Measuring and Improving Consistency in Pretrained Language Models

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze and Yoav Goldberg

TACL @ EMNLP 2021

NLP These Days: (1)

WIKIPEDIA
The Free Encyclopedia

I am very good at completing masked [MASK] in very long sentences.

NLP These Days: (2)

The movie was amazing

Model's Failure Mode

How many birds?	**A:** 1
Is there 1 bird?	**A:** no
Are there 2 birds?	**A:** yes
Are there any birds?	**A:** no

Ribeiro et al., 2019

Model's Failure Mode

Kublai originally named his eldest son, Zhenjin, as the Crown Prince, but he died before Kublai in 1285.

(c) Excerpt from an input paragraph, **SQuAD dataset**.

Q: When did Zhenjin die?	**A:** 1285
Q: Who died in 1285?	**A:** Kublai

Ribeiro et al., 2019

Consistency

Consistency in Models

- End-task models suffer from inconsistency
- Today's standard pipeline is: Pretrain -> Finetune
- **Our theseis**: *Inconsistency of the PLM, will be realized also in the downstream tasks*

Consistency in Models: This Talk

1. Why would we care about consistency
2. ParaRel🤘: a new resource that enables us to measure consistency
3. A framework for measuring (In)Consistency in Language Models
    - In the context of factual knowledge
4. A proposal to improve consistency in LMs.

Setup: LMs as Knowledge Bases

Language Models as Knowledge Bases?

