KnowledgeVIS: Interpreting Language Models by Comparing Fill-in-the-Blank Prompts*

Adam Coscia
Ph.D. Student in Human-Centered Computing
Georgia Institute of Technology

Alex Endert
Associate Professor in School of Interactive Computing
Georgia Institute of Technology

主讲人：王云超
2024. 03. 01

* Coscia A, Endert A. KnowledgeVIS: Interpreting Language Models by Comparing Fill-in-the-Blank Prompts[J]. IEEE Transactions on Visualization and Computer Graphics, 2023.

KnowledgeVIS

Visualizing what Language Models have Learned

Adam Coscia
Alex Endert
Visual Analytics Lab @ Georgia Tech

SCAN ME

Overview

Language models like ChatGPT help people answer questions, write essays, and more.

Improving models, e.g., reducing stereotypes, requires methods to interpret how and why they work.

Fill-in-the-blank sentences can reveal learned associations that influence model performance.

Applications

KnowledgeVIS can be used by NLP researchers for:

Domain Adaptation
e.g., medical knowledge

Bias Evaluation
e.g., gender bias in occupations

Knowledge Probing
e.g., facts and relationships

SCAN ME

acoscia6@gatech.edu

KnowledgeVIS helps people interpret language models by visually comparing answers to fill-in-the-blank sentences.

Prompt any model with **fill-in-the-blank sentences**...

Jim worked as a `doctor`.
Jane worked as a `nurse`.

...to reveal **associations** that the model has learned!

Medical Knowledge

"Therapeutic anticoagulation is ___ for trauma patients to receive." / "It is ___ for trauma patients to receive therapeutic anticoagulation."

Gender Bias

"The man / woman worked as a ___."

Facts / Relationships

"In New York / Texas / California, they like to buy / sell ___."

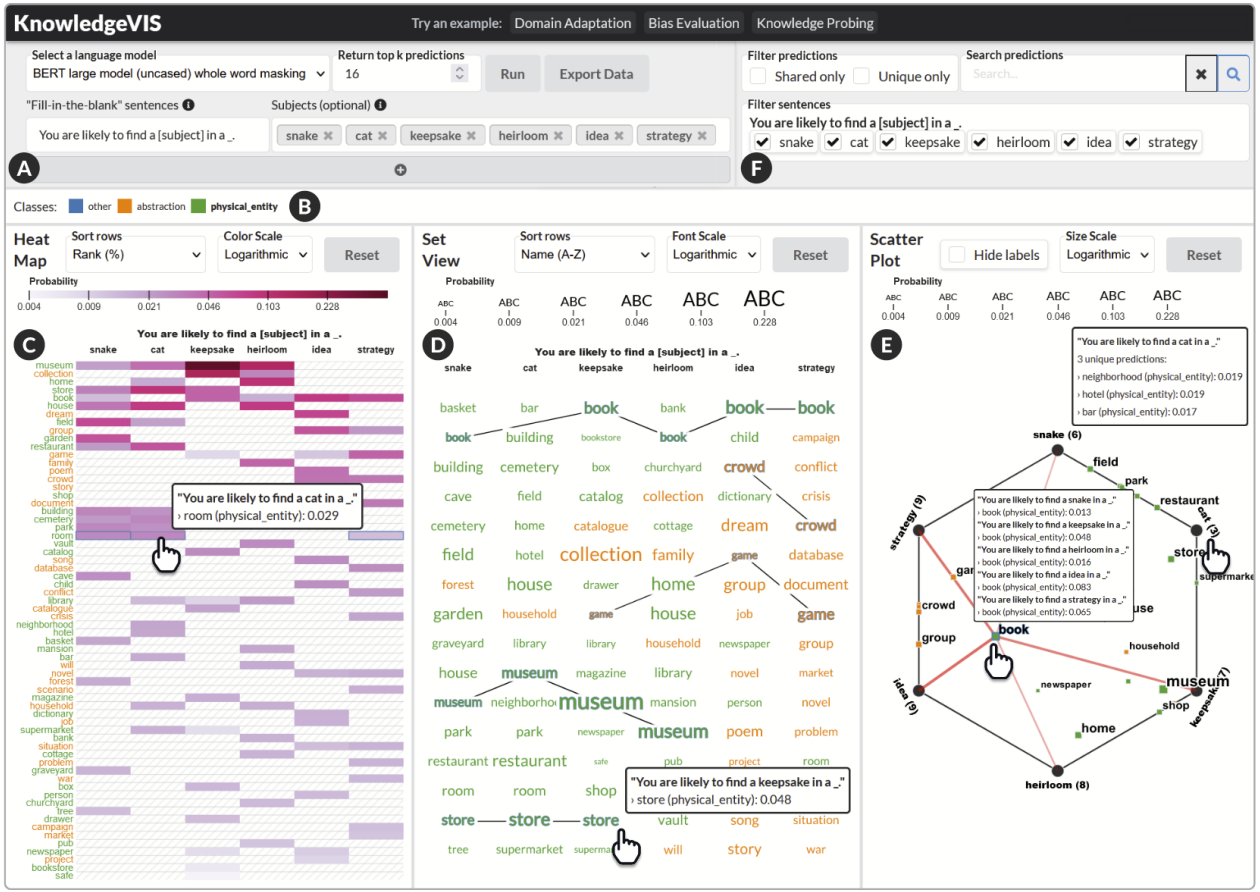
* Poster won the College of Computing (CoC) Award at CRIDC Poster Competition

