Vol.2021-NL-249 No.3
2021/7/27

事前学習済み言語モデルにおける否定の理解能力の調査

田代 真生[1,a)]　上垣外 英剛[1]　船越 孝太郎[1]　高村 大也[2]　奥村 学[1]

1. はじめに

近年，事前学習済み言語モデルは自然言語処理の様々なタスクにおいて性能の向上に大きく貢献している．事前学習済み言語モデルには，事実知識を大量のラベルなしコーパスから獲得可能であり，かつそれらを柔軟に取り出し可能であるという利点が存在し，その適用例としては，ファクトチェックへの利用 [8] や常識推論タスクでの利用 [20] が挙げられる．

一方で事前学習済み言語モデルからの事実知識の取り出しにおける課題を指摘する研究も存在しており，その一つに否定の理解能力のなさを指摘したものが挙げられる [6]．事前学習済み言語モデルにおける否定の理解能力に関しては依然として明確な結論はなく疑問が残っている [15]．Kassner ら [6] や Ettinger[3] の研究では事前学習済み言語モデルが否定を考慮せずにマスクされたトークンを予測している可能性が示唆されているのに対し，Talmor ら [19] の研究では否定語の予測が可能であることから事前学習済み言語モデルが否定を理解している可能性が示された．

そこで本研究では，これらの実験条件の違いを考慮し，否定が事前学習済み言語モデルの出力に変化を与える条件を純粋な否定考慮の設定 (3 節) と事実知識が絡んだ否定考慮の設定 (4 節) で探ることで，事前学習済み言語モデルにおける否定の理解において，1. 事実知識の想起が絡むか，2. モデルのパラメータ量や学習データ量 の条件が影響を与えるかについて調査した．これらの観点から検証を行うことにより，本研究は既存の相反する研究結果に対して一貫した説明を提供しており，否定を理解可能なモデルの将来的な作成を支援すると考えられる．

2. 関連研究

事前学習済み言語モデルは大量のラベルなしコーパスを用いて言語モデルを訓練することで言語的，意味的な知識を多く獲得しており，それによって言語理解の総合的な能力を試す GLUE ベンチマーク [21] や文章読解能力を試す SQuAD ベンチマーク [12] などで性能の向上に貢献している [2]．事前学習済み言語モデルは基本的に大量のラベルなしコーパスで言語モデルを学習する事前学習と，目的のタスクのラベル付きコーパスでタスク用のモデルを学習するファインチューニングの二段階の学習によって利用される [2]．しかし，事前学習済み言語モデルをファインチューニングせずに，入力を工夫することで目的の出力を得るプロンプティング (Prompting) という手法も提案されており，Jiang ら [5] や Shin ら [16] は自動的に高性能な言語的プロンプト (prompt) を発見する手法を提案している．

2.1 事前学習済み言語モデル内の知識の利用

Petroni ら [10] は，事前学習済み言語モデルが事実知識や常識を必要とする穴埋め式質問応答タスクをファインチューニングなしで解けることから，事実知識や常識を事前学習において獲得していることを明らかにした．事前学習済み言語モデル内の知識は，大量のラベルなしコーパスから獲得可能であり，かつ柔軟な取り出しが可能であるという利点があり，ファクトチェック [8] や常識推論 [20]，常識データセットの構築 [1] など様々なタスクで利用が進んでいる．

2.2 事前学習済み言語モデルにおける否定の理解

Kassner ら [6] や Ettinger[3]，Ribeiro ら [13] は自然言語処理において重要な否定の理解について，近年の高い自然言語処理能力を持つモデルが否定を理解していないという課題を指摘している．Kassner らや Ettinger は穴埋め式の質問応答タスクにおいて，問題に否定語を挿入した時としていない時で事前学習済み言語モデルにおける出力が変化しない事象を発見した．また，Ribeiro らは否定語を含む文章においてテキスト分類タスク，自然言語推論タスク，

[1] 東京工業大学
[2] 産業技術総合研究所
[a)] masaki@lr.pi.titech.ac.jp

読解タスクの性能が落ちる事象を発見した．

一方で，事前学習済み言語モデルにおいて否定を理解した挙動をする例も発見されており，Talmor ら [19] は Masked Language Model (MLM) が not の有無を正しく予測するタスクを一定の性能で解けることを発見した．また，Warstadt ら [22] は NPIs(否定極性項目) と呼ばれる否定と共起する必要のある項目 (ever など) に関わる文書の文法的正しさを一定の性能で判定できることを示し，事前学習済み言語モデルにおける否定に関する知識の存在を示した．本研究はこのような既存研究において存在する実験条件の差異を考慮し，事前学習済み言語モデルにおける否定の理解能力に影響を与える条件を探っているという点で既存研究との相違点がある．

