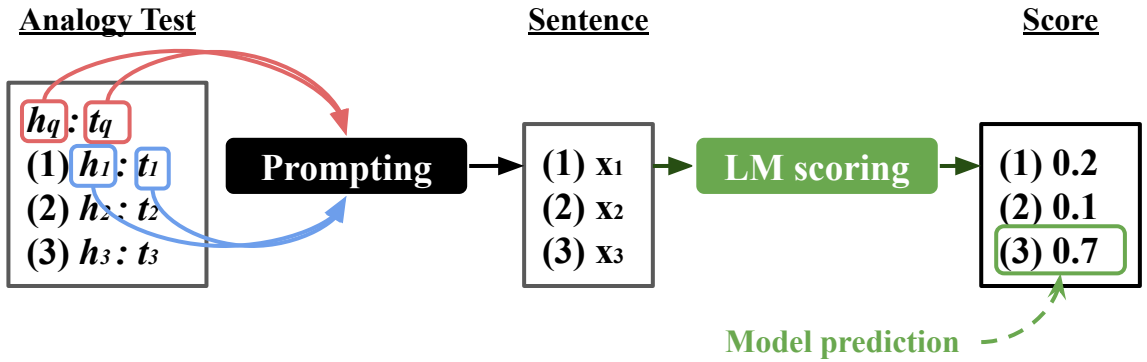
So, LMs are poor at understanding analogy... 😨

Model	**Accuracy on SAT [Turney 2003], [link]**
LRA [Turney 2005]	56.4
Word2Vec [Mikolov 2013]	42.8
GloVe [Pennington 2014]	48.9
FastText [Bojanowski 2017]	49.7
GPT3 (zeroshot) [Brown 2020]	53.7
GPT3 (fewshot) [Brown 2020]	65.2
Human [Turney 2005]	57.0

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

LM to Solve Analogy

Based on GPT3 approach, but generalize the perplexity to LM scoring with multiple template types.



Template 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 $[w_4]$.

LM Scoring

- ***Perplexity (or pseudo-perplexity for masked LMs [Wang 2019])***

$$f(\boldsymbol{x}) = \exp\left(-\sum_{j=1}^{m} \log P_{\text{auto}}(x_j|\boldsymbol{x}_{j-1})\right)$$

- ***Perplexity-based PMI***

$$r(t_i|h_i, h_q, t_q) = \log P(t_i|h_q, t_q, h_i) - \alpha \cdot \log P(t_i|h_q, t_q)$$

Parameter to control the effect of marginal likelihood [Feldman 2019].

*Conditional likelihood: $P(t_i|h_q, t_q, h_i) = -\frac{f(\mathcal{T}_t(h_q, t_q, h_i, t_i))}{\sum_{k=1}^{n} f(\mathcal{T}_t(h_q, t_q, h_i, t_k))}$

Text converted from (*hq : tq :: hi : ti)* with template *t*.

*Marginal likelihood: $P(t_i|h_q, t_q) = -\frac{\sum_{k=1}^{n} f(\mathcal{T}_t(h_q, t_q, h_k, t_i))}{\sum_{k=1}^{n}\sum_{l=1}^{n} f(\mathcal{T}_t(h_q, t_q, h_k, t_l))}$

- ***Marginal likelihood biased Perplexity (mPPL)***

$$s_{mPPL}(t_i, h_i|h_q, t_q) = \log f(\mathcal{T}_t(h_q, t_q, h_i, t_i)) - \alpha_t \cdot \log P(t_i|h_q, t_q) - \alpha_h \cdot \log P(h_i|h_q, t_q)$$

Marginal likelihood of tail and head with parameters to control the effect.