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com

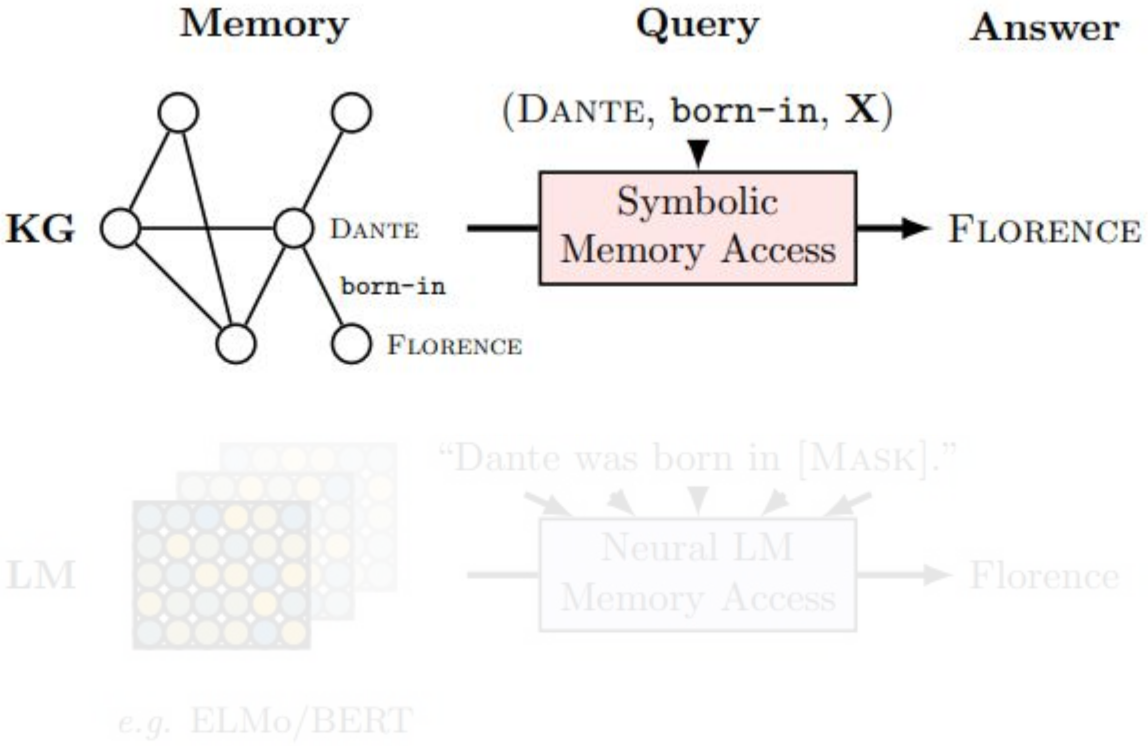


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

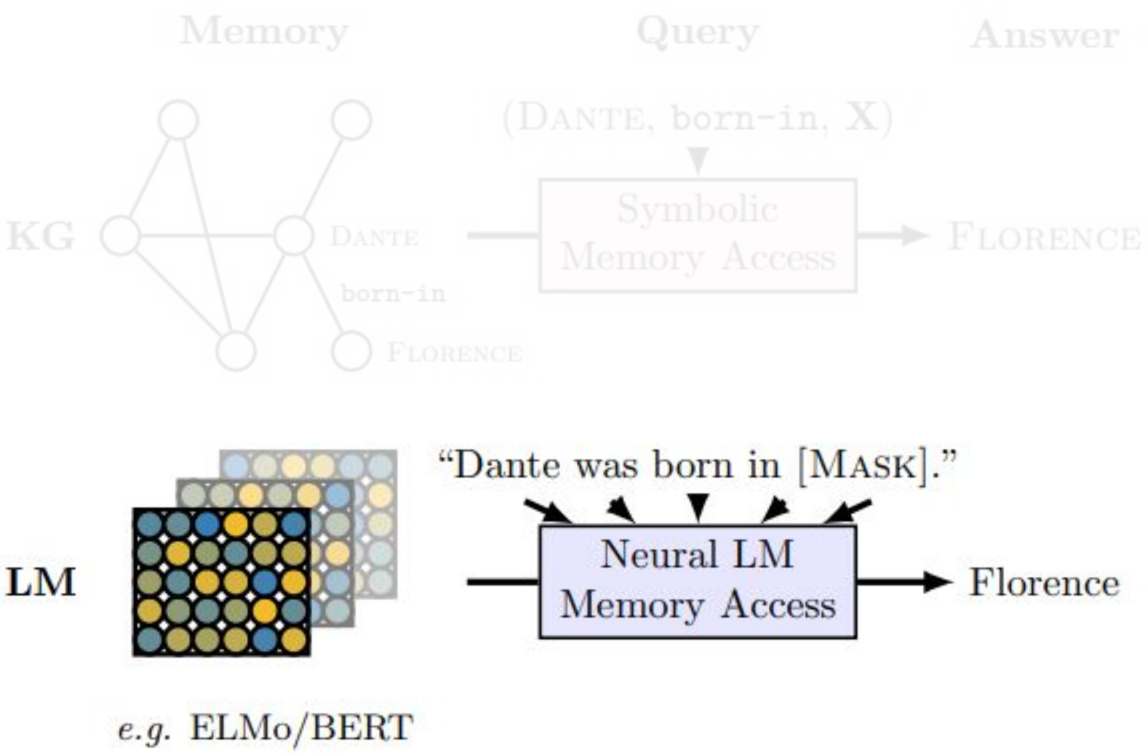


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Using Patterns to Query LMs

- Born-In: “[X] was born in [Y] .”
  - *Barack Obama was born in [MASK].*
- Broadcasting Channel: “[X] was originally aired on [Y] .”
  - *Lost was originally aired on [MASK].*

Language Models as KBs - Setup

- The data is of the form <subject, pattern, object>
- subject, object are entities in the world
- ‘pattern’ is a linguistic expression that expresses a relation
- E.g. <“Barack Obama”, “X was born in Y”, “Hawaii”>
- Given the subject and relation, the task is to predict the object
    - E.g. <“Barack Obama”, born-in> -> “Hawaii”
    - In Petroni et al., 2019, used 1 pattern for every relation

Language Models as KBs

- LMs were trained on large sources of knowledge (e.g. Wikipedia)
- Can capture (memorize) some of these facts as part of the pretraining objective

