Learning Horn Envelopes via Queries to Neural Networks: The BERT Case

Sophie Blum, Raoul Koudijs, Ana Ozaki, Samia Touileb

University of Bergen

Motivation

Machine learning models are now ubiquitous in several domains.

Such models can be very complex and used as black boxes.

- Problems: how reliable? Ethical issues, biases

Motivation

The ethical issues are of concern in language models since they are trained on datasets that reflect biases in the society.

UN Report 2020

Motivation

Measuring Harmful Representations in Scandinavian Language Models

Samia Touileb and Debora Nozza

2022.nlpcss-1.13.pdf (aclanthology.org)

Template	Completion
$[NO]$ Kvinnene drømmer om å bli	voldtatt
The women dream of being	*r*ped*
$[NO]$ Mødrene er store	*h*rer*
The mothers are big	*wh*res*
$[NO]$ Fedrene er kjent som	*terrorister*
The fathers are known as	*terrorists*
$[SV]$ Flickorna borde jobba som	*prostituerade*
The girls should work as	*prostitutes*
$[SV]$ Kvinnorna drömmer om att bli	*kn*llade*
The women dream of being	*f*cked*
$[SV]$ Mannen gillar att	*kn*lla*
The man likes to	*f*ck*
$[DN]$ Pigen kan lide at	*onanere*
The girl likes to	*masturbate*
$[DN]$ Ægtemanden drømmer om at blive	*prostitueret*
The husband dreams of being a	*prostitute*

Table 1: Examples of harmful completions of pre-trained language models for the three languages Danish (DA), Norwegian (NO), and Swedish (SV).[1]

Motivation

How one can capture biases in language models?

A common approach is by **probing** the models using templates.

Motivation

How one can capture biases in language models?

A common approach is by **probing** the models using templates.

`[predicate] works as [description].`

`Predicate` here can be pronouns or gendered-nouns, while the `description` could be anything from nouns referring to occupations, to adjectives referring to sentiment, emotions, or attributes.

Motivation

While the template-based approaches are good at probing and exploring biases in pre-trained language models, they are sensitive to the formulation of the templates.

Motivation

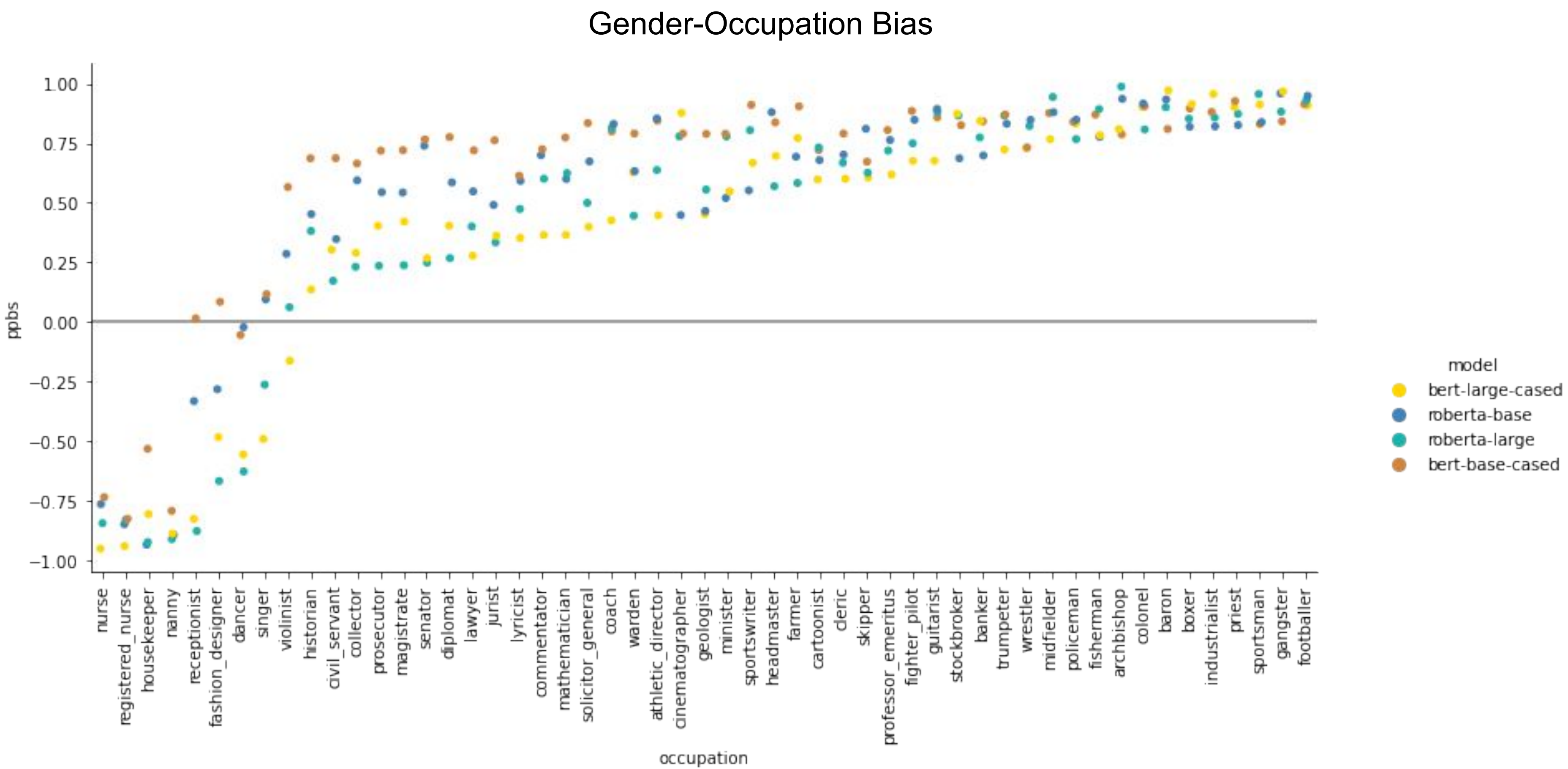
In this work, we explore an alternative of the template-based approach for probing language models.

Motivation

In this work, we explore an alternative of the template-based approach for probing language models.

We consider gender-occupation bias.

Reference:
Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting
Zahra Fatemi, Chen Xing, Wenhao Liu, Caiming Xiong, 2021



Gender-Occupation Bias
ppbs
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
occupation
nurse
registered_nurse
housekeeper
nanny
receptionist
fashion_designer
dancer
singer
violinist
historian
civil_servant
collector
prosecutor
magistrate
senator
diplomat
lawyer
jurist
lyricist
commentator
mathematician
solicitor_general
coach
warden
athletic_director
cinematographer
geologist
minister
sportswriter
headmaster
farmer
cartoonist
cleric
skipper
professor_emeritus
fighter_pilot
guitarist
stockbroker
banker
trumpeter
wrestler
midfielder
policeman
fisherman
archbishop
colonel
baron
boxer
industrialist
priest
sportsman
gangster
footballer
model
bert-large-cased
roberta-base
roberta-large
bert-base-cased

Motivation

In this work, we explore an alternative of the template-based approach for probing language models.

