2019年度　卒 業 論 文

評価データの言い換えに伴う
汎用言語モデルの頑健性の検証

2020年1月22日

宮森研究室
(学生証番号: 644943)

西浦　大介

京都産業大学コンピュータ理工学部

概 要

近年、自然言語処理において BERT という汎用言語モデルが注目を集めている。このモデルは、質問応答やテキスト分類、テキスト要約などの様々なタスクにおいて高い精度を達成したモデルである。このモデルの登場以降、BERT を派生したモデルが次々と提案されている。その中でも、新たな自然言語モデルの RoBERTa は、BERT の学習方法を改善することで、短い学習時間で高い精度を実現させた最新の言語モデルである。RoBERTa は、様々な自然言語タスクのデータセットに対する精度を調査するに留まっており、敵対的なデータに対するモデルの性能は検証されていない。そこで本論文では、RoBERTa に対して、評価データセットの内容を意図的に変更し、敵対的なデータを作成することでモデルの頑健性を検証する。まず、これまで RoBERTa で検証されてきた各データセットにおいて、モデルが正答したデータおよび誤答したデータを取得する（第 1 段階）。その後、正答データおよび誤答データのそれぞれの回答の特徴を元に、難易度の異なる 2 種類の敵対的なデータを作成する（第 2 段階）。実験より、第 1 段階において正答したデータセット（つまり、正答率が 100%だったデータセット）から作成した、敵対的なデータセットでは、最低でも 8 割の正答率を保つことを確認した。一方、第 1 段階で誤答したデータセット（つまり、正答率が 0%だったデータセット）から作成した、敵対的なデータセットでは、最高で 5 割の正答率を達成することが分かった。第 1 段階で誤答したデータのテキストの意味を変えずに、文法の変更や、主語と目的語の入れ替えをすることで作成した新たなデータを、モデルが正答できたことから、RoBERTa が、必ずしも汎用的に自然言語を扱えているとはいえないことが考えられる。

目 次

第1章　はじめに

近年、ある機械読解モデルの入力データに対して、意図的に変更を加えることで「敵対的サンプル」を作成し、モデルの性能を検証する研究 [?][?] が行われている。敵対的サンプルとはモデルが正しく処理できるサンプルに対して誤答させるように変更を加えたサンプルのことである。例えば、下図 1.1 のように、モデルが「パンダ」と認識した画像にテナガザルのノイズを加えた敵対的サンプルを入力データとして機械に読み取らせると、画像はパンダなのにモデルが「テナガザル」と回答してしまう。

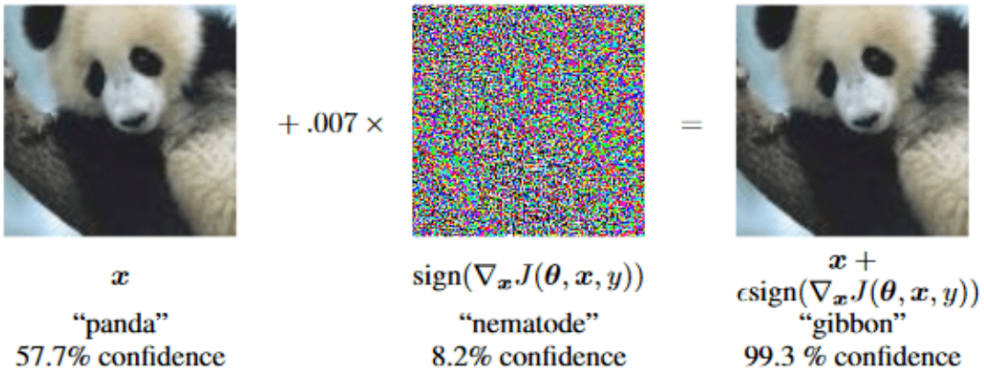

図 1.1: 敵対的なデータの具体例（[?] の文献より引用）

これに対して、近年の研究において、自然言語処理という「テキスト」を入力データとして機械に読み込ませる研究が行われている。この研究は、データセットに対する精度を調査するに留まっており、実際には敵対的なデータに対してモデルの性能を検証することはできていない。本稿では、質問応答用のデータセットの内容を意図的に変更し、敵対的なデータを作成することで、機械読解モデルの性能を検証する。本稿の構成は以下の通りである。2 章では関連研究について述べる。 3 章では、データセットの内容を意図的に変更し、敵対的なデータを作成することで、モデルの頑健性を検証する。 4 章では実験結果を示す。 5 章では実験結果を踏まえた考察を述べる。最後に、 6 章で全体のまとめと今後の課題について述べる。