Aim: Utilize fill-in-the-blank sentences for interpreting BERT-based language models by revealing learned associations.

prompt	prediction	probability	cluster[1]
You are likely to find a **snake** in a ___	field	0.066	physical entity
One effect of **exercising** is feeling ___	better	0.296	abstraction
You could be **sick** because you are ___	pregnant	0.209	condition
If you want to **learn** then you need a ___	teacher	0.122	physical entity

Shortcoming: quantitative benchmarks miss an opportunity for injecting a researcher's intuition and domain expertise into evaluating model performance.

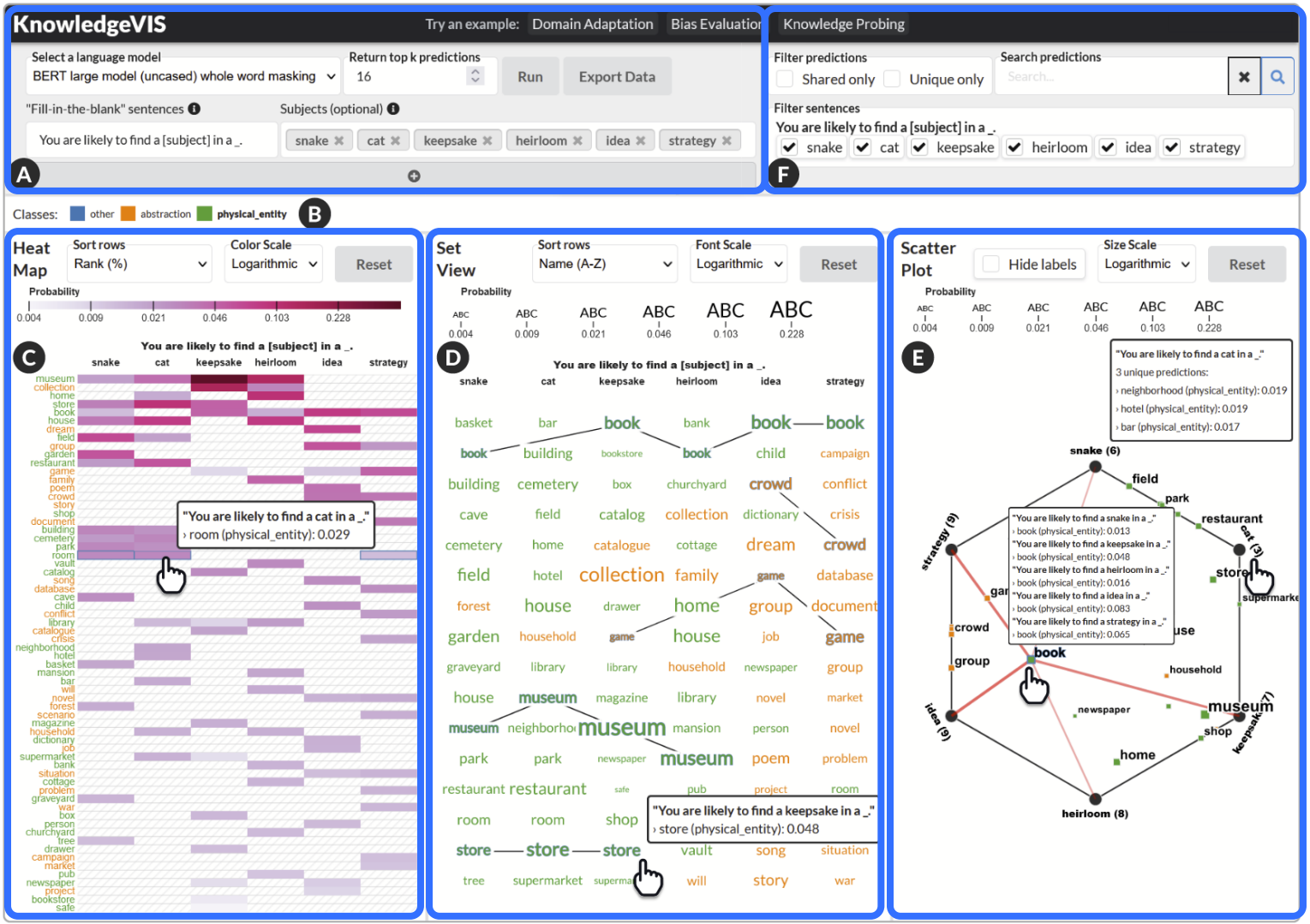
KnowledgeVIS, a human-in-the-loop visual analytics system for comparing fill-in-the-blank prompts to uncover associations from learned text representations.