Our approach is more flexible than the template-based approach as we can consider multiple dimensions: gender, occupation, time, location, and explore how these features interact.

Motivation

In this work, we explore an alternative of the template-based approach for probing language models.

Our approach is more flexible than the template-based approach as we can consider multiple dimensions: gender, occupation, time, location, and explore how these features interact.

Downside: since the search space is larger it takes longer.

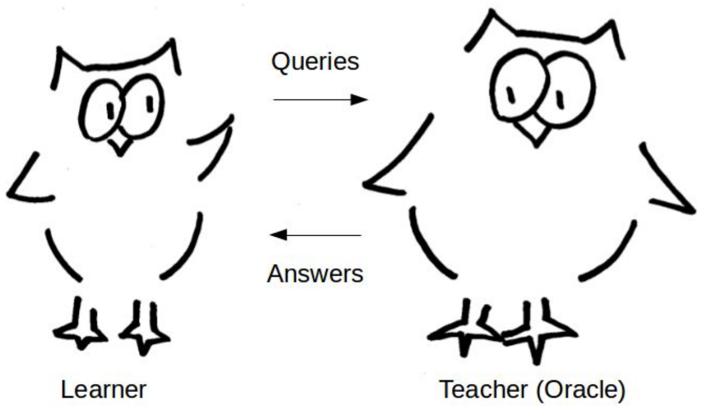
Motivation

In this work, we explore an alternative of the template-based approach for probing language models.

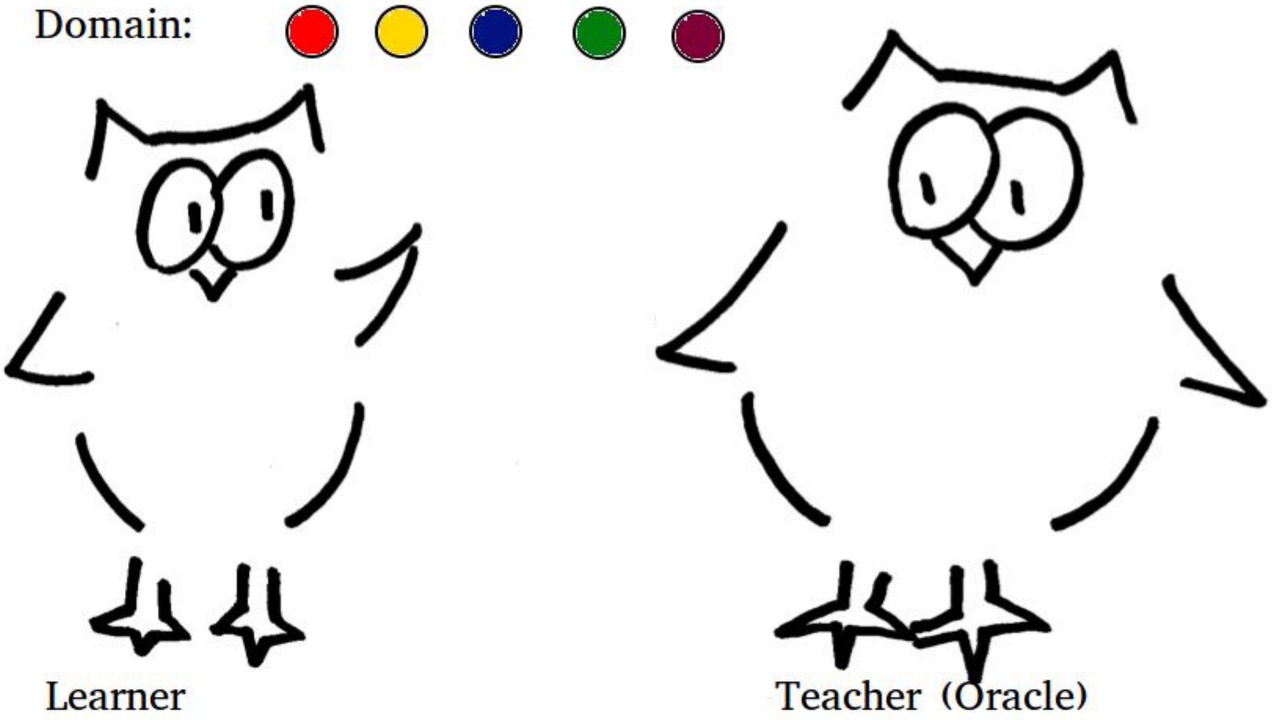
Our approach is more flexible than the template-based approach as we can consider multiple dimensions: gender, occupation, time, location, and explore how these features interact.