第2章　関連研究

近年、自然言語処理において、Google の研究チームが開発した BERT[?] という汎用言語モデルが注目を集めている。このモデルは、質問応答やテキスト分類などの様々なタスクにおいて高い精度を達成した [?][?]。仕組みとしては、Transformer[?](図 2.1) という、入力層の各単語に双方向の注意機構を向けた技術を採用する。Transformer は、RNN や CNN といった従来型の深層ネットワークを使わない手法であり、これによって長くて複雑な文章の依存関係を捉えることができる。

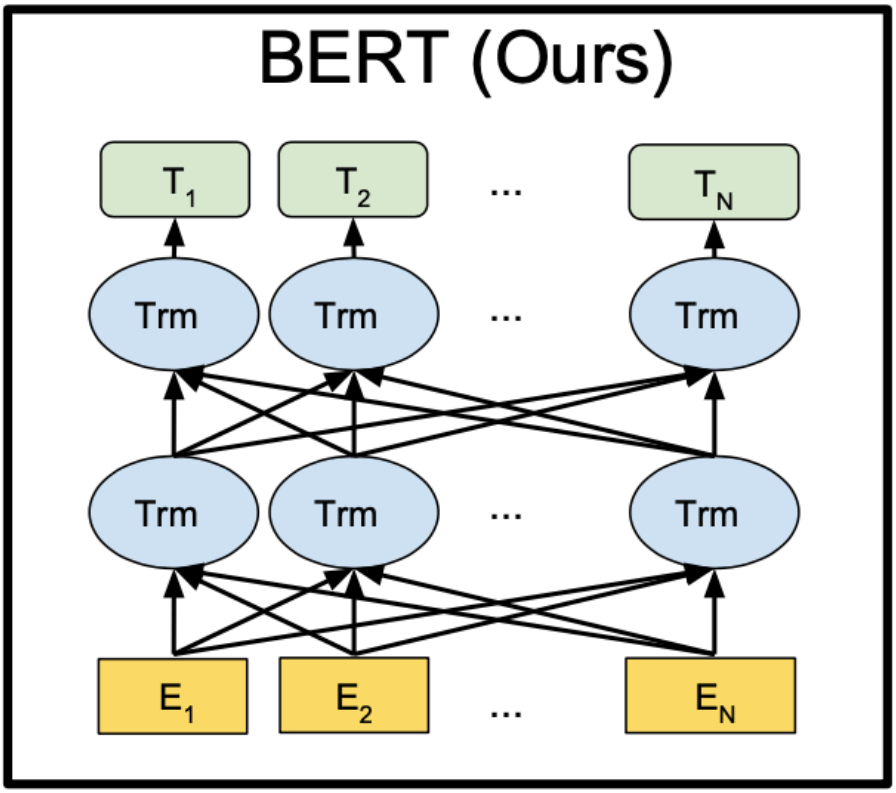

図 2.1: BERT の仕組み ([?] から引用)

2.1 BERT の学習の仕組み

本節は、BERT の学習方法について説明する。はじめに、学習用のデータとしては、wikipedia[2]などの大量のコーパスを用いる。例えば、コーパスの中に図 2.2[3]のような入力データ (input) が存在したとする。この時、先頭には [CLS]、2 つの文章 A、 B の間に [SEP] という特殊トークンを挿入して連結した構造が入力データになる。次に、入力データを投入した直後、トークンの ID が Token Embedding、文章 A、 B の区分が Segment Embedding、文章 A、B 内の位置が Position Embedding と、それぞれ事前学習の過程で学習される H 次元の埋め込み表現に置き換えられ、それらを加算したベクトルが Transformer へ送られる。

