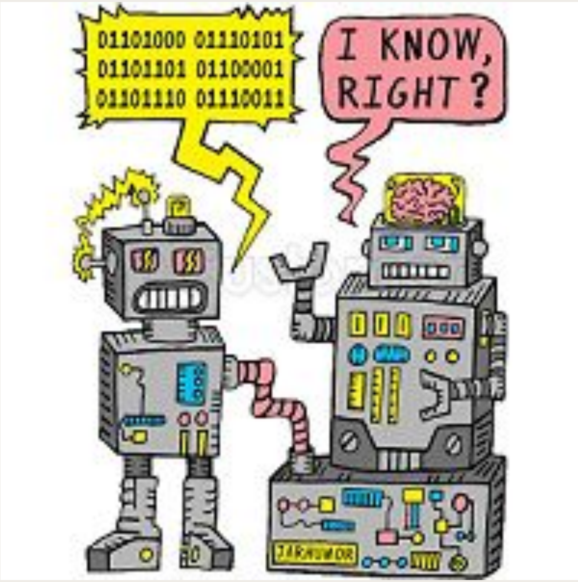


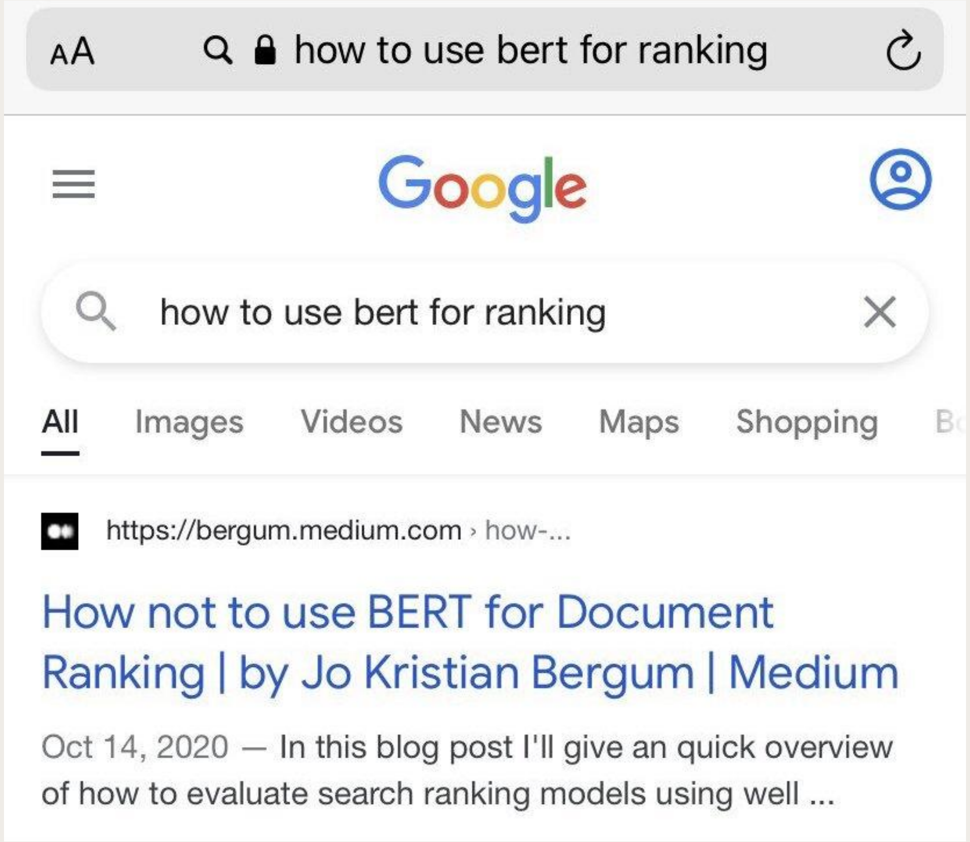
Attention Heads and Negation Focus Detection