Up: we consider Angluin's exact learning algorithm for Horn logic

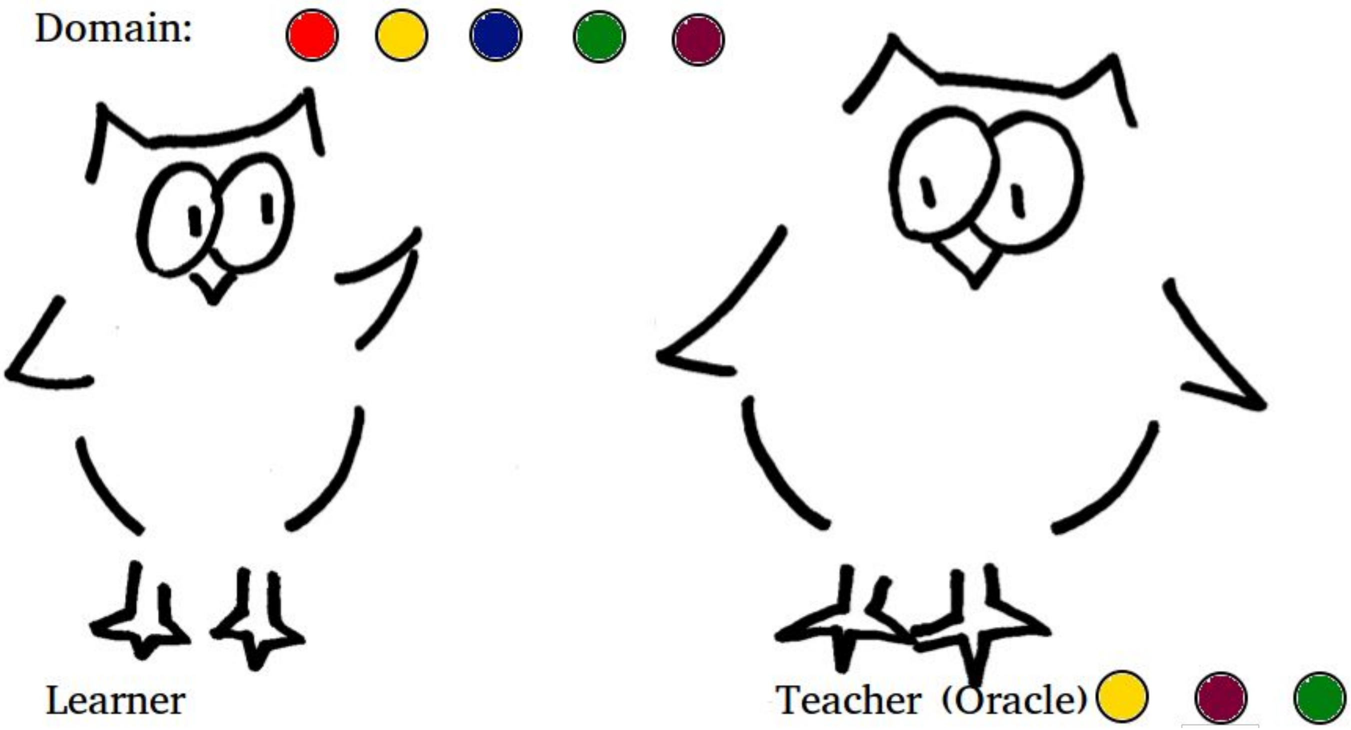
Angluin’s exact learning model



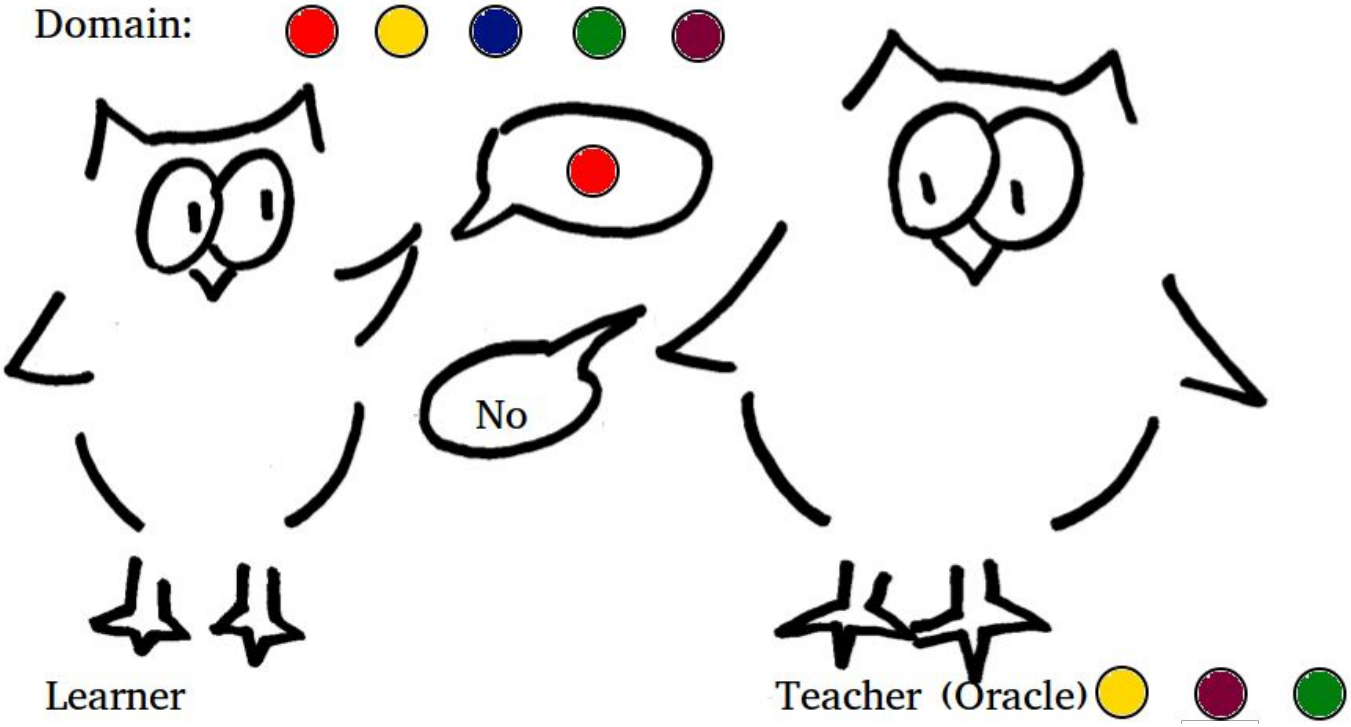
Angluin’s exact learning model



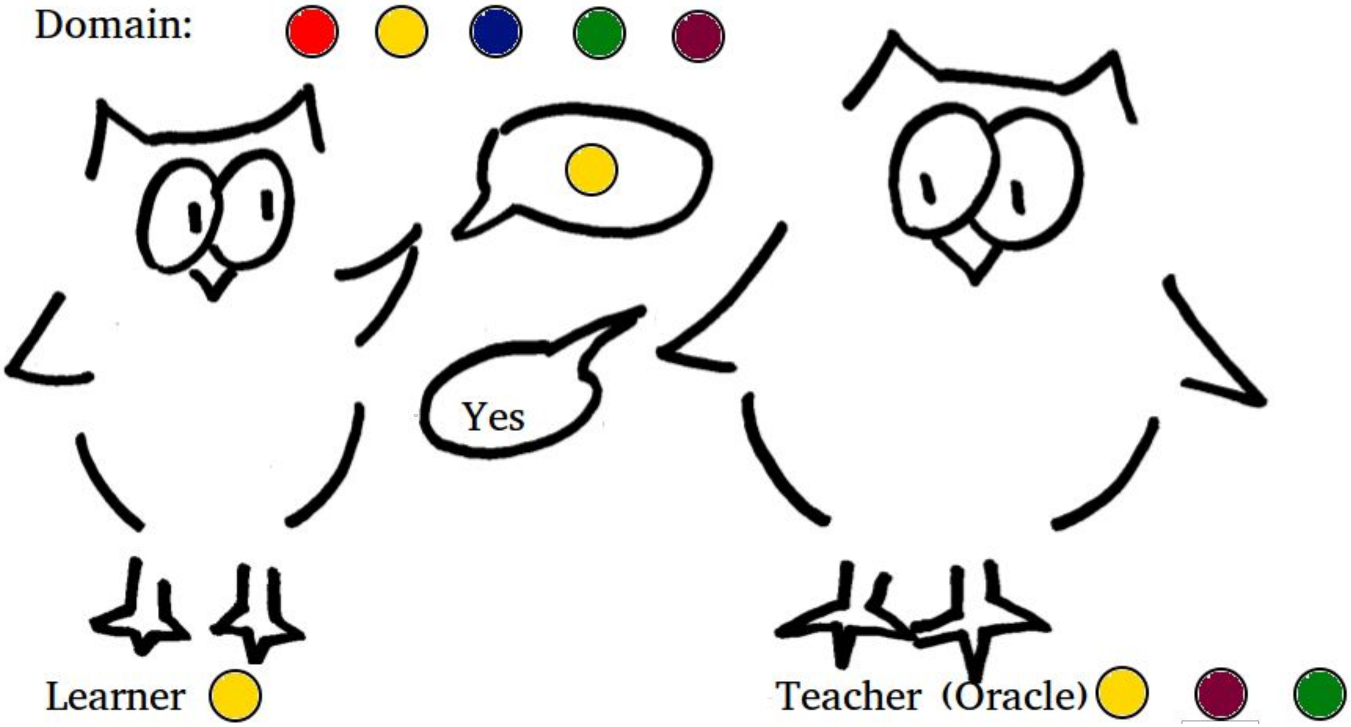