Pretraining a Language Model

Background

Early life of Barack Obama

Main articles: Early life and career of Barack Obama and Ann Dunham

People who express doubts about Obama's eligibility or reject details about his early life are often informally called "birthers", a term that parallels[23] the nickname "truthers" for adherents of 9/11 conspiracy theories.[24][25] These conspiracy theorists reject at least some of the following facts about his early life:

Barack Obama was born on August 4, 1961, at Kapi'olani Maternity & Gynecological Hospital (now called Kapi'olani Medical Center for Women & Children) in Honolulu, Hawaii,[26][27][28][29] to Ann Dunham,[30] from Wichita, Kansas,[31] and her husband Barack Obama Sr., a Luo from Nyang'oma Kogelo, Nyanza Province (in what was then the Colony and Protectorate of Kenya), who was attending the University of Hawaii. Birth notices for Barack Obama were published in *The Honolulu Advertiser* on August 13 and the *Honolulu Star-Bulletin* on August 14, 1961.[26][31] Obama's father's immigration file also clearly states Barack Obama was born in Hawaii.[32] One of his high school teachers, who was acquainted with his mother at the time, remembered hearing about the day of his birth.[30]

Pretraining a Language Model

Background

Early life of Barack Obama

Main articles: Early life and career of Barack Obama and Ann Dunham

People who express doubts about Obama's eligibility or reject details about his early life are often informally called "birthers", a term that parallels[23] the nickname "truthers" for adherents of 9/11 conspiracy theories.[24][25] These conspiracy theorists reject at least some of the following facts about his early life:

Barack Obama was born on August 4, 1961, at Kapi'olani Maternity & Gynecological Hospital (now called Kapi'olani Medical Center for Women & Children) in Honolulu, Hawaii,[26][27][28][29] to Ann Dunham,[30] from Wichita, Kansas,[31] and her husband Barack Obama Sr., a Luo from Nyang'oma Kogelo, Nyanza Province (in what was then the Colony and Protectorate of Kenya), who was attending the University of Hawaii. Birth notices for Barack Obama were published in *The Honolulu Advertiser* on August 13 and the *Honolulu Star-Bulletin* on August 14, 1961.[26][31] Obama's father's immigration file also clearly states Barack Obama was born in Hawaii.[32] One of his high school teachers, who was acquainted with his mother at the time, remembered hearing about the day of his birth.[30]

Pretraining a Language Model

And it works!

LM predictions

#1 mask:Tel Aviv is located in **[MASK]**.

	bert_large_cased
0	Israel
1	Jerusalem
2	Palestine
3	Haifa
4	Egypt
5	Europe
6	Ukraine
7	Lebanon
8	Jordan
9	Germany

Pretraining a Language Model

Well, sometimes…

LM predictions

#1 mask:Barack Obama was born in **[MASK]**.

	bert_large_cased
0	Chicago
1	Philadelphia
2	Detroit
3	Houston
4	Atlanta
5	Georgia
6	Boston
7	Texas
8	Paris
9	Dallas

Language Models as KBs

- This factual knowledge cannot appear from thin air
- So what is the problem?
- The way we would use the LM-as-KB:
    - Query via natural language, which varies across users, without a specific schema

Language Models as KBs

So the real question is

Does It Generalize?

Language Models as KBs - Consistency?

We'd like that an LM would make the same prediction across paraphrases

E.g.:

"*Seinfeld* was aired on [Y]."

- ↔️ "*Seinfeld*, that was aired on [Y],"
- ↔️ "[Y]'s series *Seinfeld*,"

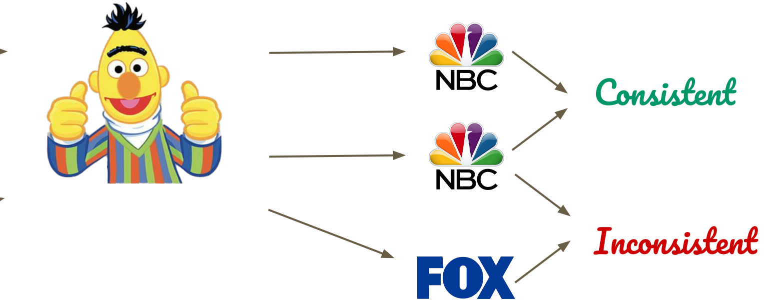
Language Models as KBs - Consistency?