3. 事実知識を必要としない問題における否定の考慮

本実験では事実知識を必要としない問題における否定の有無が出力に与える影響を調べるために，**表 1** のようなマスク部の予測に事実知識がいらない問題を Wikipedia を用いて用意し実験した．ここで事実知識とは Storks らに倣い [18] 世界に関する明示されやすい知識 (DBpedia など) を指しており，語彙の言語的な関わりを示す語彙的な知識 (WordNet など) と区別している．[*1]本実験を Kassner らや Ettinger の実験と比較すると，Kassner らや Ettinger の実験が問題を解くために否定の考慮に加えてモデルに事実知識を要求するものであったのに対し，本実験ではそれを要しない設定にしてあるため，純粋に事前学習済み言語モデルが否定の考慮を行えるのかを調べているという違いがある．具体的には，Kassner らの実験では ‘Barack Obama was born in [MASK].’ と ‘Barack Obama was not born in [MASK].’ の予測の違いを見ているが，この例にて否定を適切に考慮するためにはモデル内で事実知識の想起と否定の考慮を組み合わせる必要があり，純粋な否定の理解能力を見ることができない．また，Talmor ら [19] の実験と比較すると ‘not’ を予測するタスクではないため，モデルの予測に対して ‘not’ が文脈としてどのような影響を及ぼすのかを調べられるという違いがある．

3.1 実験設定

データセット

この実験は Talmor らの実験を参考にしており，Talmor らの実験で利用していた ‘It was [MASK] fast, it was really slow.’ のような例を，‘not’ を含む ‘It is not fast. In short, it is [MASK].’ と not を含まない ‘It is fast. In short, it is [MASK].’ に変形し，マスク箇所に対する出力の違いを見ることによって ‘not’ の有無が出力に与える影響を調査した．ここでテンプレート文の作成に関しては Jiang ら [5] の議論などがあるが，本実験では人手でテンプレート文を作成した．その際に後の文 (‘it is not [MASK].’) が前の文 (‘it is fast.’) と同じ内容になるような予測をするために言語的な知識に基づいて ‘In short,’ という語句を挿入した．また，Talmor ら [19] の実験では，ConceptNet[17]，WordNet[4]，Google Books Corpus から，類義語，対義語対を出現頻度によってフィルタリングし抽出していたが，Kassner らが用いていた NegatedLAMA においては対義語がないもの，出現頻度が少ないものを対象にしていたため，本研究では対義語を持たない単語 (**表 2** を参照) や低頻度語に関しても対象にし，Wikipedia から無作為に単語を抽出した．

評価方法

変化の評価の指標としては正解率と，Kassner らの実験を参考にトップ 1 予測の一致率，スピアマンの順位相関を利用している．正解率とは，出力が否定によってどのように変化したかを調べるために，トップ 1 の予測が正解か不正解かどちらでもないかをラベル付けし．正解の割合を調べたものである．ここで，否定を含む入力に対しては抽出した単語の対義語，同位語を正解にし，類義語を不正解にした．一方で，否定を含まない入力に対しては類義語を正解にし，対義語・同位語を不正解にした．また，出力が正解にも不正解にも含まれない場合はどちらでもないとラベル付けした．また，トップ 1 予測の一致率と，スピアマンの順位相関では否定を挿入する前と後でのトップ 1 の予測が変化したか，予測対での順位相関がどうかを見ており，どちらも低い方が否定を考慮して予測を変化させたと捉えられる．ただし，上記の指標をそのまま利用すると，否定の効果が不明な出力の変化 (fast → good など) を多く検知してしまうという課題を考慮し，先述の正解と不正解からなる候補の内のトップ 1 の予測の変化と順位相関を調査することで，調査したい対象内の出力の変化のみを見られるようにした

実験に利用する事前学習済み言語モデルとしては Huggingface の Transformers ライブラリ [*2]にて公開されている事前学習済み言語モデルを利用した．モデルの種類としては様々な事前学習方法の影響を調べるために BERT[2]，RoBERTa[9]，ALBERT[7]，GPT-2[11] を調査した．また，モデルサイズの影響を調べるために，それぞれのモデルについて公開されている全てのサイズで調査した．さらに Warstadt ら [23] の調査から，事前学習データ量が否定の理解に影響を与えるという仮説を立て，Warstadt らが公開している様々なデータ量で事前学習された RoBERTa についても調査を行った．