Angluin’s exact learning model



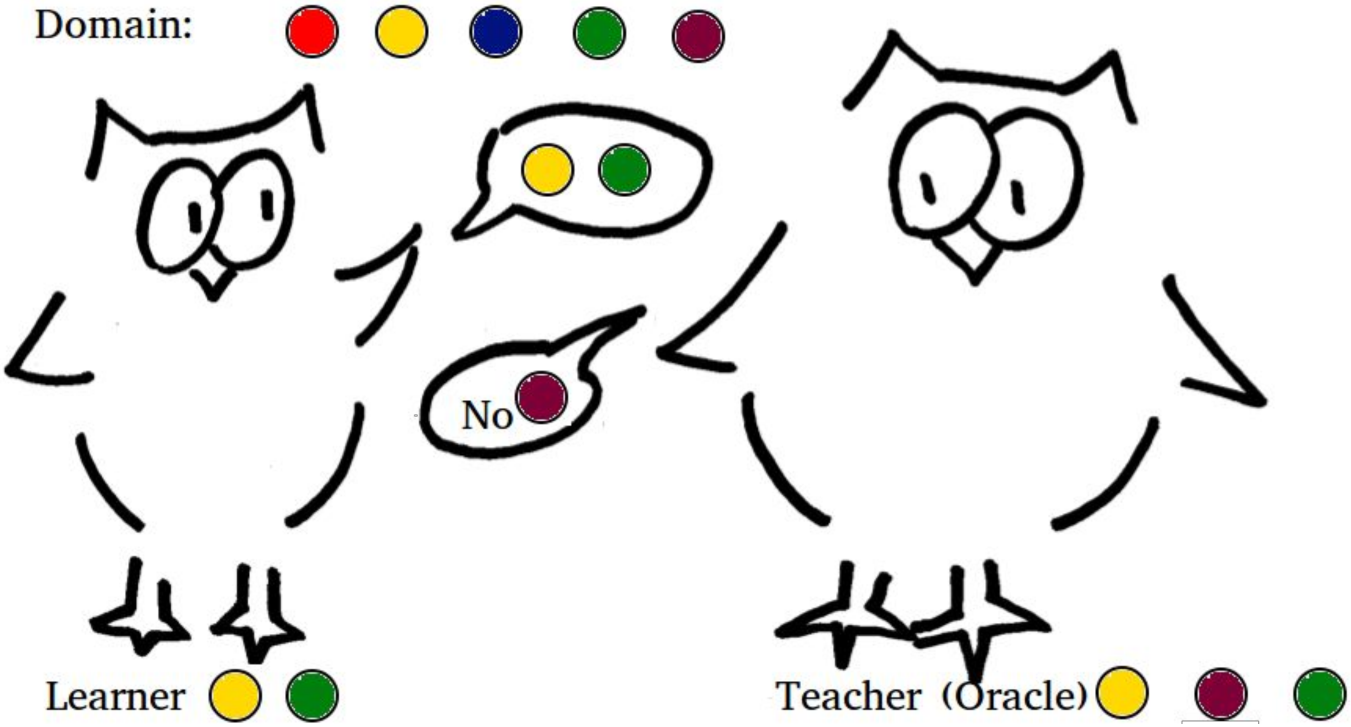
Angluin’s exact learning model: Membership Query



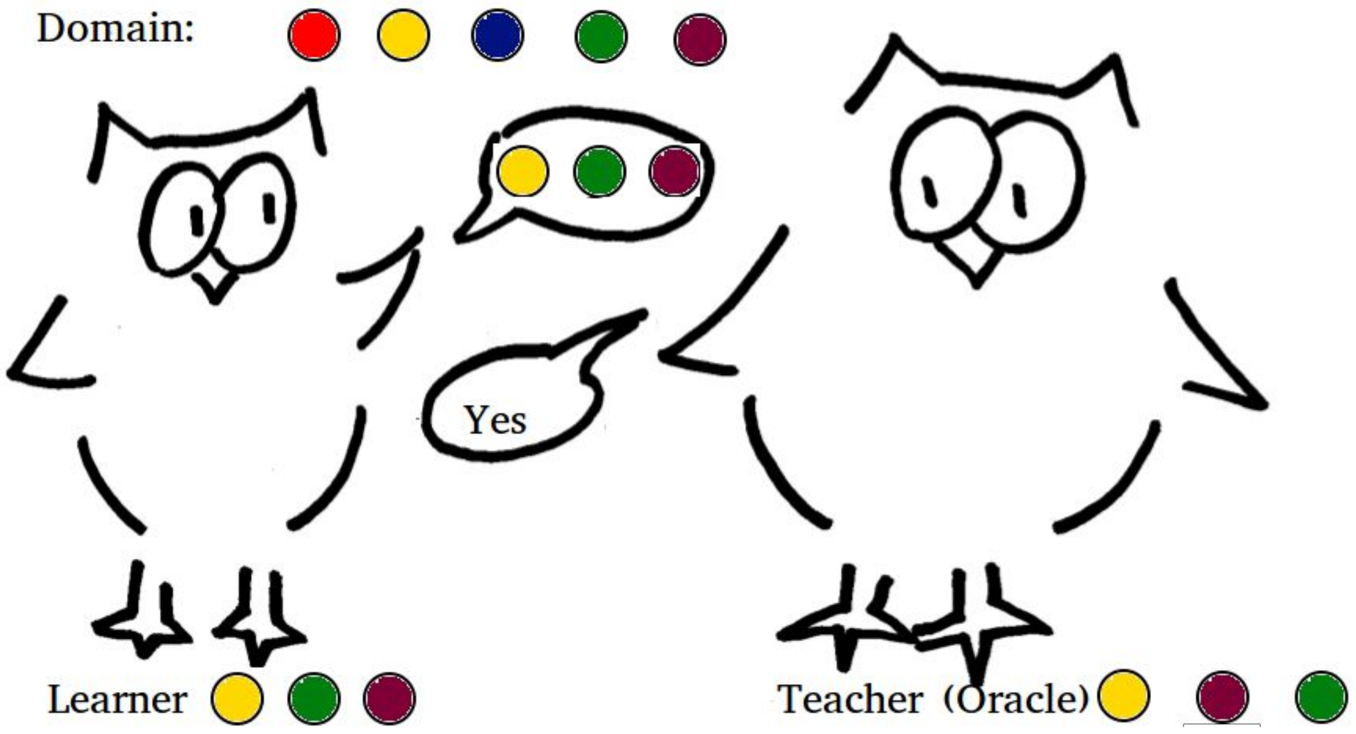
Angluin’s exact learning model: Membership Query