We'd like that an LM would make the same prediction across paraphrases

E.g.:

"*Seinfeld* was aired on [Y]."

- ↔️ "*Seinfeld*, that was aired on [Y],"
- ↔️ "[Y]'s series *Seinfeld*,"



ParaRel 🤘

Language Models as KBs - ParaRel 🤘

But where can we get these patterns?

We build a new resource:

ParaRel🤘 (**Para**phrase **Rel**ations)

ParaRel 🤘 - Creation

- For every relation, we manually build a set of patterns that are paraphrases of each other, in 4 steps:
  - Starting with the single pattern from LAMA (*Petroni et al., 2019*)
  - Augmenting with automatically extracted patterns from LPAQA (*Jiang et al., 2020*)
  - Searching for patterns in wikipedia using SPIKE (*Shlain et al., 2020*)
  - Additional patterns using linguistic expertise of the authors

ParaRel 🤘 - Creation

- Was manually collected by the authors of this paper
- 2 additional authors verified the patterns, while engaging in discussion to reach an agreement (discarding otherwise)
- Human Eval: Sampled 156 pairs, and asked NLP grad students to annotate. Reaching **95.5%** agreement (and later fixed the errors)

ParaRel 🤘 - Summary

# Relations	38
# Patterns	328
Min # patterns	2
Max # patterns	20
Avg # patterns	8.63

Setup & Evaluation

Consistency - Setup

Data Pairs (D)

$(D_1, r_1, P_1), \ldots, (D_i, r_i, P_i), \ldots, (D_n, r_n, P_n)$

Patterns (P)

D_1
(*Lou Reed, Brooklyn*)
(*Masako Natsume, Tokyo*)
...

...

D_i
(*Seinfeld, NBC*)
(*Homeland, Showtime*)
...

...

r_i =*originally-aired-on*

Homeland originally aired on *[MASK]*
Homeland premiered on *[MASK]*
...
Seinfeld originally aired on *[MASK]*
Seinfeld premiered on *[MASK]*

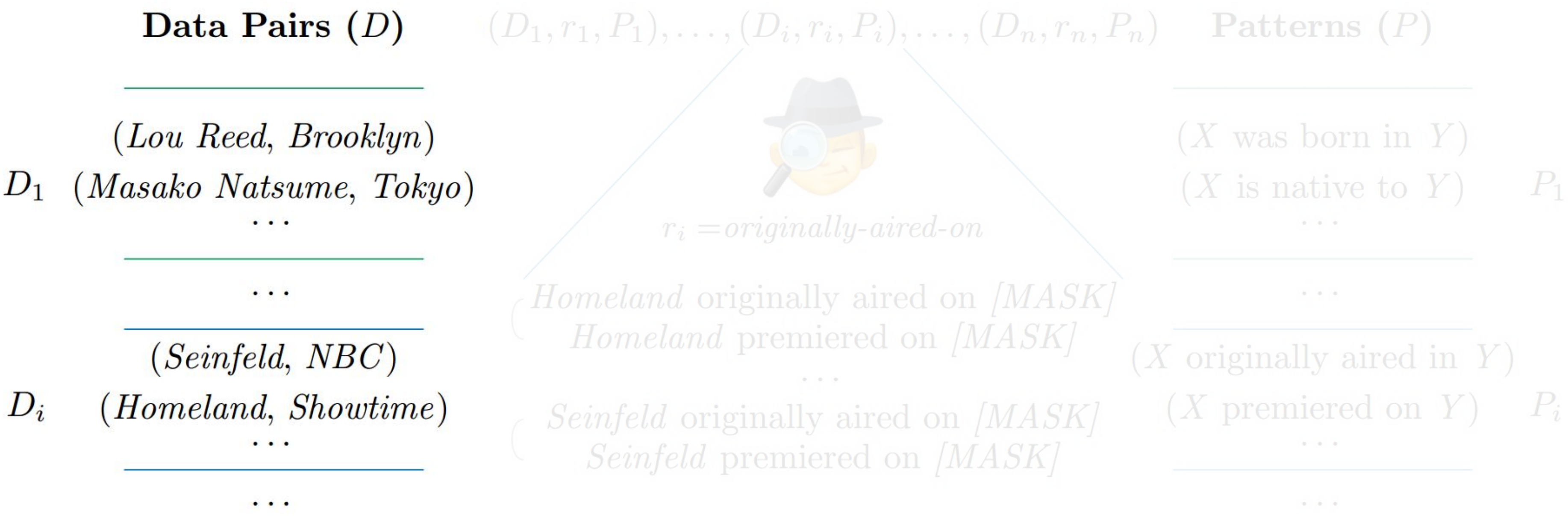
P_1
(X was born in Y)
(X is native to Y)
...