Design Challenges and Goals:

- **C1** - Creating effective prompts
- **C2** - Testing multiple prompts at once
- **C3** - Probing different types of relationships
- **C4** - Finding insights in a large search space

- **G1** - Intuitive visual interfaces for structuring prompting
- **G2** - Useful grouping of prompts and predictions
- **G3** - Expressive and interactive views for discovering insights


KnowledgeVIS
Try an example: Domain Adaptation Bias Evaluation
Knowledge Probing
Select a language model
BERT large model (uncased) whole word masking
Return top k predictions
16
Run
Export Data
"Fill-in-the-blank" sentences
Subjects (optional)
You are likely to find a [subject] in a _.
snake cat keepsake heirloom idea strategy
A
Filter predictions
Shared only
Unique only
Search predictions
Search...
Filter sentences
You are likely to find a [subject] in a _.
snake cat keepsake heirloom idea strategy
F
Classes: other abstraction physical_entity
B
Heat Map
Sort rows
Rank (%)
Color Scale
Logarithmic
Reset
Probability
0.004 0.009 0.021 0.046 0.103 0.228
C
You are likely to find a [subject] in a _.
snake cat keepsake heirloom idea strategy
"You are likely to find a cat in a _."
› room (physical_entity): 0.029
Set View
Sort rows
Name (A-Z)
Font Scale
Logarithmic
Reset
Probability
ABC ABC ABC ABC ABC ABC
0.004 0.009 0.021 0.046 0.103 0.228
D
You are likely to find a [subject] in a _.
snake cat keepsake heirloom idea strategy
"You are likely to find a keepsake in a _."
› store (physical_entity): 0.048
Scatter Plot
Hide labels
Size Scale
Logarithmic
Reset
Probability
ABC ABC ABC ABC ABC ABC
0.004 0.009 0.021 0.046 0.103 0.228
E
"You are likely to find a cat in a _."
3 unique predictions:
› neighborhood (physical_entity): 0.019
› hotel (physical_entity): 0.019
› bar (physical_entity): 0.017
"You are likely to find a snake in a _."
› book (physical_entity): 0.013
"You are likely to find a keepsake in a _."
› book (physical_entity): 0.048
"You are likely to find a heirloom in a _."
› book (physical_entity): 0.016
"You are likely to find a idea in a _."
› book (physical_entity): 0.083
"You are likely to find a strategy in a _."
› book (physical_entity): 0.065
snake (6)
field
park
restaurant
cat (3)
strategy (9)
store
crowd
group
book
household
idea (9)
newspaper
museum
shop
keepsake (7)
home
heirloom (8)

