Questions
Key Ideas

Negated LAMA: Birds cannot fly

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

Abstract

Pretrained language models have achieved remarkable improvements in a broad range of natural language processing tasks, including question answering (QA). To analyze pretrained language model performance on QA, we extend the LAMA (Petroni et al., 2019) evaluation framework by a *component that is focused on negation*. We find that pretrained language models are equally prone to generate facts ("birds can fly") and their negation ("birds cannot fly"). This casts doubt on the claim that pretrained language models have adequately learned factual knowledge.

1 Introduction

Pretrained language models like Transformer-XL (Dai et al., 2019), ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have emerged as universal tools that capture a diverse range of linguistic and factual knowledge.

Recently, Petroni et al. (2019) introduced LAMA (LAnguage Model Analysis) to investigate to what extent pretrained language models have the capacity to recall factual knowledge without the use of fine-tuning. The training objective of pretrained language models is to predict masked tokens in a sequence. With this "fill-in-the-blank" scheme, question answering tasks can be reformulated as cloze statements. For example, "Who developed the theory of relativity?" is reformulated as "The theory of relativity was developed by [MASK].". This setup allows for unsupervised open domain question answering. Petroni et al. (2019) find that, on this task, pretrained language models outperform supervised baselines using traditional knowledge bases with access to oracle knowledge.

This work analyzes the understanding of pretrained language models of factual and commonsense knowledge stored in negated statements. To this end, we introduce the *negated LAMA dataset*. We construct it by simply inserting negation elements (e.g., "not") in LAMA cloze statement (e.g., "The theory of relativity was not developed by [MASK]."). In our experiments, we query the pretrained language models with both original LAMA and negated LAMA statements and compare their predictions in terms of rank correlation and overlap of top predictions. We find that the predicted filler words often have high overlap. Thus, negating a cloze statement does not change the predictions in many cases – but of course it should as our example "birds can fly" vs. "birds cannot fly" shows. We identify and analyze a subset of cloze statements where predictions are different. We find that BERT handles negation best among pretrained language models, but it still fails badly on most negated statements.

2 Data

A cloze statement is generated from a subject-relation-object triple from a knowledge base and from a templatic statement for the relation that contains variables X and Y for subject and object (e.g, "X was born in Y"). We then substitute the subject for X and MASK for Y. The triples are chosen such that Y is always a single-token answer.

LAMA covers different sources: The Google-RE[1] set covers the three relations "place of birth", "date of birth" and "place of death". T-REx (ElSahar et al., 2018) consists of a subset of Wikidata triples covering 41 relations. ConceptNet (Li et al., 2016) combines 16 commonsense relationships between words and/or phrases. The underlying Open Mind Common Sense corpus provides matching statements to query the language model. SQuAD (Rajpurkar et al., 2016) is a standard question answering dataset. LAMA contains a subset of 305

[1] https://code.google.com/archive/p/relation-extraction-corpus/

Corpus	Relation	Statistics		LM									
		Facts	Rel	Txl		Eb		E5b		Bb		Bl	
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	Total	11458	16	91.1	32.0	96.8	63.5	96.2	53.5	89.9	34.9	88.6	31.3
SQuAD	Total	305	-	91.8	46.9	97.1	62.0	96.4	53.1	89.5	42.9	86.5	41.9

Table 1: Mean spearman rank correlation and the mean percentage of overlap in first ranked predictions between the original cloze template queries and the negated statement for Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl).

context-insensitive questions and provides manually reformulated cloze-style statements to query the model.

We created negated versions of Google-RE, T-REx and SQuAD by manually inserting a negation element in each template or statement. We did the same for a subset of ConceptNet that is easy to negate. We selected this subset by filtering for sentence length and extracting common queries.

3 Models

We use the source code provided by Petroni et al. (2019) and Wolf et al. (2019) [2] and evaluate using Transformer-XL large (Txl), ELMo original (Eb), ELMo 5.5B (E5B), BERT-base (Bb) and BERT-large (Bl).