...

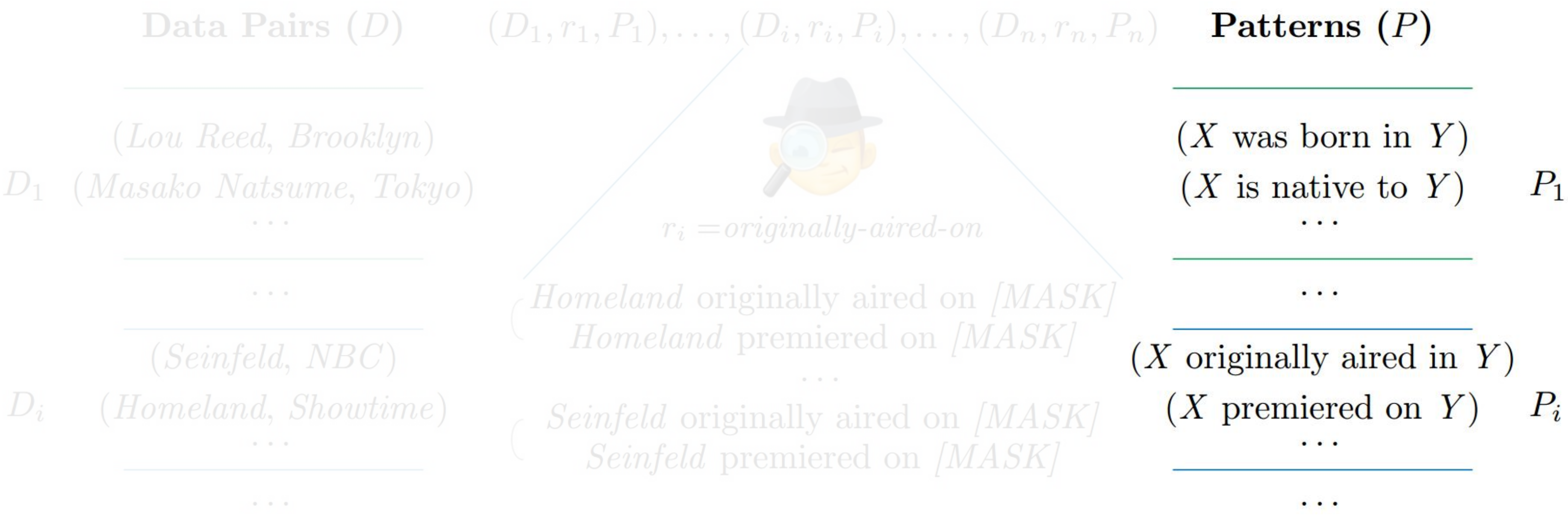
P_i
(X originally aired in Y)
(X premiered on Y)
...

...