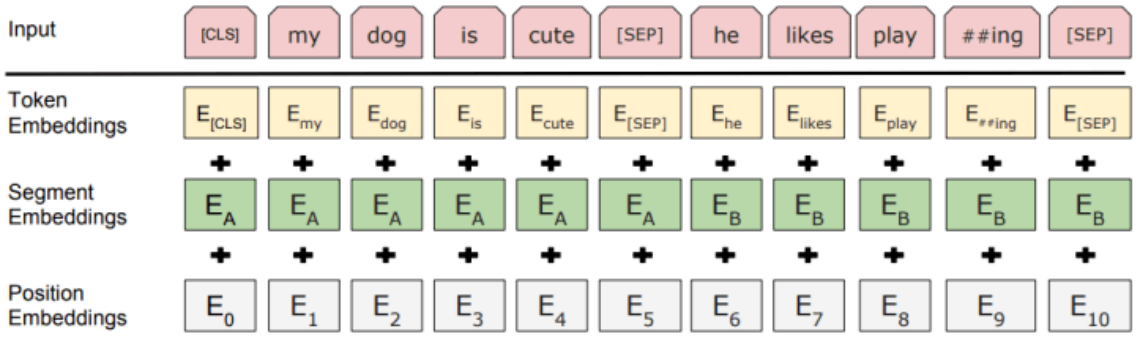

図 2.2: BERT において入力データを Transformer に送るときの仕組み

2.2 事前学習

2.2.1 MLM(Masked Language Model)

MLM とは、コーパス内から特定のトークンをランダムに選び、MASK トークンに置き換えて隠した状態にする手法である。例えば、「my dog is cute.」という文章をランダムに選んだとして、「my dog [MASK] cute」のように置き換える。この文に Transformer を適用し、[MASK] に相当するトークンを正しく推測できるように学習する。

2.2.2 NSP(Next Sentense Prediction)

NSP とは、コーパス内の 50%を文章 A、文章 B が連続したものを [MASK] 化したもの (正例)、50%の文章 A、文章 B を不連続なもの (負例) として学習を行う。例えば、「my dog [MASK] cute. [SEP]he likes playing.[SEP]」(正例) と「[CLS]my dog [MASK] cute. [SEP]I [MASK] watch ##ing TV yesterday.[SEP]」(負例) に分ける。この文に Transformer を適用し、[MASK] に相当するトークンを正しく推測できるように学習する。この手法によって、質問応答やテキスト分類において正解の特徴、不正解の特徴を識別することが可能となる。

[2]https://ja.wikipedia.org/wiki/メインページ
[3]https://arxiv.org/pdf/1810.04805.pdf から引用

2.3 Fine-tuning

前節の事前学習の内容を再利用して学習を行うことを Fine-tuning という。これにより、図 2.3 のように質問応答やテキスト分類などの様々なタスクに応用することができる。

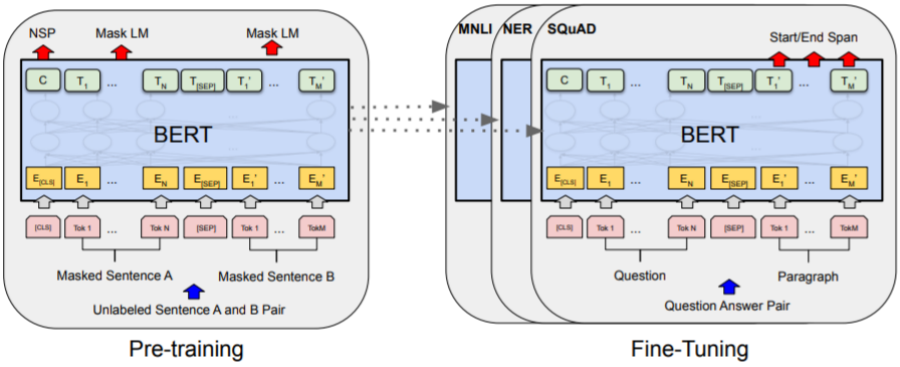

図 2.3: 事前学習と Fine-tuing の関係図

2.4 RoBERTa

前節の BERT モデルの登場以降、BERT を派生したモデル [?][?] が次々と提案されている。その中でも、新たな自然言語モデルの RoBERTa[?] は、BERT の事前学習の方法を改善して Fine-tuning を行うことで、短い学習時間で高い精度を実現させた最新の言語モデルである。主な改善方法としては、4 つある。1 つ目は、前節の BERT の MLM のマスクを静的 (static) から動的 (dynamic) に変更した。前者は学習前に [MASK] を一度作成したら、それを変更せずに使い回す手法をとっている。後者は [MASK] を作成したら使い回さず、毎回 [MASK] が適用される位置を変更する手法をとっている。(図 2.4[4]参照)