LING575 - Spring 2022

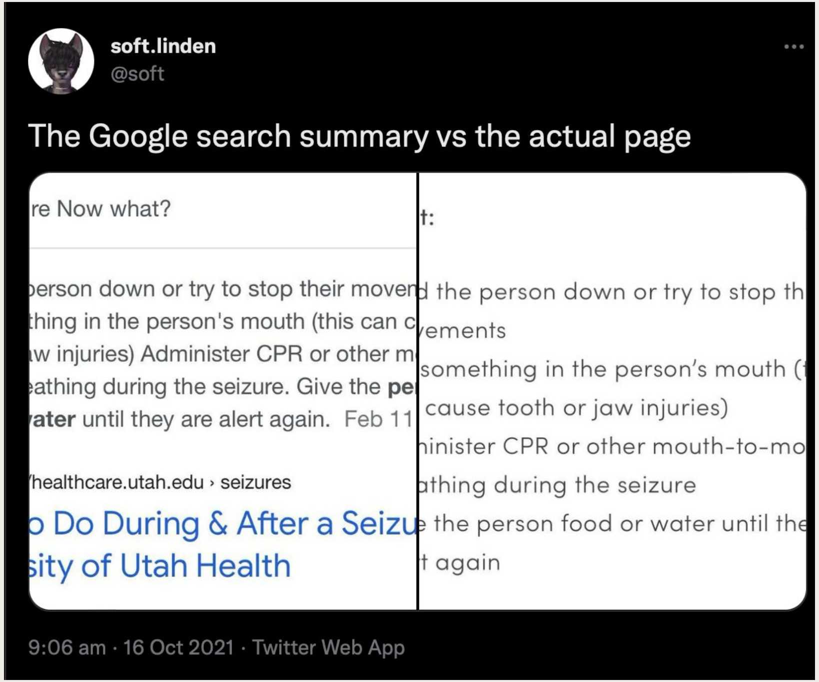
Group 7

Chirag Soni, Amélie Reymond, Sanjana Sharma, Soma Dhavala

Isn’t negation trivial?



But can have serious consequences…



Source: https://twitter.com/soft/status/1449406390976409600?s=21&t=TwRf97eyVfWIUJfZ4RlocA

Had a seizure Now what?

Hold the person down or try to stop their movements. Put something in the person's mouth (this can cause tooth or jaw injuries) Administer CPR or other mouth-to-mouth breathing during the seizure. Give the **person food or water** until they are alert again. Feb 11, 2021

https://healthcare.utah.edu › seizures

What to Do During & After a Seizure | University of Utah Health

Do not:

- Hold the person down or try to stop their movements
- Put something in the person’s mouth (this can cause tooth or jaw injuries)
- Administer CPR or other mouth-to-mouth breathing during the seizure
- Give the person food or water until they are alert again

Outline

1. Background
- Why care about negation?
- Detecting negation
2. How do PLMs handle negation? - literature review
3. Looking at attention heads
4. Our experiment
   a. Negation Focus
   b. Negation under Factual Correctness

I. Background : Negation - why care?

- Important property in many **NLU tasks**: sentiment analysis, QA and natural language inference
- “All human systems of communication contain a representation of negation” (Horn, 1989)s

I. Background : Negation - why care?

- **Frequent** phenomenon in language: approx. 25% of English sentences, depending on domain and genre (Hossain et al. 2020)
- Psychologically more difficult to process (e.g.: Just and Carpenter, 1971)

I. Background : Negation detection

- From a linguistic perspective: negation has scope and focus, crucial to capture its semantics
- Negation cue = tokens that express negation (no, not, never, n’t, …)
- Scope = part of the meaning that is negated
- Focus = part of the scope that is most prominently or explicitly negated

I. Background : Negation detection in NLP

a. [John had] **never** [said as much before]

b. John had never said {as much} before

- Never is the negation **cue**
- In [ ] brackets: **scope**
- In { } brackets: **focus**

I. Background : Negation detection in NLP

a. The government didn't release the UFO files {until 2008}

→ the government didn't release the UFO files but {not after 2008}

I. Background : Negation detection in NLP

- Negation is tricky for NLP
- In logic two negatives cancel each other out: $A \equiv \sim(\sim A)$
- Not always the case in natural language:

  “She is not unhappy” =/= “She is happy”

  → “She is not fully unhappy but not really happy either”

- Sometimes implicit meaning

  “Cows do not eat meat” → “Cows eat something else”

II. So… how do PLMs handle Negation?

A literature review on analyzing negation in pre-trained language models

- What BERT is not, Ettinger, 2019
- Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly, Kassner & Schütze, 2020
- An Analysis of Natural Language Inference Benchmarks Through the Lens of Negation, Hossain et al., 2020
- Investigating Negation in Pre-trained Vision-and-language Models. Dobreva & Keller, 2021

II. How do PLMs handle Negation?

What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models

Allyson Ettinger
Department of Linguistics
University of Chicago
aettinger@uchicago.edu

- comparison with experiments in psycholinguistics

II. How do PLMs handle Negation?

Results

- When the statement is affirmative, BERT assigns a higher probability to the true completion to 100% of items
- But for negative statements BERT assigns a higher probability to 0% of the true completion!