Consistency - Setup



Consistency - Setup



Consistency - Setup

$$(D_1, r_1, P_1), \ldots, (D_i, r_i, P_i), \ldots, (D_n, r_n, P_n)$$

Data Pairs (D)

D_1: (*Lou Reed*, *Brooklyn*), (*Masako Natsume*, *Tokyo*), ...

...

D_i: (*Seinfeld*, *NBC*), (*Homeland*, *Showtime*), ...

...

r_i =*originally-aired-on*

Homeland originally aired on *[MASK]*
Homeland premiered on *[MASK]*

...

Seinfeld originally aired on *[MASK]*
Seinfeld premiered on *[MASK]*

Patterns (P)

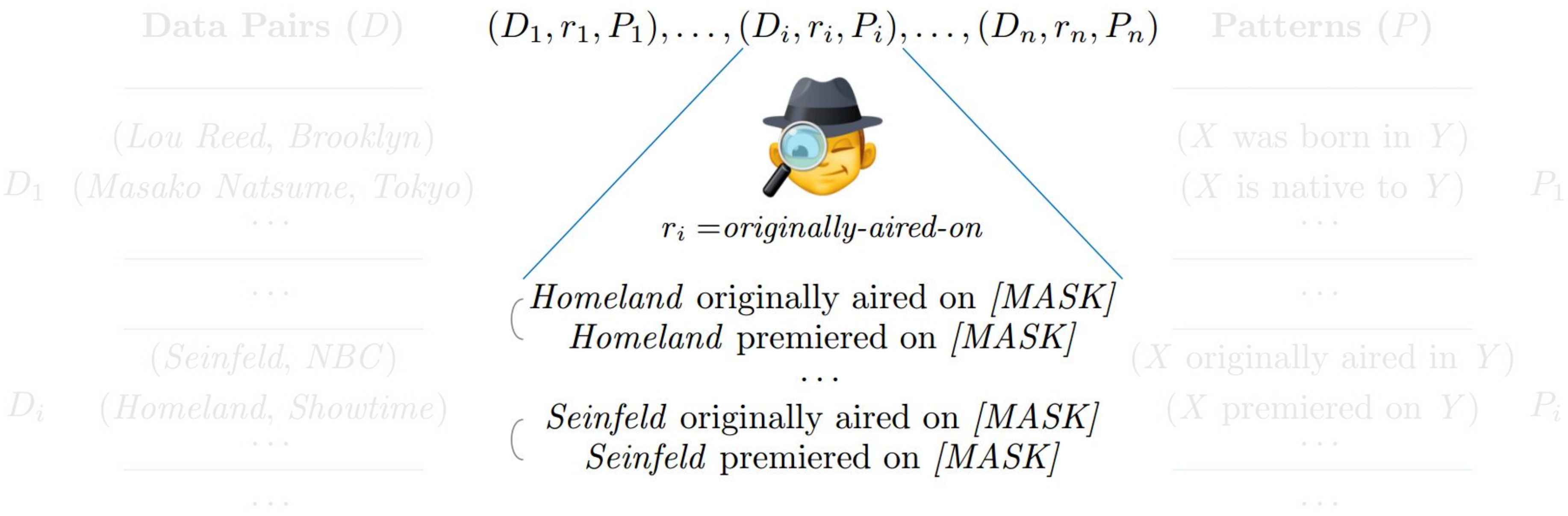
P_1: (X was born in Y), (X is native to Y), ...

...

P_i: (X originally aired in Y), (X premiered on Y), ...

...