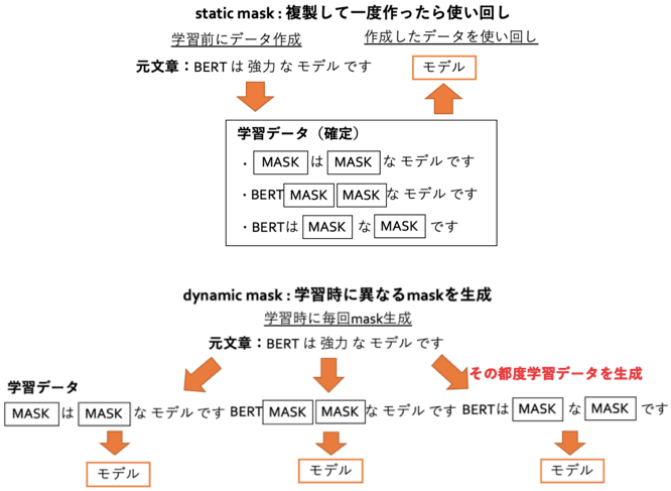

図 2.4: static mask と dynamic mask の違い

[4]https://ai-scholar.tech/others/roberta-ai-230 から引用

2 つ目は、前節の NSP を削除した。3 つ目は、学習データを 10 倍に増加させた。具体的には、BERT で使用されていた学習データ (16[GB]) にオープンデータを加えて 10 倍の 160[GB] のデータを用いた。それに伴い、学習率とバッチサイズと学習ステップ数も適切なものに変更している。4 つ目は、図 2.1 のように入力文字列のトークナイズを文字単位からバイト単位のものに変更した (BPE)。BPE は文字レベルと単語レベルの双方を考慮したエンコーディングで、単語全体だけを考慮するより、サブワードに着目することができる。例えば、recycle という英単語を BPE で読み込む時に文字レベルを re と cycle、単語レベルを recycle とすることで、双方を考慮したエンコーディングを行える。

表 2.1: RoBERTa のエンコーディングの例

手順	従来の encode
入力	明日は
辞書を使った分割	明 / 日 / は
辞書でマッピング	4369/4349/1948

手順	RoBERTa の encode
入力	明日は
UTF-8 でエンコード	\xe6\x98\x8e / \xe6\x97\xa5 / \xe3\x81\xaf
BPE の語彙でマッピング	48647 12736 /47954 8210 / 48549

以上の 4 点を変更することで従来の BERT の精度を上回ることができた。

第3章　検証手順

3.1　検証手順の概要

本稿では、最新手法の自然言語モデルである RoBERTa に対して、対象となる評価データセットを意図的に変更することで、モデルの頑健性を検証する。具体的には、RoBERTa に評価データセットの内容を読み取らせて、正解したデータと不正解になったデータに分類する。(第 1 段階)。そして、それぞれのデータを意図的に変更し、敵対的なデータを作成する。それを再びモデルで評価してモデルの性能を検証する (第 2 段階)。

3.2　データセットの概要

3.2.1　CommonsenseQA

CommonsenseQA[**?**] とは、一般常識 (Commonsense) に基づいた多肢選択式の質問応答データセットであり、Alon Talmor らによって作成された。各データの数は下記の表 3.1、検証に使用する評価用データの内容と質問の型は下記の表 3.2、質問としては、下記の図 3.1 のようなものである。

表 3.1: CommonsenseQA の各データ数

データセット名	学習用データ	評価用データ	テスト用データ
CommonsenseQA	9741	1221	1140

表 3.2: CommonsenseQA の評価用データの詳細

質問の型	What	Where	How	Why	Which	Who	When	その他	合計
型の数	761	346	42	34	13	8	4	13	1221
割合 (%)	62.33	28.34	3.44	2.78	1.06	0.66	0.33	1.06	100

```
{"answerKey": "A", "id": "1afa02df02c908a558b4036e80242fac",
 "question": {"question_concept": "revolving door",
 "choices": [{"label": "A", "text": "bank"},
{"label": "B", "text": "library"},
{"label": "C", "text": "department store"},
{"label": "D", "text": "mall"},
{"label": "E", "text": "new york"}],
 "stem": "A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?"}}
```


図 3.1: CommonsenseQA の質問

3.2.2 GLUE

GLUE(the General Language Understanding Evaluation)[**?**] とは、質問応答や感情分析、テキスト分類など一般的な言語理解評価のためのデータセットであり、Alex Wang らによって作成された。各データセットの概要とデータ数は、それぞれ下記の表 3.3、表 3.4 である。また、検証で用いる各データセットの評価用データの内訳が表 3.5、各データセットの問題例が表 3.6 である。