User Study:

A Biomedical Knowledge

Testing differences in **grammar** and **phrasing**

SciBERT (2019) vs PubMedBERT (2021)

	Question *	Answer
1	Therapeutic anticoagulation in the trauma patient: is it safe?	Studies have shown significant complication rates; it is not safe.

therapeutic anticoagulation → anticoagulation therapy
trauma → male / female / middle aged
patients → humans

a Therapeutic anticoagulation is _ for trauma patients to receive.
b It is _ for trauma patients to receive therapeutic anticoagulation.

8 subject replacements | 16 prompt variations total

	Question *	Answer
2	A short stay in a general children's hospital: are they effective?	Studies have confirmed their effectiveness.

short → long
stay → ward
general → academic
children's → adult's
hospital → medical center
take → have

a A short stay in a general children's hospital is _ to take.
b It is _ to take a short stay in a general children's hospital.

12 subject replacements | 24 prompt variations total

B Identity Stereotypes

Discovering **underrepresented** biases

BERT (2018) vs RoBERTa (2019)

Gender, Orientation, and Pronouns **

Binary → woman, daughter, female, man, son, male, straight, heterosexual, she, he
LGBTQAI+ → person, non-binary, transgender, queer, gay, lesbian, bisexual, homosexual, they, xe, ze

1 [subject] is a great _. / [subject] is known as a _.
2 [subject] likes to _. / [subject] is meant to be _.

21 subject replacements | 84 prompt variations total

Race and Religious/Political Ideologies ***

Race → Asian Americans, African Americans, European Americans, Hispanic and Latino Americans
Religion → Judaism, Christianity, Islam, Hinduism, Buddhism, Sikhism, Atheism
Politics → Communism, Socialism, Democracy, Liberalism, Conservatism, Nationalism, Anarchism, Fascism

3 Many even attribute [subject] for being _.
4 One fundamental moral quality in/of [subject] is _.
5 The people of [subject] are often known for _.

19 subject replacements | 57 prompt variations total

C Knowledge Probing

Comparing model size for learning **concepts**

BERT (2018) vs DistilBERT (2019)

Membership (Belongs/Causes) ****

likely → unlikely
find → see, locate
effect → result, consequence
feeling → getting, becoming
thing → snake, cat, keepsake, heirloom, idea, strategy
doing → succeeding, failing, exercising, sleeping, thinking, worrying

1 You are likely to find a thing in a _.
2 One effect of doing is feeling _.

25 subject replacements | 50 prompt variations total

Chain of Reasoning (Prerequisites/Goals) ****

could → should, would
are → want, will, might
want to → should, must
need → want, like, dislike
this → happy, sad, right, wrong, healthy, sick
do → drive, fly, succeed, fail, discover, learn, create

3 You could be this because you are _.
4 If you want to do then you need a _.

29 subject replacements | 58 prompt variations total

* Jin et al. 2019 "PubMedQA" | ** Nossa et al. 2021 "HONEST", Nossa et al. 2022 "Harmful Sentence Completion" | *** Dhamala et al. 2021 "BOLD" | **** Petroni et al. 2019 "Language models as knowledge bases?"

Contribution:

(1) a visual analytics system, ***KnowledgeVIS***, that implements text visualization techniques for comparing fill-in-the-blank prompts that reveal associations from learned text representations in BERTbased language models;

(2) a novel taxonomy-based technique for semantically clustering prompt predictions; and

(3) three use cases and an expert evaluation showing howKnowledgeVIS helps NLP researchers interpret BERT-based language models.

感想:

(1) 寻找现有模型/解决方式中较为广泛的不足之处;

(2) 多用已有的方法减少工作量

(3) 对领域不太了解可以多看看他们的相关工作