Consistency - Setup



Consistency - Models

- BERT
- BERT Whole-Word-Masking
- RoBERTa
- ALBERT

And a Baseline:

- Most common object (consistent by definition)

Consistency - Evaluation

- **Accuracy**: Accurate prediction of the LAMA pattern
- **Consistency**: For each relation and tuple, compute all paraphrases pairs, and test if the predictions are equal: n(n-1)/2 pairs
- **Consistent-Acc**: Consistent and accurate prediction of all paraphrases

Results

Consistency - Results

Model	Accuracy	Consistency	Consistent-Acc
majority	23.1+-21.0	100.0+-0.0	23.1+-21.0
BERT-base	45.8+-25.6	58.5+-24.2	27.0+-23.8
BERT-large	48.1+-26.1	**61.1**+-23.0	**29.5**+-26.6
BERT-large-wwm	**48.7**+-25.0	60.9+-24.2	29.3+-26.9
RoBERTa-base	39.0+-22.8	52.1+-17.8	16.4+-16.4
RoBERTa-large	43.2+-24.7	56.3+-20.4	22.5+-21.1
ALBERT-base	29.8+-22.8	49.8+-20.1	16.7+-20.3
ALBERT-xxlarge	41.7+-24.9	52.1+-22.4	23.8+-24.8

Consistency - Results

Model	Accuracy	Consistency	Consistent-Acc
majority	23.1+-21.0	100.0+-0.0	23.1+-21.0
BERT-base	45.8+-25.6	58.5+-24.2	27.0+-23.8
BERT-large	48.1+-26.1	**61.1**+-23.0	**29.5**+-26.6
BERT-large-wwm	**48.7**+-25.0	60.9+-24.2	29.3+-26.9
RoBERTa-base	39.0+-22.8	52.1+-17.8	16.4+-16.4
RoBERTa-large	43.2+-24.7	56.3+-20.4	22.5+-21.1
ALBERT-base	29.8+-22.8	49.8+-20.1	16.7+-20.3
ALBERT-xxlarge	41.7+-24.9	52.1+-22.4	23.8+-24.8

Consistency - Results

Model	Accuracy	Consistency	Consistent-Acc
majority	23.1+-21.0	100.0+-0.0	23.1+-21.0
BERT-base	45.8+-25.6	58.5+-24.2	27.0+-23.8
BERT-large	48.1+-26.1	**61.1**+-23.0	**29.5**+-26.6
BERT-large-wwm	**48.7**+-25.0	60.9+-24.2	29.3+-26.9
RoBERTa-base	39.0+-22.8	52.1+-17.8	16.4+-16.4
RoBERTa-large	43.2+-24.7	56.3+-20.4	22.5+-21.1
ALBERT-base	29.8+-22.8	49.8+-20.1	16.7+-20.3
ALBERT-xxlarge	41.7+-24.9	52.1+-22.4	23.8+-24.8

Consistency - Results

Model	Accuracy	Consistency	Consistent-Acc
majority	23.1+-21.0	100.0+-0.0	23.1+-21.0
BERT-base	45.8+-25.6	58.5+-24.2	27.0+-23.8
BERT-large	48.1+-26.1	**61.1**+-23.0	**29.5**+-26.6
BERT-large-wwm	**48.7**+-25.0	60.9+-24.2	29.3+-26.9
RoBERTa-base	39.0+-22.8	52.1+-17.8	16.4+-16.4
RoBERTa-large	43.2+-24.7	56.3+-20.4	22.5+-21.1
ALBERT-base	29.8+-22.8	49.8+-20.1	16.7+-20.3
ALBERT-xxlarge	41.7+-24.9	52.1+-22.4	23.8+-24.8

Consistency - Results

Model	Accuracy	Consistency	Consistent-Acc
majority	23.1+-21.0	100.0+-0.0	23.1+-21.0
BERT-base	45.8+-25.6	58.5+-24.2	27.0+-23.8
BERT-large	48.1+-26.1	**61.1**+-23.0	**29.5**+-26.6
BERT-large-wwm	**48.7**+-25.0	60.9+-24.2	29.3+-26.9
RoBERTa-base	39.0+-22.8	52.1+-17.8	16.4+-16.4
RoBERTa-large	43.2+-24.7	56.3+-20.4	22.5+-21.1
ALBERT-base	29.8+-22.8	49.8+-20.1	16.7+-20.3
ALBERT-xxlarge	41.7+-24.9	52.1+-22.4	23.8+-24.8