表 3.3: GLUE の各データセットの概要

	データセット名	概要
GLUE	MRPC	2 つの文章のペアは同じ意味かを判定。
	QNLI	質問と文章のペアは正しい答えを含んでいるかを判定。
	QQP	2 つの質問のペアは同じ意味かを判定。
	MNLI	2 つの文章のペアは含意・中立・矛盾か判定。
	SST-2	映画に対しての感情文が良いか悪いかを判定。
	CoLA	英文が構文的に正しいかを判定。

表 3.4: GLUE の各データセットのデータ数

	データセット名	学習用データ	評価用データ	テスト用データ
GLUE	MRPC	3668	408	1725
	QNLI	104743	5463	5463
	QQP	363870	40431	390965
	MNLI	392702	9815	9796
	SST-2	67349	872	1821
	CoLA	8551	1043	1061

表 3.5: 各評価データセットの内訳

データセット	問題数	正解ラベル / 件数		正解ラベル / 件数		正解ラベル / 件数	
MRPC	408	意味的に同じ	279	意味的に違う	129	ー	ー
QNLI	5463	含意	2702	含意でない	2761	ー	ー
QQP	40430	意味的に同じ	24756	意味的に違う	15674	ー	ー
MNLI	9815	含意	3479	中立	3123	矛盾	3213
SST-2	872	良い感情	444	悪い感情	428	ー	ー
CoLA	1043	正しい文法	721	正しくない文法	322	ー	ー

表 3.6: GLUE の各データセットの問題例

	データセット名	問題例
GLUE	MRPC	I'm 22 years old. Today is my 22nd birthday. (意味的に同じ) I'm 22 years old. Tomorrow is my 22nd birthday.(意味的に違う)
	QNLI	Where does Mike live? Mike lives in America.(含意) Where does Mike live? Mike'brother lives in America.(含意でない)
	QQP	Why is life difficult? Why is life severe? (意味的に同じ) Why is life difficult? Why is life easy?(意味的に違う)
	MNLI	I like banana. My favorite food is banana.(含意) I went to Kyoto.Hokkaido is so cold.(中立) I don' know why. I know the reason.(矛盾)
	SST-2	The movie is so amazing. (良い感情) Main character was not cool. (悪い感情)
	CoLA	This is a pen.(正しい文法) I are tired.(正しくない文法)

3.3 事前準備

3.3.1 環境構築

本研究では検証を計算機サーバ上で行うために、python3.59 を用いて環境を構築する。具体的な構築方法としては、anyenv を用いて pyenv をインストールする。その後、pyenv で python3.59 の仮想環境を設定する。環境構築が終了したら、検証に必要なモジュール (表 3.7) をインストールする。

表 3.7: 仮想環境に必要なモジュール一覧

モジュール名	バージョン
fairseq	0.8.0
torch	1.3.1
numpy	1.17.4
example	0.1.0
cffi	1.13.2
fastBPE	0.1.0
tokenizer	2.0.3
scipy	1.4.1

モジュール名	バージョン
regex	2019.12.19
sacrebleu	1.4.3
tqdm	4.40.2
pycparser	2.19
portalocker	1.5.2
typing	3.7.4.1
spacy	2.2.3
sentensepiece	0.1.85

3.3.2 事前学習

本実験では、wikitext-103-raw[5]を学習用のコーパスとして使用する。学習方法としては、2 章の RoBERTa の関連研究で説明した方法を用いる。

3.3.3 Fine-tuning

本検証では、BERT-Large を派生した事前学習モデルを使用して、各データセットの Fine-tuning を行う。下記の表 3.8 は各データセットの Fine-tuning の際に必要なパラメータとその設定値である。

表 3.8: CommonsenseQA,GLUE における Fine-tuning 時のパラメータ

データセット名	CommonsenseQA
epoch 数	5
学習率	1e-05
batch size	16
学習ステップ数	3000

MRPC	QNLI	QQP	MNLI	SST-2	CoLA
10	10	10	10	10	10
1e-5	1e-5	1e-5	1e-5	1e-5	1e-5
16	32	32	32	32	16
2296	33212	113272	123873	20935	5336