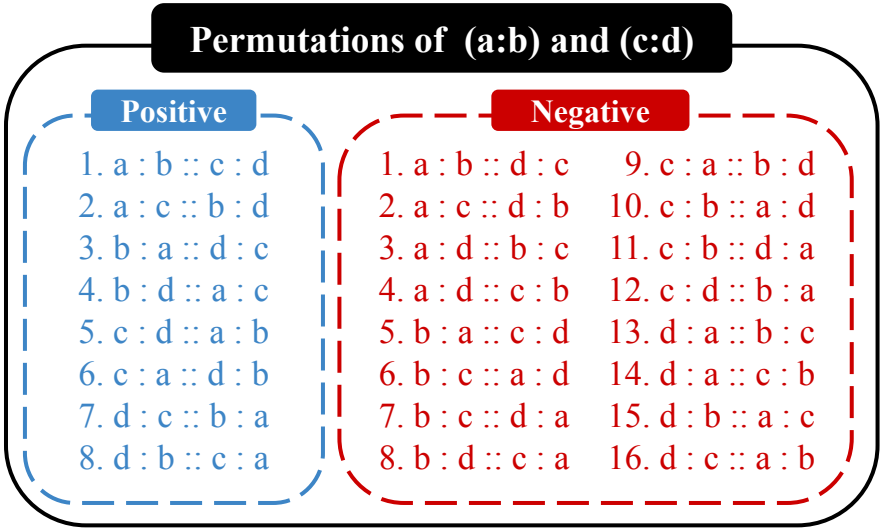
Perplexity

Permutation of Analogical Proportion

Analogical Proportion (AP)
→ *"a" is to "b" as "c" is to "d"* or *a : b :: c : d*

Permutation Invariance [Barbot 2019]

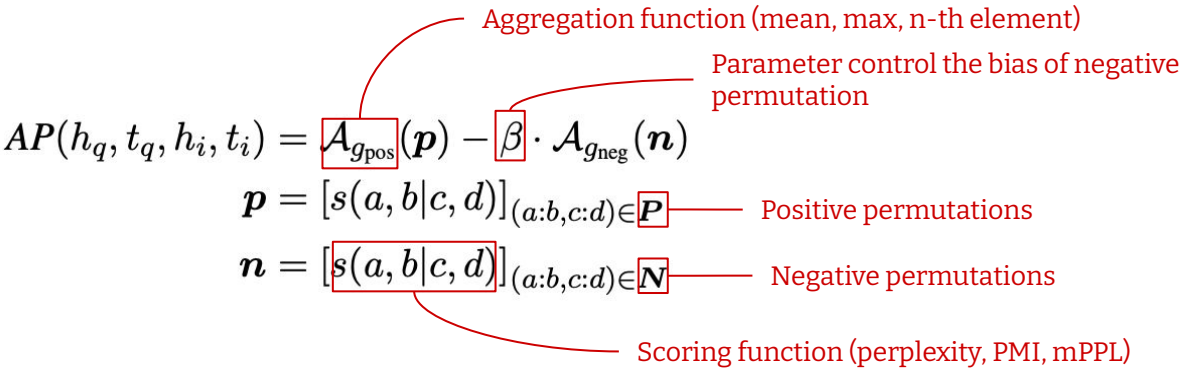
- Positive permutation:
  - Statement is still the same analogy.
- Negative permutation:
  - Statement is **not** the same analogy.



Example:

- *"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"*

Analogical Proportion (AP) Score



AP Score: Relative gain of LM score from negative permutations to positive permutations.

Key assumption: Better scoring function should give low score on the negative permutations and high score on the positive permutations.

Analogy Question 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

SAT [Turney 2003]: US college admission test

UNIT2 [Boteanu 2015]: US college applicants in grade of 4 to 12 (age 9 onwards)

UNIT4: Same as UNIT2 but organized by high-beginning (SAT level), low-intermediate, high-intermediate, low-advanced, and high-advanced (GRE level).

Google [Mikolov 2013]: A mix of semantic and morphological relations (*capital-of*, *singular-plural*, *etc*)

BATS [Gladkova 2016]: A mix of *lexicographic*, *encyclopedic*, and *derivational* and *inflectional morphology*

Results (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
		s_{PMI}		27.0	32.0	31.2	74.0	59.1	44.7
		s_{mPPL}							
	GPT-2	s_{PPL}		35.9	41.2	44.9	80.4	63.5	53.2
		s_{PMI}		34.4	44.7	43.3	62.8	62.8	49.6
		s_{mPPL}							
	RoBERTa	s_{PPL}		42.4	49.1	49.1	90.8	69.7	60.2
		s_{PMI}		35.9	42.5	44.0	60.8	60.8	48.8
		s_{mPPL}							
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 (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