→ BERT's strong insensitivity to the meaning negation

	Affirmative	Negative
$\text{BERT}_{\text{BASE}}$	100	0.0
$\text{BERT}_{\text{LARGE}}$	100	0.0

Table 12: Percent of NEG-136-SIMP items with true completion assigned higher probability than false

II. How do PLMs handle Negation?

Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly

Nora Kassner, Hinrich Schütze
Center for Information and Language Processing (CIS)
LMU Munich, Germany
kassner@cis.lmu.de

- Result: PLMs have difficulty distinguishing between positive negative sentences
- More later

II. How do PLMs handle Negation?

An Analysis of Natural Language Inference Benchmarks through the Lens of Negation

Md Mosharaf Hossain,[θ] **Venelin Kovatchev,**[3] **Pranoy Dutta,**[θ] **Tiffany Kao,**[θ]
Elizabeth Wei,[θ] and **Eduardo Blanco**[θ]
[θ]University of North Texas [3]University of Barcelona
mdmosharafhossain@my.unt.edu vkovatchev@ub.edu
{PranoyDutta,TiffanyKao,ElizabethWei}@my.unt.edu eduardo.blanco@unt.edu

- ~ 25% of English sentences contain negation
- under-represented in common NLI benchmarks (RTE, SNLI)
- Creation of new benchmark for NLI by adding more negation to the original benchmarks: 4500 pairs of text-hypotheses containing negation

	#sents.	% w/ neg.
General English		
Online Reviews		
books	4,845,154	22.64
movies	616,287	28.97
Conversations		
oral	538,973	27.43
written	510,458	29.92
Wikipedia	2,735,930	8.69
Books	1,809,184	28.45
OntoNotes	63,918	17.14
NLI benchmarks		
RTE	16,389	7.16
SNLI	1,138,598	1.19
MNLI	883,436	22.63

II. How do PLMs handle Negation?

Experiment

- BERT, XLNet, RoBERTa
- Cannot solve the task on the new test set when training on original benchmarks (RTE, SNLI, MNLI)
- Only slight improvement when fine-tuned on new text-hypothesis pairs (depends on the original benchmark)

→ negation is **still a challenge for NLI** despite what it may seems from looking at common benchmarks!

II. How do PLMs handle Negation?

Investigating Negation in Pre-trained Vision-and-language Models

Radina Dobreva and **Frank Keller**
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
r.dobreva@ed.ac.uk, keller@inf.ed.ac.uk

- Natural Language Visual Reasoning for Real (NVLR2) task
- Two images and a sentence: is the sentence true of the images?
- New benchmark: same images but added negation (9.6% only in original dataset)

The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

One image shows exactly two brown acorns in back-to-back caps on green foliage.

II. How do PLMs handle Negation?

Results

- Both models L&V models used show a drop in performance on the negation samples

  → Language & vision models also find it hard to handle negation

	LXMERT		UNITER$_{paired-attn}$		UNITER$_{triplet}$	
	negative	positive	negative	positive	negative	positive
Verbal (content)	28.72	69.23	43.62	73.63	43.62	71.43
Verbal (existential)	30.56	82.41	50.0	77.77	44.44	66.66
NP (nonexistential)	44.83	67.86	48.28	64.29	55.17	50.0
NP (existential)	34.55	80.0	50.91	85.45	32.73	87.27
NP (number-to-none)	54.17	72.22	51.39	77.77	55.56	76.39
Sentence-wide	38.55	66.27	31.33	69.87	38.55	65.06
Overall	36.96	73.5	45.35	76.5	44.22	71.5

Table 3: Accuracy on the negation test set and the corresponding non-negated (positive) examples.

II. How do PLMs handle Negation?

... Not very well!

But why? Is it something to do with the detection of **scope**?

III. Looking at Attention Heads in BERT

How does BERT's attention change when you fine-tune?
An analysis methodology and a case study in negation scope

Yiyun Zhao
Department of Linguistics
University of Arizona
yiyunzhao@arizona.edu

Steven Bethard
School of Information
University of Arizona
bethard@arizona.edu

- Intuition: if a word is within negation cue, its maximal attention will be on negation cue
- Clark et al. : inspection of pre-trained transformers' attention mechanism
- Some syntactic properties are encoded intuitively
- E.g.: the maximum attention of a dependent is on its syntactic head

III. Looking at Attention Heads in BERT

- Zhao & Bethard argue that it is important to show that encodings are enhanced after fine-tuning on tasks that require linguistic knowledge
- If that is not the case then that means model could use another mechanism