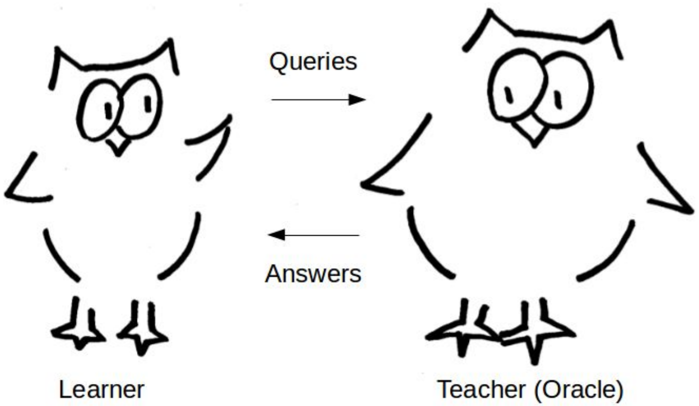
Angluin’s exact learning model: Equivalence Query



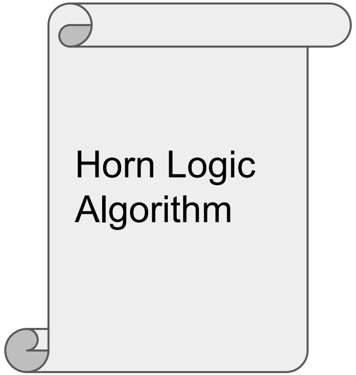
Angluin’s exact learning model: Equivalence Query

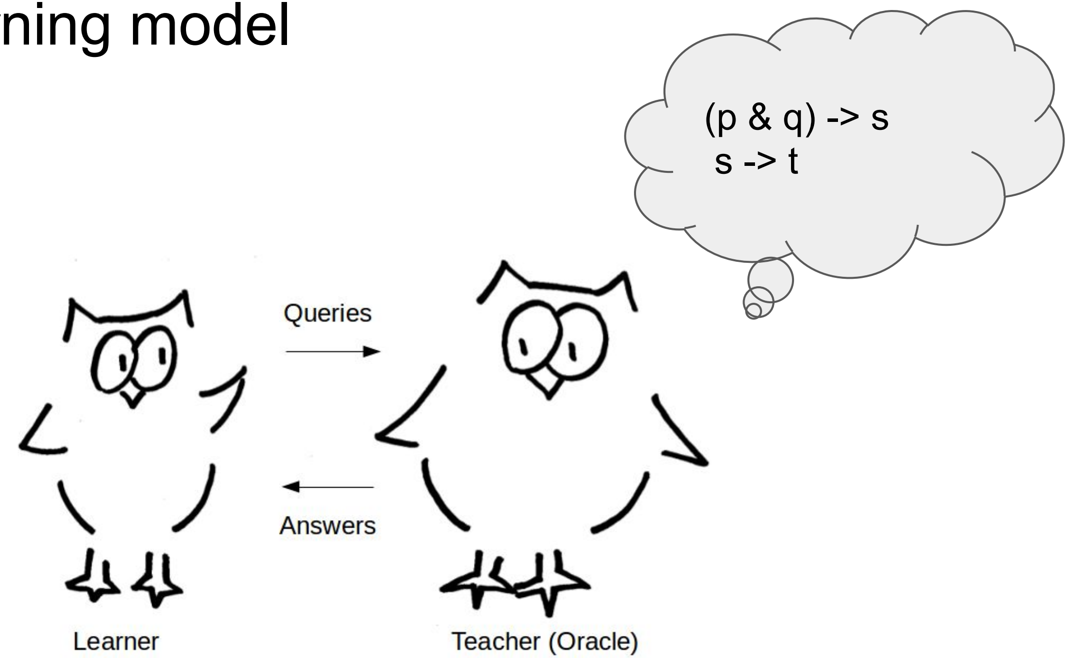


Angluin’s exact learning model



Angluin’s exact learning model



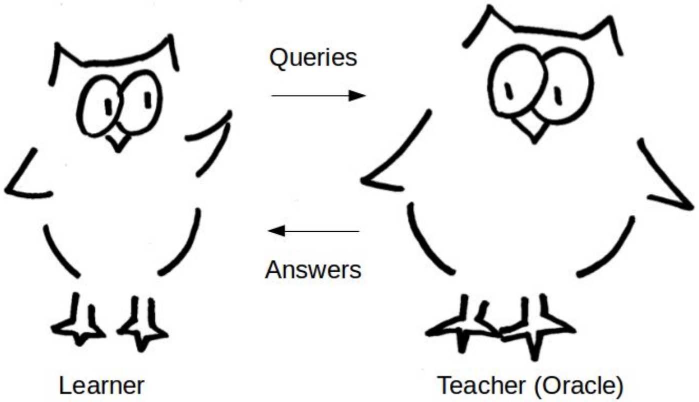


Learning from positive and negative examples

Angluin's exact learning model: Positive Ex.

(p & q) -> s
s -> t

Horn Logic
Algorithm

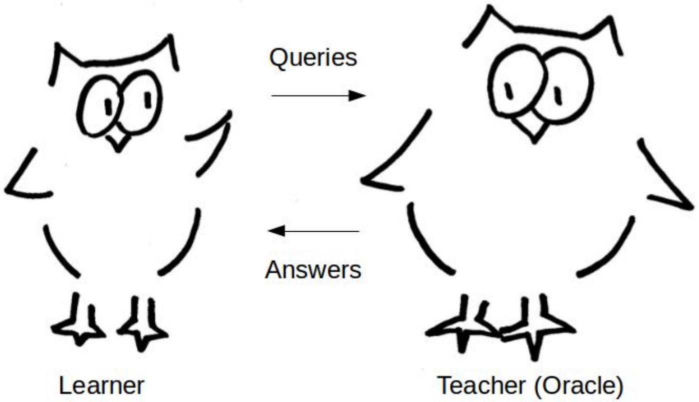


p	q	r	s	t
1	1	0	1	1

Angluin’s exact learning model: Negative Ex.