[5]https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset

3.4 CommonsenseQA の検証手順

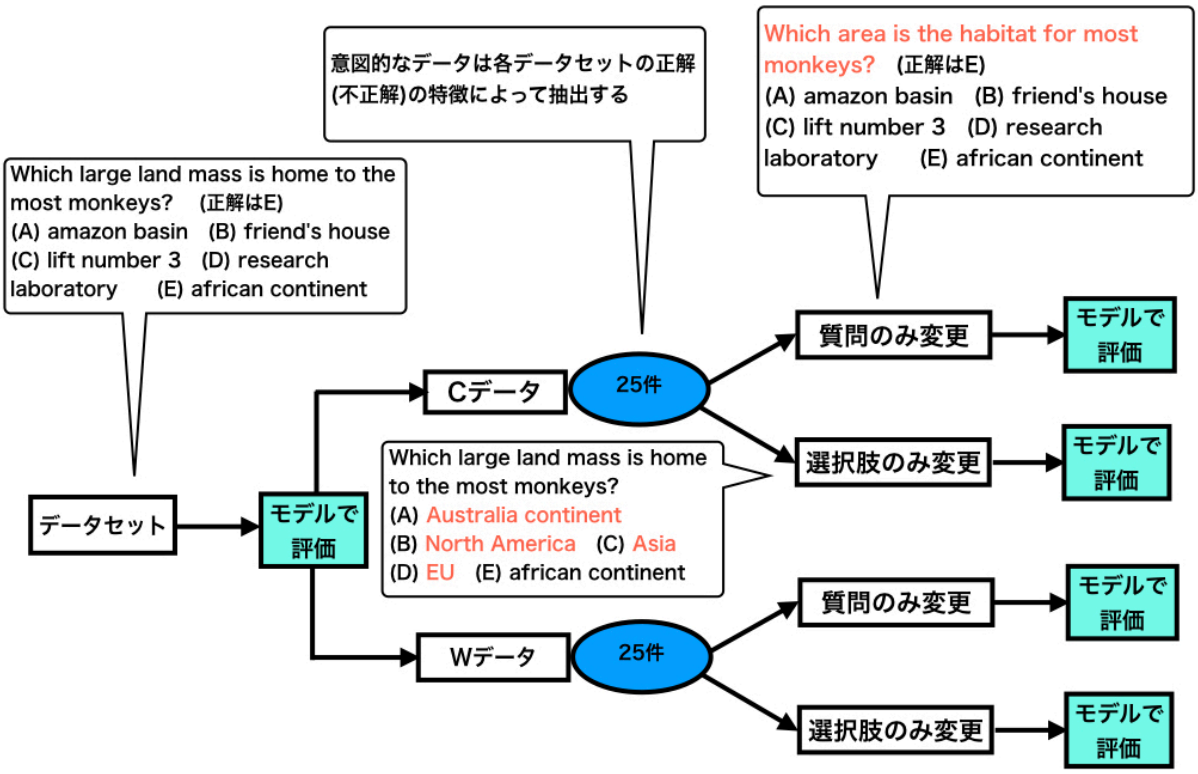

図 3.2: CommonsenseQA の検証手順 (全体図)

3.4.1 第 1 段階の概要

はじめに、CommonsenseQA の評価用データを github[6]上に公開されている RoBERTa の評価用プログラムを用いて上図 3.2 のように C データと W データに分類する。分類した結果が表 3.9 である。

[6]https://github.com/pytorch/fairseq/tree/master/examples/roberta/

3.4.2 CommonsenseQA の第 1 段階の結果

表 3.9: CommonsenseQA の第 1 段階におけるモデルの結果 (C データ)

データセット名	質問の型	What	Where	How	Why	Which	Who	When	その他	合計
CommonsenseQA	数	738	346	68	34	9	8	4	14	1221
	正答数	577	264	59	25	8	7	3	10	953
	正答率 (%)	78.2	76.3	86.8	73.5	88.9	87.5	75.0	71.4	78.0

3.4.3 第 2 段階の概要

第 1 段階では、評価用データを C データと W データに分類した。第 2 段階では、各データの特徴をもとに質問と選択肢のペアを手動で 50 件抽出する。この 50 件をそれぞれ「質問」のみを変更したもの、「選択肢」のみを変更したもの、計 100 件作成する。なお、下記の表 3.10、表**??**は抽出する際の各データの質問の型の件数と抽出する際の各データの特徴である。抽出が完了したら、表 3.12 のように各パターンにおいて意図的に単語を変更する。変更が終わったら、再び評価用プログラムにかける。