4 Results

Table 1 compares the predictions of original LAMA and negated LAMA. As true answers of the negated statements are highly ambiguous, our measures are spearman rank correlation and overlap in rank 1 predictions between the original and negated dataset. Table 2 gives examples of BERT-large predictions.

We observe rank correlations of more than 0.85 in most cases and a high overlap in first ranked predictions like: “Birds can fly.” and “Birds cannot fly.”. BERT has slightly better results than the other models.

Our interpretation of the results is that BERT mostly did not learn the meaning of negation. The impressive results in QA suggest that pretrained language models are able to memorize aspects of specific facts; but, apparently, they ignore negation markers in many cases and rely on the co-occurrence of the subject with the original relation only. One reason for the poor performance we observe probably is that negated statements occur much less in training corpora than positive statements.

A key problem is that the LAMA setup does not allow to refrain from giving an answer. Generally, prediction probabilities drop in the negated statements, which would suggest the existence of a threshold to filter answers. But a closer look at the probabilities of correct and incorrect predictions shows that they fall into the same range. No common threshold can be found.

Given that negation has little effect on most queries, it is interesting to look at the small number of queries where pretrained language models make correct predictions, i.e., they solve the cloze task as a human subject would do. We give two examples of such patterns. The pattern “X did not die in Y” always results in the generic top ranked predictions: “battle”, “office”, “prison” whereas the original pattern is likely to rank cities first. This seems appropriate since a statement of the form, say, “X did not die in New York” is rare in text corpora, but statements characterizing the situation in which the death occurred (“he did not die in prison”) sound more natural. For the template “X was born in Y”, cities are predicted. In contrast, for “X was not born in Y”, countries are predicted. Both times it refers to a more specific statement, more likely to occur in the training corpus. People would refer more often to a person being born in a city and not born in a country, giving you in both cases more precise information.

5 Related Work

Pretrained embeddings have pushed baselines on a variety of question answering datasets (Kwiatkowski et al., 2019; Alberti et al., 2019). Generally, the pretrained models are fine-tuned to

[2] https://github.com/facebookresearch/LAMA

	Query	Answer	Generation
(A)	Marvano was born in [MASK].	Belgium	Rome (-2.0), Milan (-2.2), Italy (-2.5)
	Marvano was not born in [MASK].		Italy (-1.6), Rome (-3.2), Chicago (-3.2)
	Odofredus died in [MASK].	Bologna	Rome (-0.8), office (-3.0), c (-3.3)
	Odofredus did not die in [MASK].		battle (-1,5), office (-1.7), captivity (-3.2)
(B)	The capital of Azerbaijan is [MASK].	Baku	Baku (-0.0), Yerevan (-6.9), Azerbaijan (-7.5)
	The capital of Azerbaijan is not [MASK].		included (-1.9), mentioned (-2.2), shown (-2.2)
	Havana is the capital of [MASK] .	Cuba	Cuba (-0.0), Havana (-5.8), Spain (-7.9)
	Havana is not the capital of [MASK] .		Cuba (-0.0), Havana (-5.5), Nicaragua (-5.8)
(C)	Birds can [MASK].	fly	fly (-0.5), sing (-2.3), talk (-2.8)
	Birds cannot [MASK].		fly (-0.3), sing (-3.6), speak (-4.1)
	Cat requires [MASK].	meat	feeding (-1.2), feed (-1.5), shelter (-3.1)
	Cat does not require [MASK].		food (-3.0), feeding (-3.2), breeding (-4.1)
(D)	The theory of relativity was developed by [MASK].	Einstein	Einstein (-0.3), Newton (-1.6), Maxwell (-5.2)
	The theory of relativity was not developed by [MASK].		Einstein (-0.6), Newton (-2.0), Galileo (-4.0)
	Chloroplasts need [MASK] to replicate.	light	time (-2.3), enzymes (-2.4), energy (-2.4)
	Chloroplasts do not need [MASK] to replicate.		enzymes (-1.7), oxygen (-1.8), water (-2.4)

Table 2: Examples of generation for BERT-large for **(A)** Google-RE, **(B)** T-REx, **(C)** ConceptNet, **(D)** SQuAD. The last column reports the top three tokens generated together with the associated log probability (in brackets).

the specific task (Liu et al., 2019; Devlin et al., 2019) but recent work has applied the models without the fine-tuning step (Radford et al., 2019; Petroni et al., 2019).