III. Looking at Attention Heads in BERT

Overview of the methodology used by Zhao & Bethard:

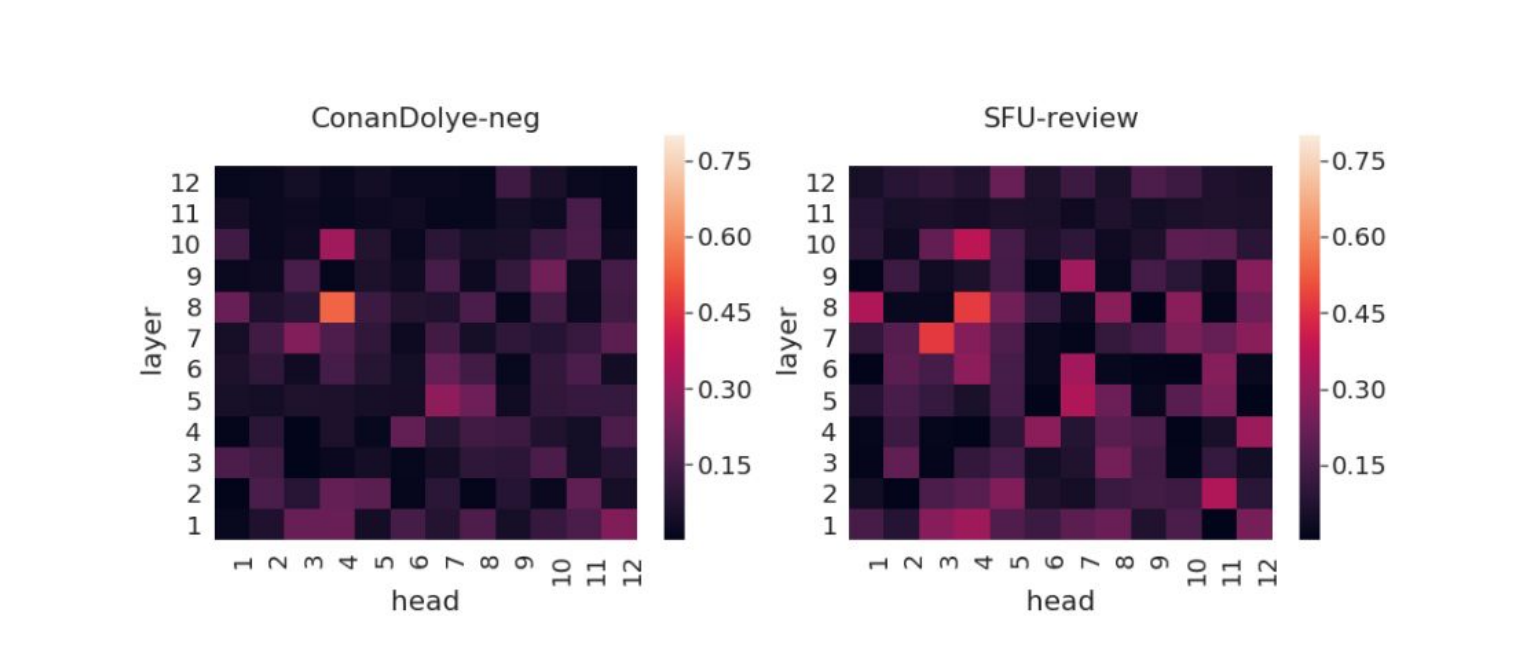
- **Hypothesize a representation** of phenomenon of interest (here: negation scope)
- Identify a relevant **downstream** task (supervised negation scope problems)
- Design a **control task** where the phenomenon is irrelevant, learnable without any knowledge of the linguistic phenomenon (word types to binary labels)
- **Differences** between fined-tuned models on control and downstream task

III. Looking at Attention Heads in BERT

Results

- Fine-tuning does improve for the heads that are already good at detecting negation
- But only for BERT-base and RoBERTA-base
- Weaker evidence for the larger versions: possible other mechanism to encode negation