[*1] 表 1 において Earth と Mars 等が排他的であるという知識が必要となってくる例が存在するが，そのような知識に関しては語彙的な知識の一部であると考える．

[*2] https://huggingface.co/transformers/

表 1 否定の考慮に事実知識を必要としない問題の例

	入力	正解
対義語を持つ単語	It is fast. In short, it is [MASK].	fast
対義語を持つ単語の否定	It is fast. In short, it is not [MASK].	slow
対義語を持たない単語	It is Earth. In short, it is [MASK].	Earth, world, globe
対義語を持たない単語の否定	It is Earth. In short, it is not [MASK].	Mars, Mercury, Venus

表 2 事実知識を必要としない問題における否定の考慮の度合い．True, False, None はトップ予測が正解率，不正解率，どちらでもない割合を示している．ρ はスピアマンの順位相関を，% はトップ予測の一致の割合を示している．

setting	False	True	None	ρ	%
ALBERT-B	0.904	0.017	0.079	0.936	0.966
ALBERT-L	0.827	0.014	0.159	0.822	0.900
ALBERT-XL	0.742	0.044	0.214	0.736	0.798
ALBERT-XXL	0.173	0.244	0.583	0.359	0.496
BERT-B	0.513	0.074	0.413	0.759	0.717
BERT-L-WWM	0.502	0.100	0.398	0.614	0.758
BERT-L	0.502	0.129	0.369	0.674	0.753
GPT-2	0.240	0.046	0.713	0.748	0.858
GPT-2-L	0.099	0.070	0.831	0.597	0.789
GPT-2-XL	0.033	0.115	0.852	0.481	0.720
RoBERTa-B-1M	0.145	0.000	0.855	0.809	0.797
RoBERTa-B-10M	0.720	0.010	0.270	0.919	0.957
RoBERTa-B-100M	0.745	0.069	0.186	0.851	0.852
RoBERTa-B-1B	0.669	0.101	0.230	0.822	0.857
RoBERTa-B	0.430	0.139	0.432	0.607	0.737
distilRoBERTa-B	0.478	0.130	0.391	0.746	0.767
RoBERTa-L	0.205	0.253	0.542	0.381	0.557

表 3 RoBERTa-large における否定対象による否定の考慮度の違い

品詞	対義語を持つ	accuracy	ρ	%
ADJ	True	0.368	-0.349	0.267
NOUN	False	0.194	0.850	0.717
	True	0.232	0.029	0.496
VERB	True	0.320	0.083	0.494

3.2 実験結果

表 2 は否定語を含む入力における出力の傾向を示している．否定を含む入力における正解率を見てみると，最も性能が良かった RoBERTa-large モデルでは全体の 25.3%の例で正しく否定を考慮できていることがわかる．一方で否定を含まない設定の際にはそのような正解の出力 (対義語や同位語) は 2%であった．この結果を見ると，否定の挿入によって対義語や同位語の割合が，大きく上昇していることから，特定のモデルにおいては否定語の存在によって適切な出力の反転が行われていることが理解できる．順位相関や予測の一致率について見てみると，否定を含む入力における正解率と負の相関があり，正解率と同様にモデルにおける否定の考慮の度合いを読み取ることができる．ただし，RoBERTa-B-1M や RoBERTa-B-10M では正解率が低いのにも関わらず，低い順位相関が記録されており，単純な入力の変化に対するモデルの頑健性の低さと，否定を考慮している度合いを順位相関や予測の一致度のみで区別することが難しい事例が存在することが分かる．

モデルの影響について考えると BERT, ALBERT, RoBERTa, GPT-2 においてモデルサイズが大きくなるにつれて正解率が向上していることから，これらのモデルではモデルサイズを上昇させることにより否定の理解が進んでいることがわかる．これは Devlin ら [2] や Liu ら [9] の実験において分かったモデルサイズが幅広い自然言語処理能力に与える影響と一致している．また，学習データ量の影響について考えると，RoBERTa において学習データ量を増やすにつれて正解率が向上しており，否定の理解が進んでいることがわかる．これは Warstadt らの主張である事前学習データ量を増やすことよって，事前学習済み言語モデルが言語的な情報を表層的な情報より選好するという傾向と一致する．