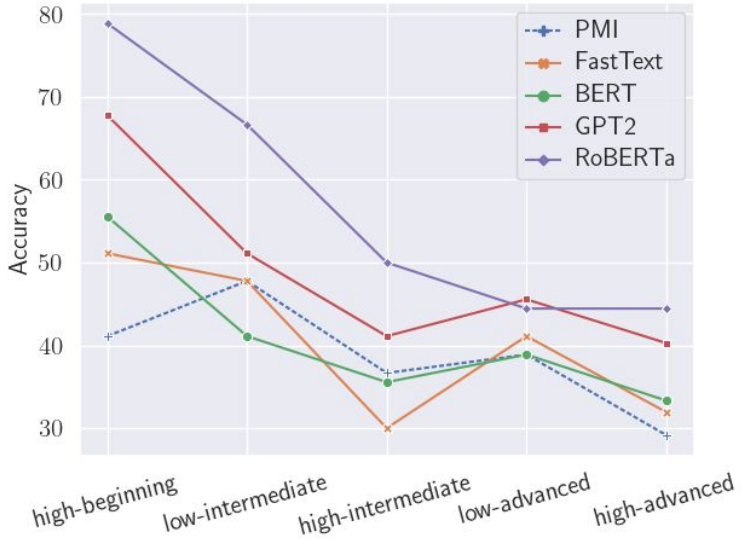
Results (SAT full)

Few-shot models outperform word embeddings and LRA.

	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)

UNIT2



UNIT4

Conclusion

GPT3 result was dissapointing given the model size, but we show some of the smaller LMs (RoBERTa and GPT2) can solve analogies in a true zero-shot setting to some extent with a carefully designed scoring function.

Language models are better than word embeddings at understanding abstract relations, but have ample room for improvement (BERT is still worse even if it is tuned).

Language models are very sensitive to hyperparameter tuning in this task, and careful tuning leads to competitive results.

Conclusion

GPT3 result was dissapointing given the model size, but we show some of the smaller LMs (RoBERTa and GPT2) can solve analogies in a true zero-shot setting to some extent with a carefully designed scoring function.

Language models are better than word embeddings at understanding abstract relations, but have ample room for improvement (BERT is still worse even if it is tuned).

Language models are very sensitive to hyperparameter tuning in this task, and careful tuning leads to competitive results.

Assumption on Relational Knowledge of LMs

LMs do have relational knowledge, but it is not easy to exploit and utilize in downstream tasks with vanilla LMs.

Conclusion

GPT3 result was dissapointing given the model size, but we show some of the smaller LMs (RoBERTa and GPT2) can solve analogies in a true zero-shot setting to some extent with a carefully designed scoring function.

Language models are better than word embeddings at understanding abstract relations, but have ample room for improvement (BERT is still worse even if it is tuned).

Language models are very sensitive to hyperparameter tuning in this tas
tuning leads to competitive results.

Assumption on Relational Knowledge of LMs

LMs do have relational knowledge, but it is not easy to exploit and utilize in downstream tasks with vanilla LMs.

What if we can fine-tune LM to explicitly extract relational knowledge?🤔

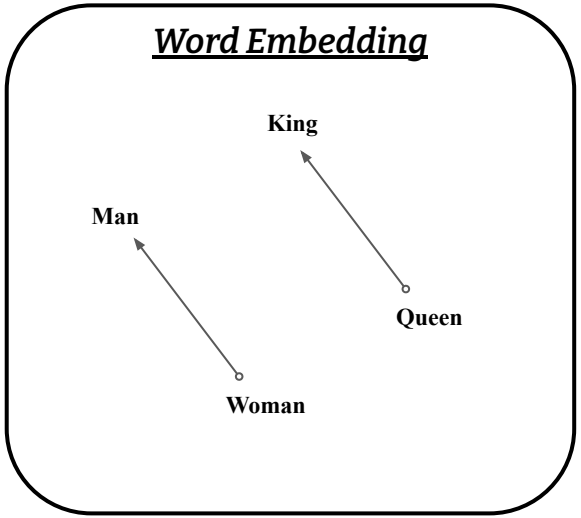
Distilling Relation Embeddings from Pre-trained Language Models

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**

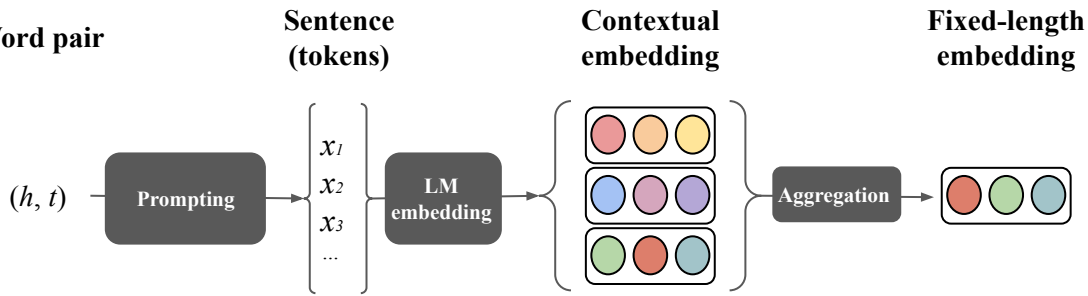
Relation Embedding with LM (RelBERT)