III. Looking at Attention Heads in BERT



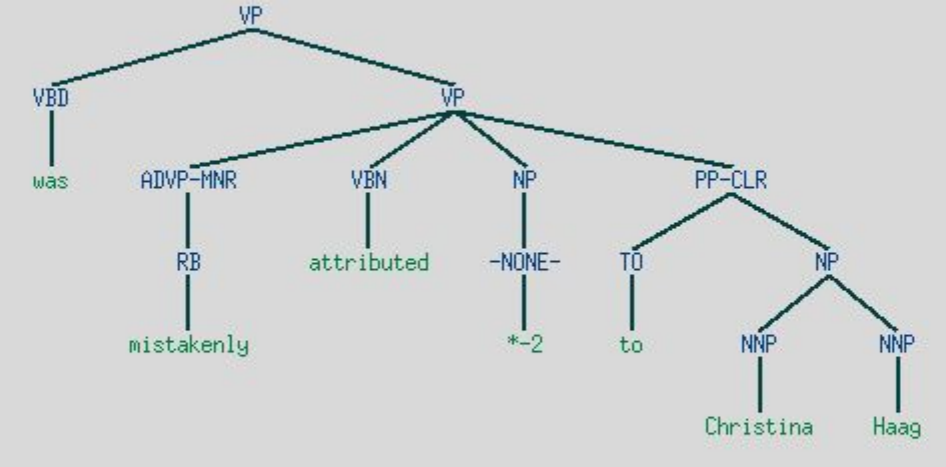
Similar results across two different datasets

Our Experiment

- IV. A: Analyzing **Attention Heads for Focus**, following Zhao and Bethard's work on negation scope
- IV B: Probing the **LMs as Knowledge Bases**, following Kassner & Schutze

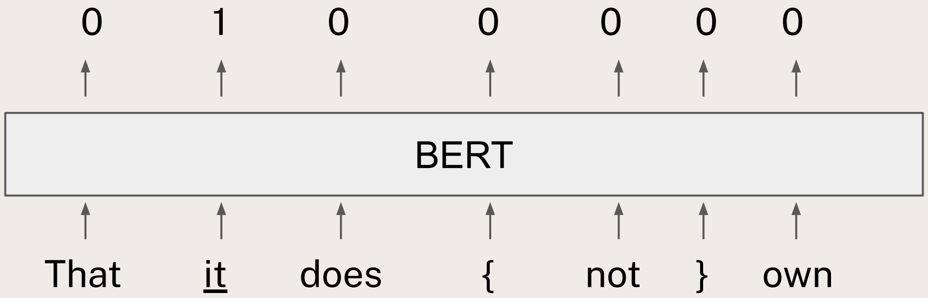
PB-FOC Dataset

- Annotates focus and negation cue in the sentence.
- Along with POS tag, NE, dependency relations, semantic roles.
- Only annotations included, not the actual words.
- Words from Penn Treebank were provided during the competition by LDC.
- No documentation about which words are expected out of the treebank.



Probing Attention for Focus

- Downstream task
  - Model as sequence tagging (token classification task)
  - 0 = not in focus, 1 = in focus



- Control task
  - Assign 0s and 1s randomly

Probing Attention for Focus

- Unsupervised probe
  - If word pays maximum attention to the negation cue => in focus

$$attendneg(i) = \begin{cases} 1 & \text{if } j_{neg} = \operatorname*{argmax}_{j=1}^{n} a_{ij} \\ 0 & otherwise \end{cases}$$

- Calculate precision, recall and F_1 on both downstream and control task.
- Compare with fixed offset baseline.
- Compare with and without fine-tuning.

IV-B. Negation under Factual Correctness

Premise:

- Languages Models act as Knowledge Bases (
    - LAMA is a dataset of cloze statements
    - Eg: Munich is the capital of Germany
    - Probe is an NLG task

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]

- PLMs do not distinguish between Negated and Non-negated cloze statements (Kassner et al)
    - Negated LAMA dataset
    - Eg: Munich is not the capital of Germany
    - Misprimed Probe is an NLG task