表 3.10: 抽出する際の各データの質問の型の件数

データセット	データ内容	パターン	抽出数	What	Where	Which	When	Who	Why	How	その他
CommonsenseQA	C データ	Question	25	5	5	2	2	2	3	3	3
		Choise	25	5	5	2	2	2	3	3	3
	W データ	Question	25	7	6	1	1	1	3	3	3
		Choise	25	7	6	1	1	1	3	3	3

表 3.11: 抽出する際の各データの特徴

データ内容	傾向	
C データ (25 件)	正答と誤答の選択肢の内容の差が大きい	
	質問	Which large land mass is home to the most monkeys?
	選択肢	(A) amazon basin (B) friend ’ s house (C) lift number 3 (D) research laboratory (E) african continent
W データ (25 件)	正答と誤答をはっきり区別出来ないことがある	
	質問	What event might one buy tickets for seats? (正解は B、モデルは D)
	選択肢	(A) park (B) show (C) auditorium (D) movies (E) rest area

表 3.12: 新たなデータを作成する方法とその具体例

変更部分	変更方法	
質問のみ	質問の意味を大きく変えないように、単語やフレーズを変更・追加	
	質問 選択肢 (変更前)	What event might one buy tickets for seats? (A) park (B) show (C) auditorium (D) movies (E) rest area
	質問 選択肢 (変更後)	What event might one get tickets for reserved seats? (A) park (B) show (C) auditorium (D) movies (E) rest area
選択肢のみ	正答以外の選択肢について, より回答が難しくなるように単語を変更	
	質問 選択肢 (変更前)	What event might one buy tickets for seats? (A) park (B) show (C) auditorium (D) movies (E) rest area
	質問 選択肢 (変更後)	What event might one buy tickets for seats? (A) subway (B) show (C) Beauty salon (D) train (E) movie ’ s preview