Prompt Generation

- Custom Template, AutoPrompt [Shin 2020], P-tuning [Liu 2021]

Embedding Aggregation

- Averaging Pooling / Mask Embedding



Relation Embedding with LM (R

Prompt Generation

- Custom Template, ~~AutoPrompt~~ [Shin 202

1. Today, I finally discovered the relation between `h` and `t` : `h` is the <mask> of photographer

2. Today, I finally discovered the relation between `h` and `t` : `t` is `h`'s <mask>

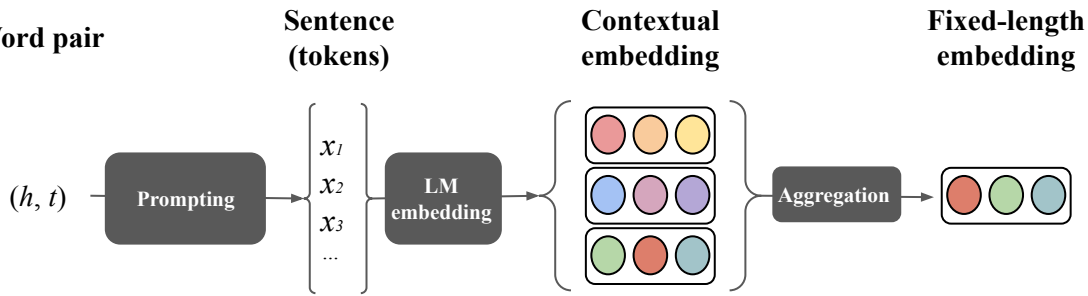
3. Today, I finally discovered the relation between `h` and `t` : <mask>

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

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

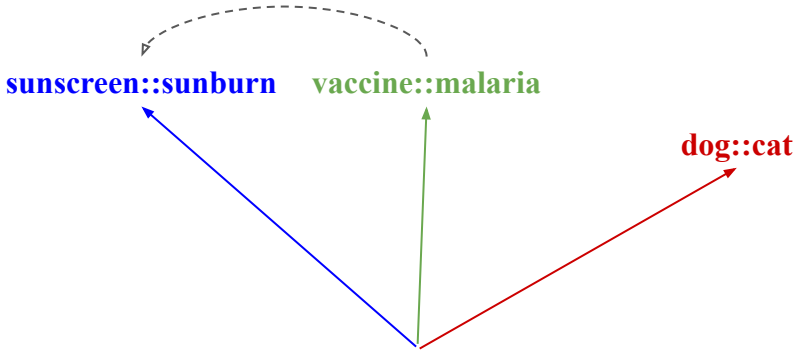
Embedding Aggregation

- Averaging Pooling / Mask Embedding



Fine-tuning

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.



Triplet and Classification Loss [Reimers 2019]

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

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

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

where W is a learnable weight.

Dataset

We create the dataset from Relation Similarity Dataset

(taxonomy).

Training: 1,580 (positive pairs) 70,207 (negative pairs)

Validation: 1,778 (positive pairs) 78,820 (negative pairs)

10 parent relations

10 child relations

The resulting taxonomy consists of 10 families or classes of relations, each with a number of specific relations as members. A brief characterization of each class follows:

1. CLASS INCLUSION: one word names a class that includes the entity named by the other word.

2. PART-WHOLE: one word names a part of the entity named by the other word, or something that is characteristically not a part.

3. SIMILAR: one word represents a different degree or form of the object, action, or quality represented by the other word.

4. CONTRAST: one word names an opposite or incompatible of the other word.

5. ATTRIBUTE: one word names a characteristic quality, property, or action of the entity named by the other word.

6. NONATTRIBUTE: one word names a quality, property, or action that is characteristically not an attribute of the entity named by the other word.

7. CASE RELATION: one word names an action that the entity named by the other word is usually involved in, or both words name entities that are normally involved in the same action in different ways, e.g., as agent, object, recipient, or instrument of the action.

8. CAUSE-PURPOSE: one word represents the cause, purpose, or goal of the entity named by the other word, or the purpose or goal of using the entity named by the other word.

9. SPACE-TIME: one word names a thing or action that is associated with a particular location or time named by the other word.

10. REPRESENTATION: one word names something that is an expression or representation of, or a plan or design for, or provides information about, the entity named by the other word.