Consistency - Results

Model	Accuracy	Consistency	Consistent-Acc
majority	23.1+-21.0	100.0+-0.0	23.1+-21.0
BERT-base	45.8+-25.6	58.5+-24.2	27.0+-23.8
BERT-large	48.1+-26.1	**61.1**+-23.0	**29.5**+-26.6
BERT-large-wwm	**48.7**+-25.0	60.9+-24.2	29.3+-26.9
RoBERTa-base	39.0+-22.8	52.1+-17.8	16.4+-16.4
RoBERTa-large	43.2+-24.7	56.3+-20.4	22.5+-21.1
ALBERT-base	29.8+-22.8	49.8+-20.1	16.7+-20.3
ALBERT-xxlarge	41.7+-24.9	52.1+-22.4	23.8+-24.8

Consistency - Results

Model	Accuracy	Consistency	Consistent-Acc
majority	23.1+-21.0	100.0+-0.0	23.1+-21.0
BERT-base	45.8+-25.6	58.5+-24.2	27.0+-23.8
BERT-large	48.1+-26.1	**61.1**+-23.0	**29.5**+-26.6
BERT-large-wwm	**48.7**+-25.0	60.9+-24.2	29.3+-26.9
RoBERTa-base	39.0+-22.8	52.1+-17.8	16.4+-16.4
RoBERTa-large	43.2+-24.7	56.3+-20.4	22.5+-21.1
ALBERT-base	29.8+-22.8	49.8+-20.1	16.7+-20.3
ALBERT-xxlarge	41.7+-24.9	52.1+-22.4	23.8+-24.8

Consistency - Summary

We have shown that:

1. The models are inconsistent
    a. Although there is a high variance between relations
2. Some models are more consistent than others

Much more analysis and experiments in the paper!!

Improved Consistency

Improved Consistency

- Can we improve the consistency of PLMs?
- We want predictions from paraphrases to be equal

$$\min_{\theta} \mathrm{sim}(\arg\max_{i} f_{\theta}(P_n)[i], \arg\max_{j} f_{\theta}(P_m)[j])$$

But, this involves argmax, and it's hard to optimize for

Improved Consistency

- We go on a softer version, and try to make the distributions alike

$$Q_n = softmax(f_\theta(P_n))$$

Improved Consistency

- We go on a softer version, and try to make the distributions alike

$$Q_n = softmax(f_\theta(P_n))$$

$$\mathcal{L}_c = \sum_{n=1}^{k} \sum_{m=n+1}^{k} D_{KL}(Q_n^{r_i} || Q_m^{r_i}) + D_{KL}(Q_m^{r_i} || Q_n^{r_i})$$

Improved Consistency

- We go on a softer version, and try to make the distributions alike

$$Q_n = softmax(f_\theta(P_n))$$

$$\mathcal{L}_c = \sum_{n=1}^{k} \sum_{m=n+1}^{k} D_{KL}(Q_n^{r_i} || Q_m^{r_i}) + D_{KL}(Q_m^{r_i} || Q_n^{r_i})$$

We also continue the pretraining objective (MLM)

$$\mathcal{L} = \lambda \mathcal{L}_c + \mathcal{L}_{MLM}$$

Improved Consistency

- We only use 3 relations
- Use their corresponding tuples (subject, pattern, object)
- Train for 3 epochs, with early stopping

Improved Consistency

Model	Accuracy	Consistency	Consistent-Acc
majority	24.4+-22.5	100.0+-0.0	24.4+-22.5
BERT-base	45.6+-27.6	58.2+-23.9	27.3+-24.8
BERT-ft	**47.4**+-27.3	**64.0**+-22.9	**33.2**+-27.0

Summary

- We created (and released) ParaRel 🤘, 328 manually written patterns for 38 relations
- We test whether popular LMs are consistent to factual knowledge...
    - and show empirically they **are not!**
- We experiment with a novel pretraining loss for improving consistency in LMs
    - And improve consistency in PLMs
    - But much work still remains!



Thanks!

Questions?