表 3 は比較的，否定の考慮ができた RoBERTa-large における否定対象による否定の考慮度の違いを示している．本実験では否定に影響を与えうる否定対象の特徴として対象の品詞と対義語の有無を取り上げ調査を行った．品詞別の正解率を見てみると形容詞においては 37%近く正しい出力を得られたのに対し，名詞に関しては 23%程度にとどまった．また，順位相関や予測の一致率を見ても形容詞と比較して動詞や名詞の否定考慮が出来ていないことがわかる．対義語の有無が与える影響について見てみると，対義語を持たない対象において予測の一致率が 71.7%程度であるのに対し，持つ対象においては 50%程度と低く対義語の有無が否定の考慮に影響を与えることが示唆された．

4. 事実知識の想起が否定の考慮に与える影響

本実験では 3 節の実験での純粋な否定の考慮を発展させ，事実知識の想起が必要なタスクにおける否定の考慮について，Kassner らの作成した Negated LAMA データセットを拡張したものを利用して調査した．Negated LAMA データセットはサブジェクトエンティティ (subject entity)，リレーション (relation)，オブジェクトエンティティ (object entity) の三つ組とそれに対応する穴埋め問題のテンプレートの集合によって構成されている．本実験は Negated LAMA データセットを用いて，事実知識に関するヒントとしての文脈の有無が否定の考慮に与える影響を調べることによって，事実知識が絡んだ否定の考慮という現実世界の設定での否定の考慮を調べることを目的とする．

4.1 実験設定

データセット

拡張の目的はオリジナルの Negated LAMA データセットには含まれていない対義語を持つ単語における否定を調査することであり，そのために wikidata から SPARQL を用いて，有名人の性別に関する知識と国の位置がどちらの半球にあるかについての知識を抽出して拡張した．本実験では**表 4** の (a)(b) のようなオリジナルの Negated LAMA データセットにおける否定の考慮に加え，事実知識を文脈として入力に接続することで，解くのに事実知識が不要な入力 (d)(e) を用意した．また，否定を含むことによって生まれた出力の変化と単純に入力に単語が挿入されたことによる出力を区別するために，否定語をほとんど変化させない単語である ‘really’ に置換した例 (c)(f) を用意した．

評価方法

本実験でも 3 節の実験と同様に，変化の評価の指標として候補内のトップ 1 の予測の変化と，スピアマンの順位相関を利用している．候補の作成に関しては，同じリレーション (relation) における正解のエンティティ (entity) が同位語になりやすいという特徴を生かし，同じリレーションの正解のエンティティ群を候補と考えた．さらに，3 節の実験で明らかになったようにモデルの否定理解の度合いとモデルの出力変化のしやすさを上記の指標のみからは見分けることが困難であるため，本実験では否定語を ‘really’ に置換した例との差分を否定考慮の効果として追加で確認した．調査対象のモデルとしては 3 節で選んだモデルと同じモデルを対象とした．

4.2 実験結果

表 6 は否定考慮タスクにおける事実知識の有無が予測の一致度や否定考慮度に与える影響を調べたものである．ここでスピアマンの順位相関は Negated LAMA データセット全体に対して調査し，各リレーション (relation) ごとに平均を取った上で全体の平均を取っている．また，否定考慮度は ‘really’ を挿入した例における予測の一致率から否定を挿入した例における予測の一致率を差し引いて計算したものであり，高い方が否定を考慮できていることを示している．表 6 の予測の一致度を見てみると，入力内に事実知識が含まれている方が全体的に予測が一致するという傾向が見られた．一方で否定考慮度を見てみると，ほとんどのモデルで事実知識なしの条件では否定の考慮ができていないのに対し，3 節の実験と同様に RoBERTa-large や ALBERT-xxlarge といったパラメータ数の大きいモデルは事実知識ありの条件で否定の考慮が出来ている傾向が読み取れる．これらの結果は，入力に事実知識が含まれない条件では含まれる条件よりも否定の挿入によって出力が変化したように見えるが，実際は入力の単純な追加による出力の変化であり否定の考慮という観点では知識の想起がない方が考慮しやすいということを示している．この事実知識の想起と否定の考慮を両方行うことが難しいという傾向は質問応答においてマルチホップ推論が難しいという既存研究結果と一致しており [14]，既存研究における否定理解の調査では否定理解の難しさとマルチホップ推論の難しさを区別できていないことを示唆している．