Dataset

We create the dataset from Relation Similarity Dataset

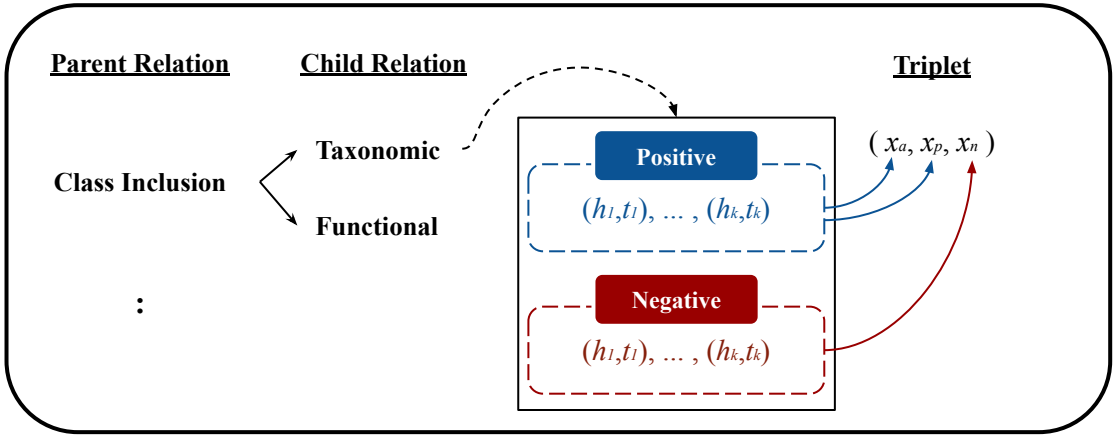
(taxonomy).

Training: 1,580 (positive pairs) 70,207 (negative pairs)

Validation: 1,778 (positive pairs) 78,820 (negative pairs)

10 parent relations

10 child relations



EXPERIMENTS

Analogy Question

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

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

Setup

- Cosine similarity in between embeddings (same as word embedding).
- No additional training or validation.
- Accuracy as the metric.
- RoBERTa large as an anchor language model for RelBERT.

Analogy Question: Results

SotA in 4 / 5 datasets 🎉

Even better than methods tuned on validation 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

Lexical Relation 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.

Classification: Results

SotA in 4 / 5 datasets in macro F1 score 🎉

SotA in 3 / 5 datasets in micro F1 score 🎉

	Model	BLESS		CogALexV		EVALution		K&H+N		ROOT09	
		macro	micro	macro	micro	macro	micro	macro	micro	macro	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

Memorization

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

Experiment: Train RelBERT without hypernym.

Result: No significant decrease in hypernym prediction.

	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)

Relation Type Breakdown

Break-down of BATS results per relation type.

- High accuracy in morphological analogies
- Poor accuracy in encyclopedic analogies

Note that the relation similarity dataset does not contain morphological relation but rather semantic relation (taxonomy).

	Relation	FastText	Manual
Encyclopedic	UK city:county	33.3	28.9
	animal:shelter	44.4	88.9
	animal:sound	80.0	86.7
	animal:young	53.3	62.2
	country:capital	82.2	37.8
	country:language	93.3	51.1
	male:female	88.9	60.0
	name:nationality	60.0	73.3
	name:occupation	86.7	75.6
	things:color	88.9	97.8
Lexical	antonyms:binary	26.7	64.4
	antonyms:gradable	44.4	88.9
	hypernyms:animals	44.4	91.1
	hypernyms:misc	42.2	71.1
	hyponyms:misc	31.1	55.6
	meronyms:member	44.4	68.9
	meronyms:part	31.1	77.8
	meronyms:substance	26.7	75.6
	synonyms:exact	17.8	80.0
	synonyms:intensity	28.9	77.8
Morphological	adj+ly	95.6	84.4
	adj+ness	100.0	97.8
	adj:comparative	100.0	97.8
	adj:superlative	97.8	100.0
	noun+less	77.8	100.0
	over+adj	75.6	84.4
	un+adj	60.0	97.8
	verb 3pSg:v+ed	100.0	75.6
	verb inf:3pSg	100.0	93.3
	verb inf:v+ed	100.0	91.1
	verb inf:v+ing	100.0	97.8
	verb v+ing:3pSg	97.8	82.2
	verb v+ing:v+ed	97.8	86.7
	verb+able	97.8	93.3
	verb+er	95.6	100.0
	verb+ment	95.6	77.8
	verb+tion	84.4	77.8
	noun:plural	78.7	87.6
	re+verb	75.6	62.2

