A RelEntLess Benchmark for Modelling Graded Relations between Named Entities

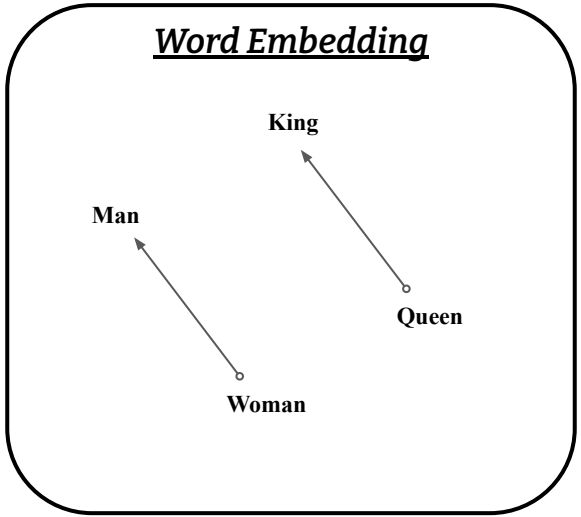
EACL 2024 Main Conference

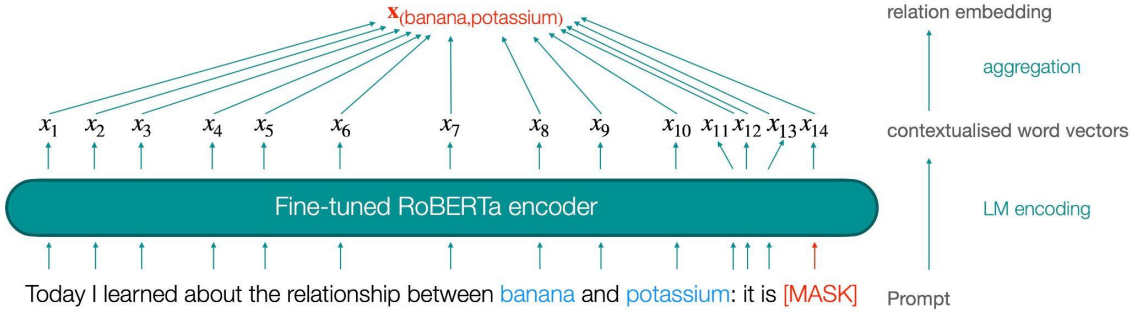
Asahi Ushio, Jose Camacho-Collados, Steven Schockaert
Cardiff NLP

Relational Knowledge

Capability to understand relationship between two words.

- Word Embedding [Mikolov (2013)]
- LMs (eg. GPT-3)
- RelBERT





Asahi Ushio, et al. "Distilling Relation Embeddings from Pretrained Language Models" (EMNLP 2021)

Word Analogy

Word analogy as a probing task of relational knowledge.

- ➢ Solvable without training.
- ➢ Different Levels
  - ○ Primary school to college
- ➢ Various Relation Types
  - ○ Named entity, common noun

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

Research Question

Word analogies are **discriminative**.

➢ (“Tokyo”, “Japan”) is *capital-of*, but (“U.K.”, “Japan”) is not.

Relations in real-world are often **graded**.

➢ (“Amazon”, “Google”) is more prototypical example of *competitor* than (“Netflix”, “Disney”).

Can LMs understand such graded relation?

Graded Relation Ranking

Relation Types	Examples (Ordered by Prototypicality)
competitor of	*[Dell, HP]* > *[Neoclassicism, Romanticism]* > *[Steve Jobs, Atlanta]*
friend of	*[Australia, New Zealand]* > *[The Beatles, Queen]* > *[KGB, CIA]*
influenced by	*[Plato, Socrates]* > *[Hip Hop, Jazz]* > *[Sauron, Shiba Inu]*
known for	*[Apple, iPhone]* > *[Apple, Apple Watch]* > *[Pixar, Novosibirsk]*
similar to	*[Coca-Cola, Peps]* > *[Christmas, Easter]* > *[Italy, Superman]*

New challenging tasks.

- **5** relation types.
- Pairs of **named entities**.
- **Rank** the pairs based on **prototypicality**.

Results

	competitor	friend	influenced	known	similar	average
Human	**75.9**	**78.0**	**70.5**	**82.0**	**80.2**	**80.2**
FastText	**25.0**	**10.0**	**7.0**	**24.0**	**20.0**	**17.0**
RelBERT	**64.0**	**20.0**	**20.0**	**44.0**	**53.0**	**40.0**
FlanT5	**74.0**	**56.0**	**44.0**	**70.0**	**66.0**	**62.0**
Flan-UL2	**79.0**	**51.0**	**47.0**	**67.0**	**57.0**	**60.0**
GPT3	**72.0**	**39.0**	**64.0**	**73.0**	**47.0**	**59.0**
GPT4	**62.5**	**55.8**	**35.9**	**60.8**	**69.3**	**56.9**

Analysis

- It scales with the model size (bigger models are often better).
- Choice of template matters.
- Few-shot improves most models (except Flan-UL2).
- Typical error involves
  - Biased by entity domain: “Rihanna” / “Stevie Wonder” for “influenced” (music domain)
  - New relationship: “OpenAI” / “Microsoft”
  - Surface similarity: “New York”/”York”

Thank you!