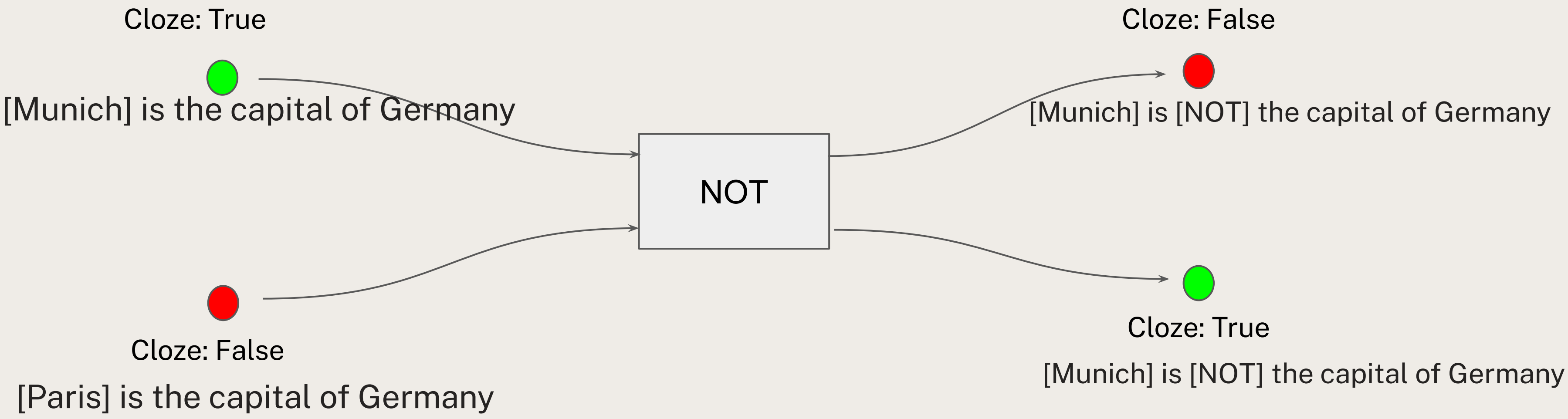
Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly

Nora Kassner, Hinrich Schütze

Our Question:

- How do Negations and Factual (In)Correctness interact?

IV-B. Negation under Factual Correctness



IV-B. PLMs as KBs

Languages Models as Knowledge Bases? Petroni et al - posited whether PLMs act and behave as KBs. They create LAMA dataset which is a collection of factual [cloze] statements, harvested from Wiki, Google-RE, ConceptNet

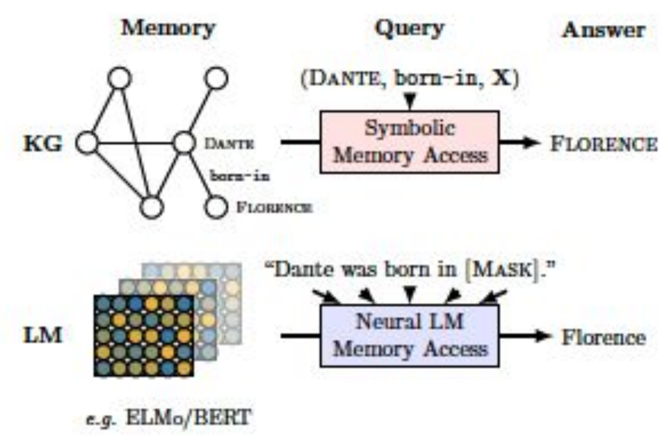


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

Eg: Munich is the capital of Germany

Task: [MASK] is the capital of Germany

PLMs as KBs

Languages Models as Knowledge Bases? Petroni et al - posited whether LLMs act and behave as KBs. They create LAMA dataset which is a collection of factual [cloze] statements, harvested from Wiki, Google-RE, ConceptNet

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Table 2: Mean precision at one (P@1) for a frequency baseline (Freq), DrQA, a relation extraction with naïve entity linking (RE_n), oracle entity linking (RE_o), fairseq-fconv (Fs), Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl) across the set of evaluation corpora.

NLG: Munich is the capital of Germany/Europe/..

BERT does better than other models in terms of

Mean P@K scores

PLMs do not act on Negation cue

Birds can not fly but can talk? Kassner et al - posited whether PLMs distinguish Negated cloze from the Non-negated

Version	Query
A	Dinosaurs? Munich is located in [MASK] .
B	Somalia? Munich is located in [MASK] .
C	Prussia? Munich is located in [MASK] .
D	Prussia? “This is great”. . . . “What a surprise.” “Good to know.” . . . Munich is located in [MASK] .

Table 1: Examples for different versions of misprimes: (A) are randomly chosen, (B) are randomly chosen from correct fillers of different instances of the relation, (C) were top-ranked fillers for the original cloze question but have at least a 30% lower prediction probability than the correct object. (D) is like (C) except that 20 short neutral sentences are inserted between misprime and MASK sentence.