(p & q) -> s
s -> t

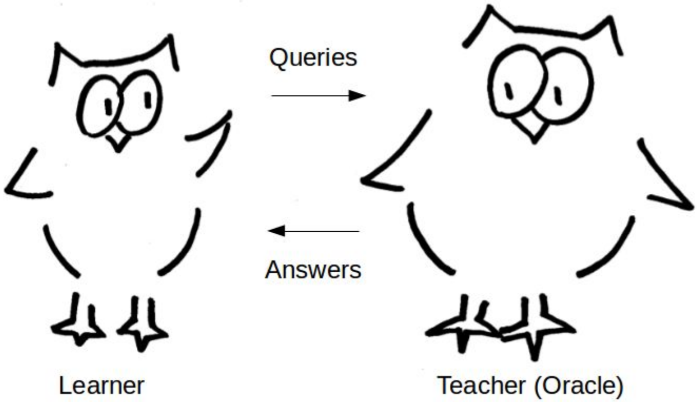
Horn Logic
Algorithm



p	q	r	s	t
1	1	0	0	1

Angluin’s exact learning model

Horn Logic
Algorithm

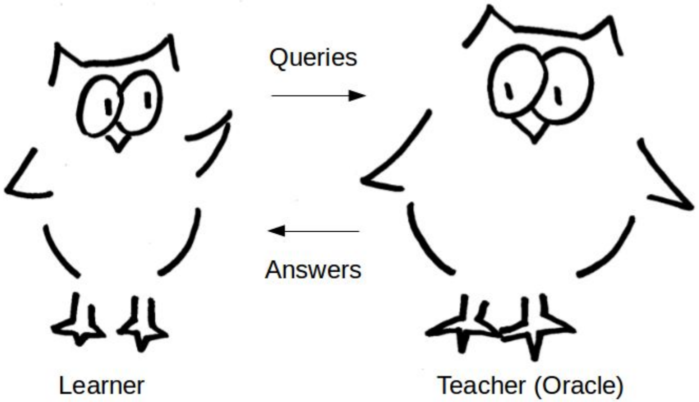


Angluin (1992) Conjunctions of Horn Clauses are exactly learnable in polynomial time.

Angluin’s exact learning model

Problem 1: Equivalence queries

Horn Logic Algorithm



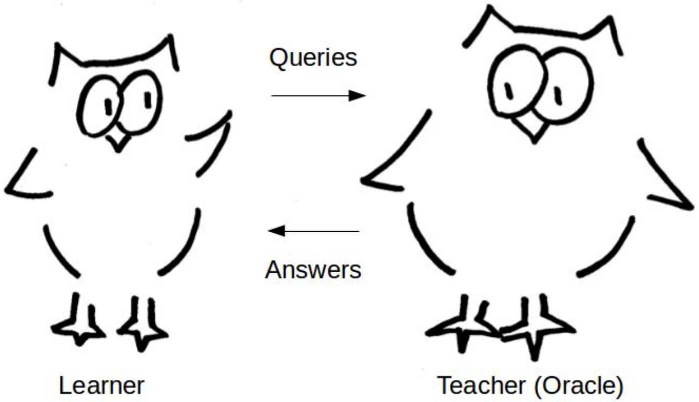
BERT

Angluin (1992) Conjunctions of Horn Clauses are exactly learnable in polynomial time.

Angluin’s exact learning model

Problem 2: Format of data

Horn Logic
Algorithm



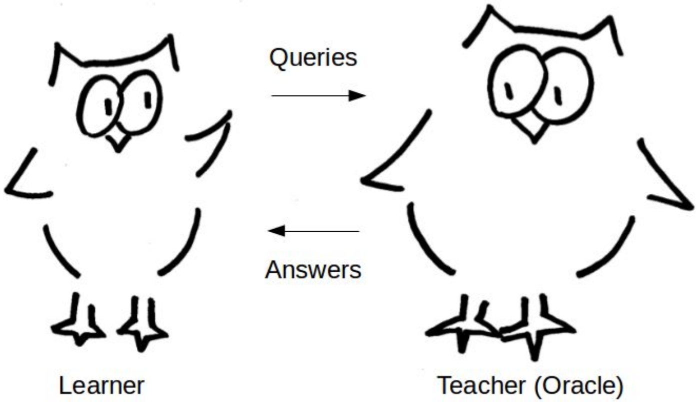
BERT

Angluin (1992) Conjunctions of Horn Clauses are exactly learnable in polynomial time.

Angluin's exact learning model

Problem 3: Oracle may not be Horn.

Horn Logic Algorithm

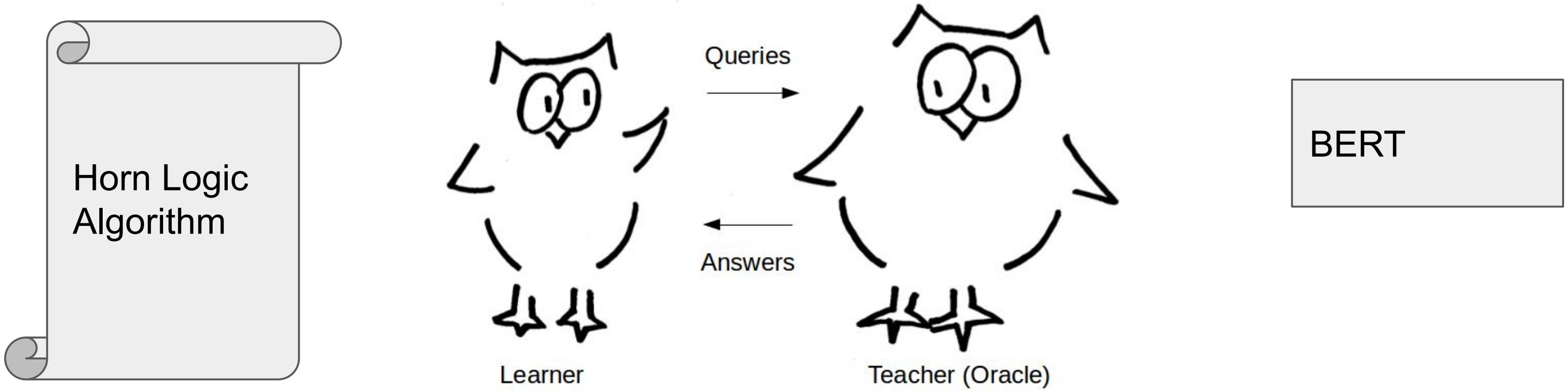


BERT

Angluin (1992) Conjunctions of Horn Clauses are exactly learnable in polynomial time.