また，実際に否定によって出力が変化しにくい文章と変化しやすい文章の例を**表 5** にまとめた．表 5 の上部は否定の挿入によって出力が最も変化しにくかった 5 つのリレーションの文章であり，下部は最も変化しやすかった 5 つのリレーションの文章である．変化しやすい文章について見てみると，3 節の実験で見たような対義語を持つ単語や国に関する単語の否定が考慮できていることがわかる．一方で，考慮できていない対象について見てみると，‘(not born [MASK])’ や ‘not a subclass of [MASK]’ など，頻出しないテンプレートを利用している傾向があることがわかる．これらより，否定の考慮については現状の事前学習済み言語モデルでは対義語を持つ単語など限られた対象については可能であり，今後はその対象やその対象を広げるための方法についてを調べる必要があると考えられる．

5. おわりに

本研究では，既存研究において事前学習済み言語モデルにおける否定の理解について相対する結果が見られている点に着目し，否定の有無が事前学習済み言語モデルを用いてマスク箇所のトークンを予測する際の出力に変化を与える条件を調べることで，事前学習済み言語モデルが否定を理解し得る条件についてを調査した．本研究では，新たに事前知識を含む文脈を追加したり，幅広いモデルを用いて否定の理解を調べることによって，事実知識の想起の必要性とモデルのパラメータ数や事前学習データ量が否定の考慮に影響を与えることを確認した．さらに，否定の理解について既存の評価指標では捉え切れていない点についても紹介し，新たな指標を利用して否定の理解についてを調査した．本研究の結果は既存の相対する事前学習済み言語モデル否定理解の研究結果に対し一貫した説明を提供しており，事前学習済み言語モデルにおける否定の考慮についてより具体的な理解を進めるものとなっている．

参考文献

[1] Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A. and Choi, Y.: COMET: Commonsense Transformers for Automatic Knowledge Graph Construction, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Association for Computational Linguistics, pp. 4762–4779 (online), DOI: 10.18653/v1/P19-1470 (2019).

[2] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019*

表 4 否定の考慮に事実知識を必要とする問題の例

入力タイプ	事実知識が必要	単語の挿入	否定を含む	入力例	正解例
(a)	✓	–	–	Barack Obama was born in [MASK].	Hawaii
(b)	✓	✓	✓	Barack Obama was not born in [MASK].	Tokyo
(c)	✓	✓	–	Barack Obama was really born in [MASK].	Hawaii
(d)	–	–	–	Barack Obama was born in Hawaii. In short, Barack Obama was born in [MASK].	Hawaii
(e)	–	✓	✓	Barack Obama was born in Hawaii. In short, Barack Obama was not born in [MASK].	Tokyo
(f)	–	✓	–	Barack Obama was born in Hawaii. In short, Barack Obama was really born in [MASK].	Hawaii

表 5 否定の考慮が出来る問題と出来ない問題の例

X	Y	否定なし予測	否定あり予測	テンプレート
Peter F. Martin	1941	1941	1941	[X] (not born [MASK]).
Buddy Holly	Brunswick	Brunswick	Brunswick	[X] is not represented by music label [Y].
Howard Florey	London	London	London	[X] did not use to work in [Y].
lenticular galaxy	galaxy	galaxy	galaxy	[X] is not a subclass of [Y].
South Asia	Asia	Asia	Asia	[X] is not part of [Y].
The Scarlet Flower	Russian	Russian	English	The original language of [X] is not [Y].
Sorengo	Italian	Italian	English	The official language of [X] is not [Y].
Iginio Ugo Tarchetti	Italian	Italian	English	[X] did not use to communicate in [Y].
Woodrow Wilson	male	male	female	Considering gender, [X] is not [Y].
Japan	northern	northern	southern	Considering location, [X] is not located in the [Y] hemisphere.