		Facts	Rels	Txl		Eb		E5b		Bb		Bl	
				ρ	%	ρ	%	ρ	%	ρ	%	ρ	%
Google-RE	birth-place	2937	1	92.8	47.1	97.1	28.5	96.0	22.9	89.3	11.2	88.3	20.1
	birth-date	1825	1	87.8	21.9	92.5	1.5	90.7	7.5	70.4	0.1	56.8	0.3
	death-place	765	1	85.8	1.4	94.3	57.8	95.9	80.7	89.8	21.7	87.0	13.2
T-REx	1-1	937	2	89.7	88.7	95.0	28.6	93.0	56.5	71.5	35.7	47.2	22.7
	N-1	20006	23	90.6	46.6	96.2	78.6	96.3	89.4	87.4	52.1	84.8	45.0
	N-M	13096	16	92.4	44.2	95.5	71.1	96.2	80.5	91.9	58.8	88.9	54.2
ConceptNet	-	2996	16	91.1	32.0	96.8	63.5	96.2	53.5	89.9	34.9	88.6	31.3
SQuAD	-	305	-	91.8	46.9	97.1	62.0	96.4	53.1	89.5	42.9	86.5	41.9

Table 2: PLMs do not distinguish positive and negative sentences. Mean spearman rank correlation (ρ) and mean percentage of overlap in first ranked predictions (%) between the original and the negated queries for Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl).

Rank correlation between LAMA and Negated LAMA is very high – implying Negation is not accounted

Interaction between Negation & Factual Correctness

Kassner & Schütze considered MLM as a task (no fine-tuning or no shallow classifier), similar to Petroni et al LAMA

Zhou & Bethard studied Negation Scope using Attention probes

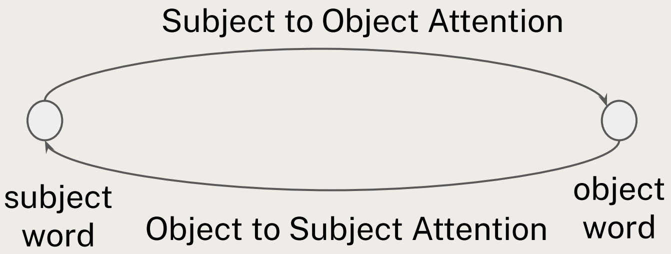
We'd like to:

- Study Negation Focus (different from Zhou & Bethard)
- Use Attention Probes (different from Kassner & Schütze)
- Create a new dataset out of Negated LAMA to provide more control tasks
- Exploratory flavour (not confirmatory)
  - Can they be factually correct under negation

Impact on Subject-Object Attention

Construct three paired sentences as follows:

	Is Factually Correct (X=1/0)	Is it negated (Y=1/0)	Cloze
A	Yes	No	Munich is the capital of Germany
B	Yes	Yes	PARIS is NOT the capital of Germany
C	No	No	PARIS is the capital of Germany
D	No	Yes	Munich is NOT the capital of Germany
E	No	No	John is the capital of Germany
F	Yes	Yes	John is not the capital of Germany



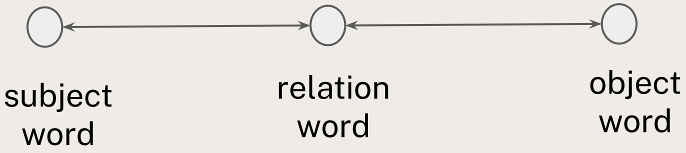
Will subject-object attention reduce in the presence of factually incorrect close due to negation?

Impact on Relation’s (negation focus) Attention

Construct three paired sentences as follows:

	Is Factually Correct (X=1/0)	Is it negated (Y=1/0)	Cloze
A	Yes	No	Munich is the capital of Germany
B	Yes	Yes	PARIS is NOT the capital of Germany
C	No	No	PARIS is the capital of Germany
D	No	Yes	Munich is NOT the capital of Germany
E	No	No	John is the capital of Germany
F	Yes	Yes	John is not the capital of Germany

Relation to Subject/Object Attention



Relation to {Subject/Object} Attention

Will Negation Cue change the Relation (Negation Focus)’s Attention on Subject-Object

Other contrasts based on Attention Score

Construct three paired sentences as follows:

	Is Factually Correct (X=1/0)	Is it negated (Y=1/0)	Cloze
A	Yes	No	Munich is the capital of Germany
B	Yes	Yes	PARIS is NOT the capital of Germany
C	No	No	PARIS is the capital of Germany
D	No	Yes	Munich is NOT the capital of Germany
E	No	No	John is the capital of Germany
F	Yes	Yes	John is not the capital of Germany

Contrast: Hypothesis

{A} -{D} < 0: PLMs can reason, not only recollect.