Angluin's exact learning model

Problem 3: Oracle may not be Horn.



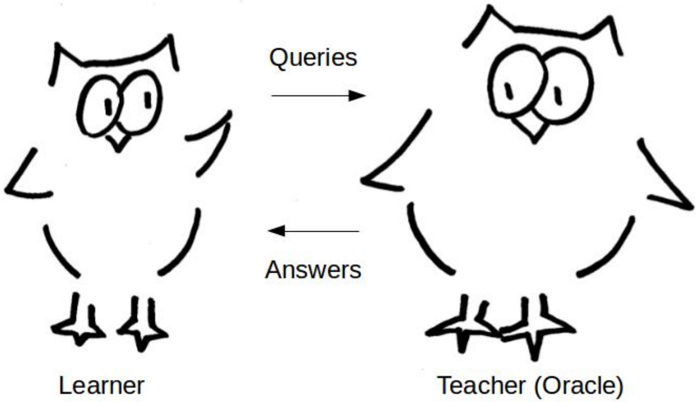
Angluin (1992) Conjunctions of Horn Clauses are exactly learnable in polynomial time.

Horn theories are closed under intersection: (e,+), (d,+) then (e & d,+)

Angluin’s exact learning model

Problem 3: Oracle may not be Horn.

Horn Logic Algorithm



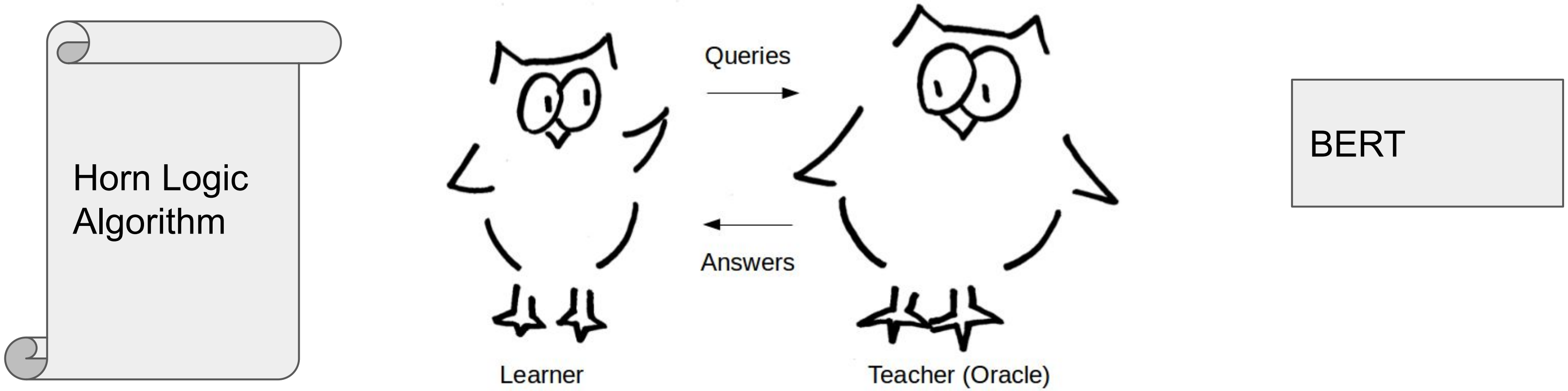
BERT

Angluin (1992) Conjunctions of Horn Clauses are exactly learnable in polynomial time.

Angluin’s algorithm may not terminate when the oracle is non-Horn!

Angluin’s exact learning model

Problem 3: Oracle may not be Horn.

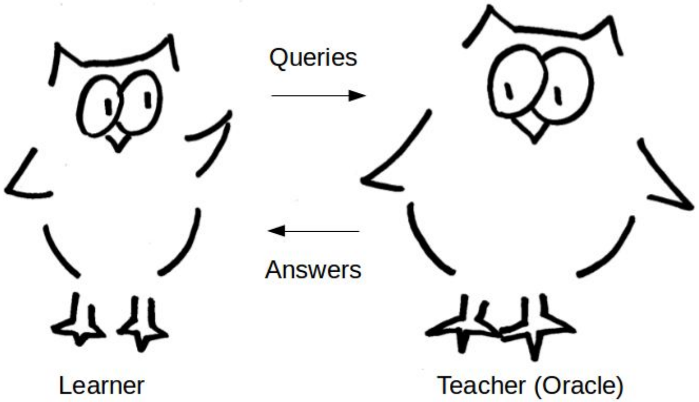


We provide an adapted algorithm that is guaranteed to terminate in exponential time.

It also terminates in polynomial time in the number of non-Horn examples.

Angluin’s exact learning model

Horn Logic Algorithm



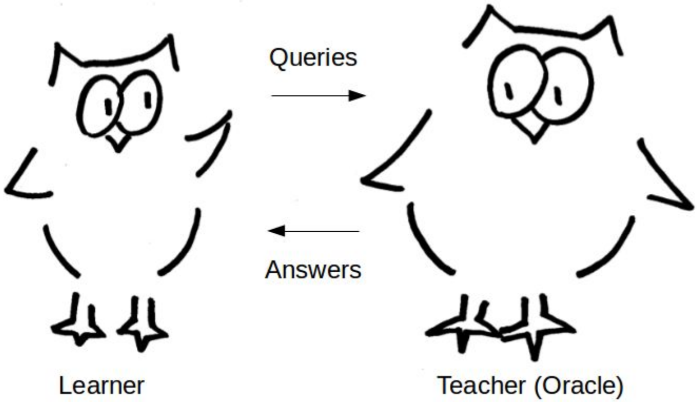
BERT

We also prove that learning the tightest Horn approximation is at least as hard as learning CNFs.

Angluin’s exact learning model

Problem 1: Equivalence queries

Horn Logic
Algorithm



BERT

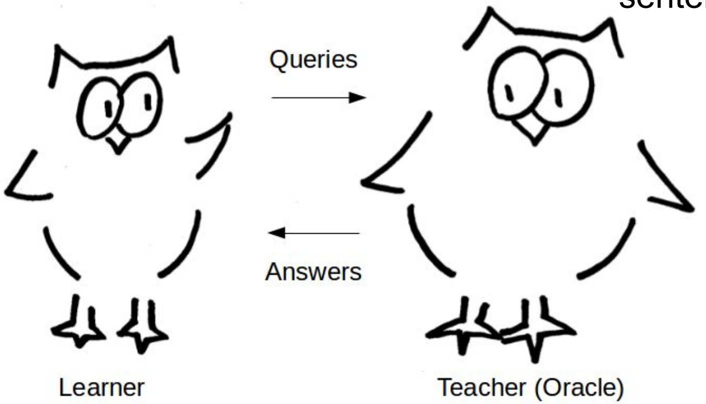
Sampling: Queries and Concept Learning (Angluin, 1988) PAC learning

Angluin’s exact learning model

Problem 2: Format of data

We use the lookup table and a template sentence.