There is a wide range of literature analyzing linguistic knowledge stored in pretrained embeddings (Jumelet and Hupkes, 2018; Gulordava et al., 2018; Giulianelli et al., 2018; McCoy et al., 2019; Dasgupta et al., 2018; Martin and Linzen, 2018; Warstadt and Bowman, 2019; Kann et al., 2019).

Concerning negation the following papers are of interest: Warstadt et al. (2019) analyze the grammatical knowledge captured by BERT. In a case study, they test for correct licensing environments for negative polarity items. They study a set of classifiers distinguishing between grammatically correct and incorrect sentences. We take a different approach by focusing on factual knowledge stored in negated statements. Grammatically correct statements can still be factually false (e.g., "General relativity, Newton develop").

Kim et al. (2019) investigate the understanding of function words – among them negation particles – using an entailment- and classification-based approach. They analyze the ability of different model architectures and training objectives to capture knowledge of single sentences. The models are fine-tuned to the task of interest. We on the other hand question to what extend factual knowledge present in negated statements is indirectly acquired during pretraining.

Ettinger (2019) defines three psycholinguistic diagnostics for language models and applies them in a case study to BERT. Negation is examined using a dataset of 72 simple sentences querying for category membership. A supplementary dataset of 16 sentences queries again for the "to be" relation only but including more natural sentence structure. Our work covers 51,329 negated statements covering a wide range of topics and relations. In the SQuAD based dataset we cover more natural language in terms of context and relation. In contrast to Ettinger (2019), we do not see a reliable preference of the true completions to false in the more natural negated statements.

Ribeiro et al. (2018) test for comprehension of minimally modified statements in an adversarial setup while trying to keep the overall semantics the same. We try to maximize the change in semantics and invert meaning.

6 Conclusion

We show that pretrained language models have problems handling negation. Output predictions for the original LAMA query and the negated statement are highly correlated.

Even though this elegant approach of querying a language model without fine-tuning allows for truly open domain question answering, promoting an answer no matter what is not always the better solution. Refraining from giving an answer can be more appropriate, making knowledge graphs currently a more reliable choice for question answering.

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A bert baseline for the natural questions. *ArXiv*, abs/1901.08634.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*.

Allyson Ettinger. 2019. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1445–1455, Berlin, Germany. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rebecca Martin and Tal Linzen. 2018. Targeted syntactic evaluation of language models.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

F. Petroni, T. Rocktäschel, A. H. Miller, P. Lewis, A. Bakhtin, Y. Wu, and S. Riedel. 2019. Language models as knowledge bases? In *To Appear in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2019. Grammatical analysis of pretrained sentence encoders with acceptability judgments. *CoRR*, abs/1901.03438.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jereti, and Samuel R. Bowman. 2019. Investigating bert's knowledge of language: Five analysis methods with npis.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Questions and (Sometimes) Answers

1. Cloze statement: A **cloze test** (also **cloze deletion test**) is an exercise, test, or assessment consisting of a portion of language with certain items, words, or signs removed (cloze text), where the participant is asked to replace the missing language item. Source: Wikipedia

2. Mean spearman rank correlation: In statistics, **Spearman's rank correlation coefficient** or **Spearman's rho**, named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as r_s, is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

   The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

   Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd, etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of −1) rank between the two variables.

   Spearman's coefficient is appropriate for both continuous and discrete ordinal variables.[1][2] Source: Wikipedia