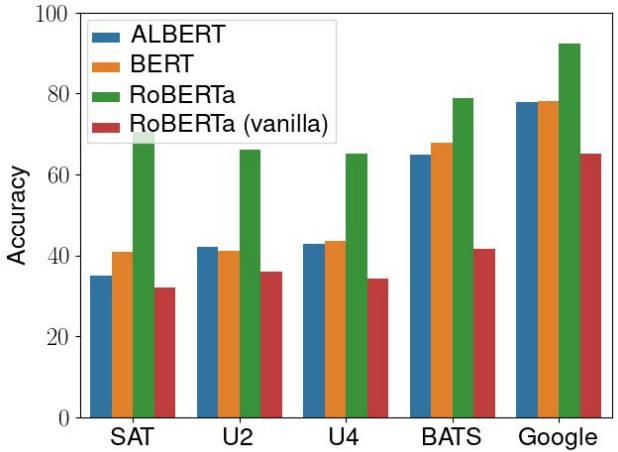
LM Variation

Train RelBERT on BERT, ALBERT in addition to RoBERTa.

→ RoBERTa is the best.

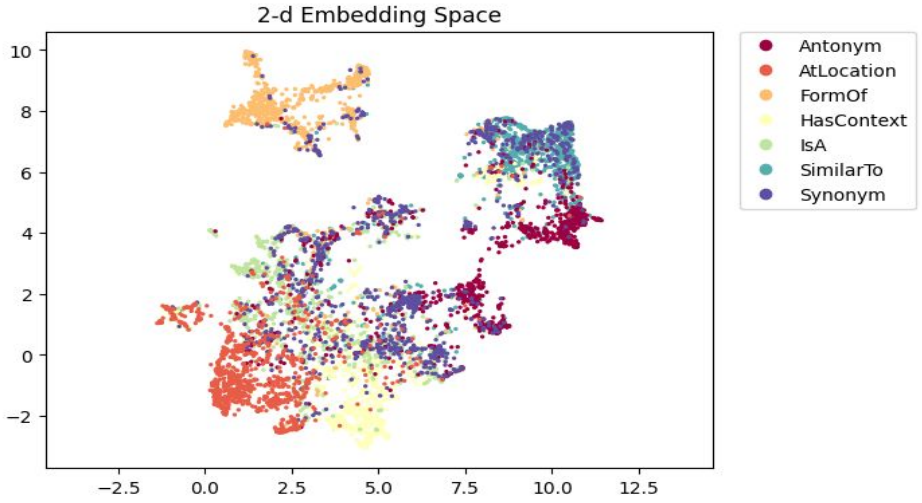
Vanilla RoBERTa (no fine-tuning).

→ Fine-tuning (distillation) is necessary.

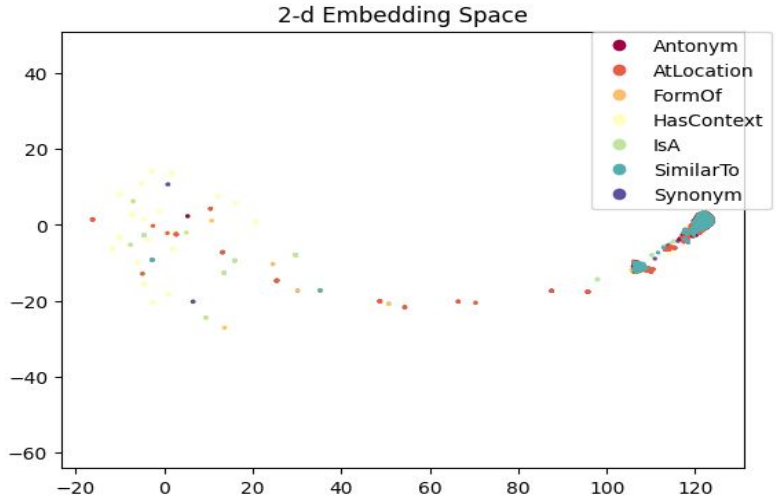


Latent Space Visualization

2-D embeddings of ConceptNet (with UMAP dim-reduction).



RelBERT



word2vec

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 embeddings outperform existing baselines, establishing SotA in analogy and relation classification.

- Model: https://huggingface.co/relbert/relbert-roberta-large
- Demo: https://huggingface.co/spaces/relbert/Analogy
- GitHub: https://github.com/asahi417/relbert

Thank You!