表 6 事実知識を必要としない問題における否定の考慮の度合い

	予測の一致度		否定考慮度	
	知識あり	知識なし	知識あり	知識なし
ALBERT-B	0.941	0.525	0.017	0.036
ALBERT-L	0.952	0.649	-0.007	-0.046
ALBERT-XL	0.964	0.564	0.017	-0.010
ALBERT-XXL	0.757	0.556	0.215	0.031
BERT-B	0.794	0.637	-0.004	-0.011
BERT-L-WWM	0.902	0.629	0.082	0.018
BERT-L	0.898	0.628	0.027	0.019
GPT-2	0.909	0.737	0.011	-0.021
GPT-2-L	0.896	0.770	0.014	-0.014
GPT-2-Xl	0.817	0.773	0.083	0.009
RoBERTa-B-1M	0.523	0.471	-0.027	-0.054
RoBERTa-B-10M	0.937	0.480	0.019	0.035
RoBERTa-B-100M	0.904	0.524	-0.000	0.013
RoBERTa-B-1B	0.760	0.617	-0.002	0.036
RoBERTa-B	0.959	0.625	0.030	-0.049
distilRoBERTa-B	0.945	0.621	-0.002	-0.041
RoBERTa-L	0.613	0.591	0.376	-0.001

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, Association for Computational Linguistics, pp. 4171–4186 (online), DOI: 10.18653/v1/N19-1423 (2019).

[3] Ettinger, A.: What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 34–48 (online), DOI: 10.1162/tacl_a_00298 (2020).

[4] Fellbaum, C.(ed.): *WordNet: an electronic lexical database*, MIT Press (1998).

[5] Jiang, Z., Xu, F. F., Araki, J. and Neubig, G.: How Can We Know What Language Models Know?, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 423–438 (online), DOI: 10.1162/tacl_a_00324 (2020).

[6] Kassner, N. and Schütze, H.: Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics, pp. 7811–7818 (online), DOI: 10.18653/v1/2020.acl-main.698 (2020).

[7] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=H1eA7AEtvS⟩ (2020).

[8] Lee, N., Li, B. Z., Wang, S., Yih, W.-t., Ma, H. and Khabsa, M.: Language Models as Fact Checkers?, *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, Online, Association for Computational Linguistics, pp. 36–41 (online), DOI: 10.18653/v1/2020.fever-1.5 (2020).

[9] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019).

[10] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y. and Miller, A.: Language Models as Knowledge Bases?, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, Association for Computational Linguistics, pp. 2463–2473 (online), DOI: 10.18653/v1/D19-1250 (2019).

[11] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I.: Language Models are Unsupervised Multitask Learners (2019).

[12] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Association for Computational Linguistics, pp. 2383–2392 (online), DOI: 10.18653/v1/D16-1264 (2016).

[13] Ribeiro, M. T., Wu, T., Guestrin, C. and Singh, S.: Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics, pp. 4902–4912 (online), DOI: 10.18653/v1/2020.acl-main.442 (2020).

[14] Richardson, K. and Sabharwal, A.: What Does My QA Model Know? Devising Controlled Probes Using Expert Knowledge, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 572–588 (online), DOI: 10.1162/tacl_a_00331 (2020).

[15] Rogers, A., Kovaleva, O. and Rumshisky, A.: A Primer in BERTology: What We Know About How BERT Works, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 842–866 (online), DOI: 10.1162/tacl_a_00349 (2020).

[16] Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E. and Singh, S.: AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Association for Computational Linguistics, pp. 4222–4235 (online), DOI: 10.18653/v1/2020.emnlp-main.346 (2020).

[17] Speer, R., Chin, J. and Havasi, C.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, AAAI Press, p. 4444–4451 (2017).

[18] Storks, S., Gao, Q. and Chai, J. Y.: Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches (2020).

[19] Talmor, A., Elazar, Y., Goldberg, Y. and Berant, J.: oLMpics-On What Language Model Pre-training Captures, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 743–758 (online), DOI: 10.1162/tacl_a_00342 (2020).

[20] Tamborrino, A., Pellicanò, N., Pannier, B., Voitot, P. and Naudin, L.: Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Association for Computational Linguistics, pp. 3878–3887 (online), DOI: 10.18653/v1/2020.acl-main.357 (2020).

[21] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, Association for Computational Linguistics, pp. 353–355 (online), DOI: 10.18653/v1/W18-5446 (2018).

[22] Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., Wang, S.-F., Phang, J., Mohananey, A., Htut, P. M., Jeretic, P. and Bowman, S. R.: Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Association for Computational Linguistics, pp. 2877–2887 (online), DOI: 10.18653/v1/D19-1286 (2019).

[23] Warstadt, A., Zhang, Y., Li, X., Liu, H. and Bowman, S. R.: Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, Association for Computational Linguistics, pp. 217–235 (online), DOI: 10.18653/v1/2020.emnlp-main.16 (2020).