{A-B} -{{C-D} + {E-F}} < 0: PLMs can reason, not only recollect, after controlling

Several Others

{{X=1-X=0}|Y=0}: PLMs reason, under no-negation

{{X=1-X=0}|Y=1}: PLMs reason, under negation

{{X=1-X=0}|Y=0} = {{X=1-X=0}|Y=1} ~= 0

PLMs reasoning ability differs under negation

Recent work in this direction

Improving negation detection with negation-focused pre-training

Hung Thinh Truong[1] **Timothy Baldwin**[1,3] **Trevor Cohn**[1] **Karin Verspoor**[2]
[1]The University of Melbourne, [2]RMIT University, [3]MBZUAI
hungthinht@student.unimelb.edu.au, tb@ldwin.net,
trevor.cohn@unimelb.edu.au, karin.verspoor@rmit.edu.au

Proposed new **negation-focused pre-training** strategies to *better* incorporate negation information and *generalizability* into language models (on strong baseline models like **NegBERT**) :

1. Targeted Data Augmentation
2. Negation masking

Recent work in this direction

Negation focused data : ***No*** serious complications such as ***hypertension***, diabetes.

Added negation cue masked data : **[CUE]** serious complications such as **[MASK]**, diabetes.

Experimental setup include evaluating following methods :

- **NegBERT**
- **AugNB** : NegBERT **plus** pre-training on negation-focused data
- **CueNB** : NegBERT **plus** pre-training on negation focused data **and** the negation cue masking objective.

Task	Same-dataset results			Cross-dataset results		
	NegBERT	AugNB	CueNB	NegBERT	AugNB	CueNB
Cue Detection	90.55	+0.36	+1.34	69.61	+2.21	+ 3.31
Scope Resolution	90.56	+ 0.59	+1.62	73.41	+0.95	+ 1.72

Table 4: Aggregated results

Ongoing debate : Does attention offer Interpretability?

Attention is not Explanation

Sarthak Jain and Byron C. Wallace, NAACL-HLT (2019). [5]

Hypotheses:

➔ Attention weights should **correlate** with feature importance measures (e.g., gradient-based measures).

➔ **Alternative** (or counterfactual) **attention weight configurations** ought to yield corresponding **changes** in prediction (and if they do not then are *equally plausible* as explanations).

Ongoing debate : Does attention offer Interpretability?

Attention is not Explanation

Sarthak Jain and Byron C. Wallace, NAACL-HLT (2019). [5]

Conclusions:

➔ **Correlation** between **standard feature importance** and **attention weights** are weak

➔ Randomly **permuting** the attention weights **doesn't** change the output significantly

Ongoing debate : Does attention offer Interpretability?

Attention is not not Explanation

Sarah Wiegreffe and Yuval Pinter (2019). [8]

One month later...

Raises the issues:

- ➔ Explanation is **ambiguous**
- ➔ Correlation studies are **insufficient**
- ➔ Adversarial attention experiments had **little to no meaning**

Ongoing debate : Does attention offer Interpretability?

Attention is ~~not~~*
~~not not~~**
maybe explanation

* Sarthak Jain & Byron C. Wallace, 2019
** Sarah Wiegreffe & Yuval Pinter, 2019

➔ Point of emphasis: "**Attention** is not **explanation**" in the same way that "**correlation** is not **causation**"?

➔ What does "**explanation**" mean to you

Might provide *plausible* explanation which can be understood by a human even if it's not faithful to how the model works.

➔ What could attention be **measuring**?

It noisily predicts input components' overall importance to a model, it is by no means a fail-safe indicator. Use it as a **sanity check** and a **tool**!

Thank you for your attention!

References

1. Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 8, 34-48.
2. Kassner, N., & Schütze, H. (2019). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. arXiv preprint arXiv:1911.03343.
3. Hossain, M. M., Hamilton, K., Palmer, A., & Blanco, E. (2020, July). Predicting the focus of negation: Model and error analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8389-8401).
4. L. R. Horn. 1989. A natural history of negation. Chicago University Press, Chicago.
5. Pullum, G. K., & Huddleston, R. (2002). The Cambridge grammar of the English language
6. Morante, R., & Blanco, E. (2012). * sem 2012 shared task: Resolving the scope and focus of negation. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) (pp. 265-274)
7. Zhao, Y., & Bethard, S. (2020, July). How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4729-4747).
8. Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., & Xiong, C. (2020). Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. arXiv preprint arXiv:2003.04985.
9. Kevin Clark and Urvashi Khandelwal and Omer Levy and Christopher D. Manning. 2019. What Does {BERT} Look At? An Analysis of BERT's Attention. CoRR Journal
10. PB-FOC, Resolving the Scope and Focus of Negation, SEM 2012 Shared Task
11. F. Petroni, T. Rocktaschel, A. H. Miller, P. Lewis, A. Bakhtin, Y. Wu and S. Riedel. 2019. Language Models as Knowledge Bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019