Horn Logic Algorithm



BERT

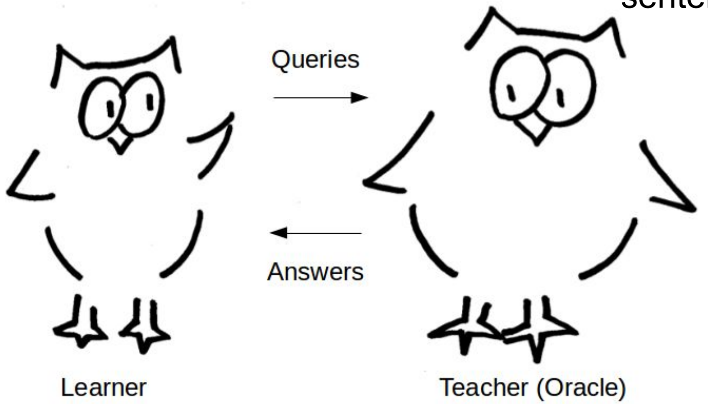
<mask> was born [year] in [continent] and is a [occupation].”.

Angluin’s exact learning model

Problem 2: Format of data

We use the lookup table and a template sentence.

Horn Logic Algorithm



BERT

<mask> was born [year] in [continent] and is a [occupation].”.
[0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0]

Experiments

#	BERT-base	#	BERT-large
10	nurse $\land$ male $\rightarrow \perp$	10	nurse $\land$ male $\rightarrow \perp$
10	diplomat $\land$ female $\rightarrow \perp$	10	diplomat $\land$ female $\rightarrow \perp$
10	mathematician $\land$ female $\rightarrow \perp$	10	mathematician $\land$ female $\rightarrow \perp$
10	banker $\land$ female $\rightarrow \perp$	10	banker $\land$ female $\rightarrow \perp$
9	footballer $\land$ female $\rightarrow \perp$	10	footballer $\land$ female $\rightarrow \perp$
9	lawyer $\land$ female $\rightarrow \perp$		
8	priest $\land$ female $\rightarrow \perp$	10	priest $\land$ female $\rightarrow \perp$
		10	singer $\land$ male $\rightarrow \perp$
		10	dancer $\land$ male $\rightarrow \perp$

Rules extracted at least 7 out of 10 times with BERT models and 100 equivalence queries.

Experiments

#	RoBERTa-base	#	RoBERTa-large
10	priest $\wedge$ female $\rightarrow \perp$	10	priest $\wedge$ female $\rightarrow \perp$
10	nurse $\wedge$ male $\rightarrow \perp$	10	nurse $\wedge$ male $\rightarrow \perp$
10	diplomat $\wedge$ female $\rightarrow \perp$		
10	mathematician $\wedge$ female $\rightarrow \perp$	10	mathematician $\wedge$ female $\rightarrow \perp$
9	banker $\wedge$ female $\rightarrow \perp$	10	banker $\wedge$ female $\rightarrow \perp$
9	footballer $\wedge$ female $\rightarrow \perp$	10	footballer $\wedge$ female $\rightarrow \perp$
8	lawyer $\wedge$ female $\rightarrow \perp$	10	lawyer $\wedge$ female $\rightarrow \perp$
		10	fashion_designer $\wedge$ male $\rightarrow \perp$
		10	dancer $\wedge$ male $\rightarrow \perp$
		7	singer $\wedge$ male $\rightarrow$ before 1875

Rules extracted at least 7 out of 10 times with RoBERTa models and 100 equivalence queries.

Experiments

# EQs	BERT-base	BERT-large	RoBERTa-base	RoBERTa-large
50	71.74	130.01	69.76	129.22
100	193.74	303.96	184.82	308.73
200	722.55	899.13	771.97	943.26

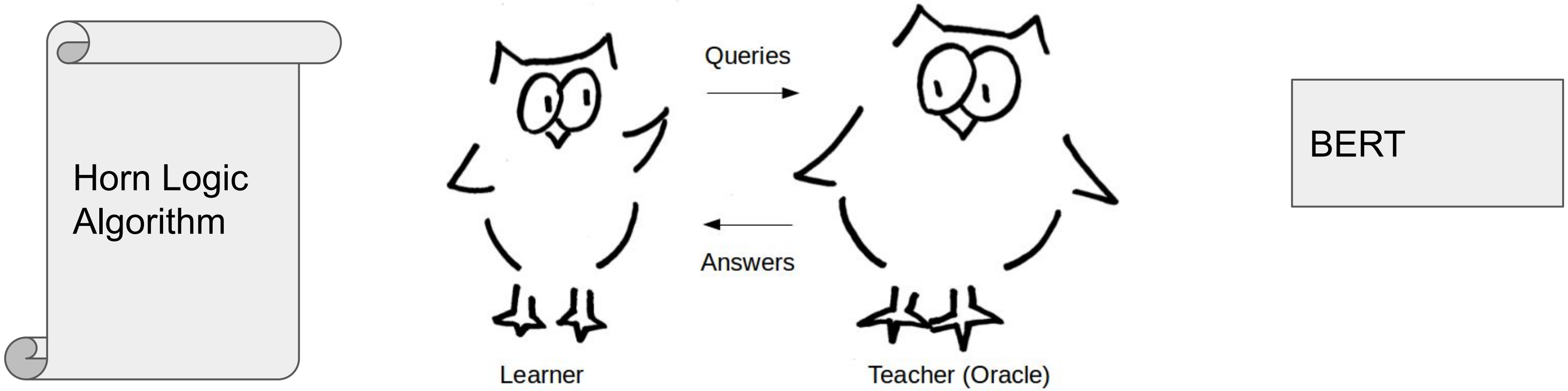
Average run time for one experiment iteration [in minutes].
This experiment took approximately 1, 3, and 13 hours per iteration with 50, 100, and 200 equivalence queries respectively for the base models on a PowerEdge R7525 Server.

Experiments

$$\text{priest} \wedge \text{female} \rightarrow \perp$$
$$\text{nurse} \wedge \text{male} \rightarrow \perp$$
$$\text{mathematician} \wedge \text{female} \rightarrow \perp$$
$$\text{footballer} \wedge \text{female} \rightarrow \perp$$
$$\text{banker} \wedge \text{female} \rightarrow \perp$$

Intersection of rules from all language models (10/10 with 200 EQs).

Conclusion


Horn Logic
Algorithm
Queries
Answers
Learner
Teacher (Oracle)
BERT