3.5　GLUE の検証手順 1

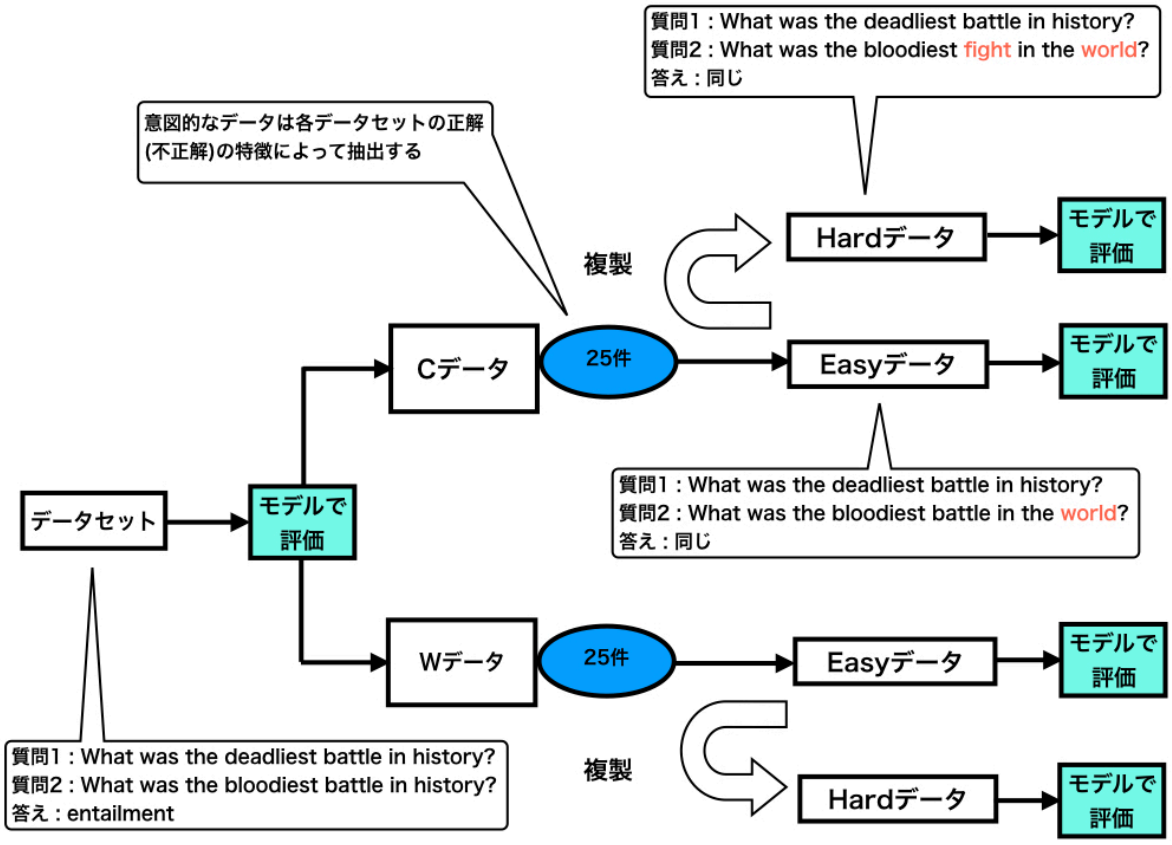

図 3.3: GLUE の検証手順 1

3.5.1　第 1 段階の概要

はじめに、GLUE の評価用データ (前節 3.3 節の表**??**～表**??**) を github[7] 上に公開されている RoBERTa の評価用プログラムを用いて上図 3.3 のように C データと W データに分類する。分類した結果が表 3.13 である。

[7]https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.glue.md

3.5.2 GLUE の第 1 段階の結果

表 3.13: MRPC,QNLI,QQP の第 1 段階におけるモデルの結果

データセット名	問題数	正解ラベル	正解数 / ラベル数	Accuracy
MRPC	408	意味的に正しい	258/279	0.90(369/408)
		意味的に違う	111/129	
QNLI	5463	含意	2569/2702	0.95(5188/5463)
		含意でない	2619/2761	
QQP	40430	意味的に正しい	23591/24756	0.92(37311/40430)
		意味的に違う	13720/15674	
MNLI	9815	含意	3097/3479	0.90(8867/9815)
		中立	2797/3123	
		矛盾	2973/3213	

表 3.14: MNLI の第 1 段階におけるモデルの誤答の結果

データセット名	問題数	正解ラベル (モデル)	誤答数	正解ラベル (モデル)	誤答数
MNLI	948	含意 (中立)	336/382	含意 (矛盾)	46/382
		中立 (含意)	166/326	中立 (矛盾)	160/326
		矛盾 (含意)	58/240	矛盾 (中立)	182/240

3.5.3 第 2 段階の概要

第 1 段階で分類した正解 (C) と不正解 (W) の各データにおいて、それぞれのデータの特徴をもとに文章のペアを手動で 50 件抽出する。抽出したペアのうち、どれかひとつの文章の名詞を、同じ意味の別の単語へ変更する。これを各データの easy データとする。さらに、easy データの内容を複製して、他の品詞の単語の類義語をできるだけ変更する。これを各データの hard データとする。なお、下記の表 3.15、表 3.18、表 3.21, 表 3.22 は各データの傾向である。これらの傾向を元に、各データの変更を次表 3.16、3.19、3.23、3.24 のように変更する。変更を完了したら、再び評価用プログラムにかける。表 3.17、3.20、3.25 は各データの抽出件数である。

表 3.15: MRPC,QQP の各データの傾向

データ内容	正解ラベル	傾向	例
C データ	意味が同じ	同じ意味と 言える言い換え	文 1 : Cisco pared spending to compensate for sluggish sales. 文 2 : In response to sluggish sales, Cisco pared spending.
	意味が違う	各品詞の意味の間違いが 顕著である	質問 1 : How do I buy used car in India? 質問 2 : Which used vehicle should I purchase in India?
W データ	意味が同じ	同じ意味とは 厳密には言えない 言い換え	文 1 : They ’ ve been in the stores for over six weeks, says Carney. 文 2 : The quarterlies usually stay in stores for between six to eight weeks, ” Carney added.
	意味が違う	フレーズの意味が 異なることを 読み取れていない	質問 1 : How do you make a pregnant belly costume? 質問 2 : How do you make an infant Batman costume?

表 3.16: MRPC,QQP の Easy,Hard の変更方法

データ内容	変更方法	
Easy	文ペアのいずれか一文を選択し、文中の名詞 1 単語を同義語に変換する	
	変更前	文 1 : Cisco pared spending to compensate for sluggish sales. 文 2 : In response to sluggish sales Cisco pared spending.
	変更後	文 1 : Cisco pared cost to compensate for sluggish sales. 文 2 : In response to sluggish sales Cisco pared spending .
Hard	Easy で作成した内容を元に、Easy で変更した箇所以外の品詞を 少なくとも 1 つ以上できるだけ多く同義語に変更する	
	変更前	文 1 : Cisco pared cost to compensate for sluggish sales. 文 2 : In response to sluggish sales Cisco pared spending.
	変更後	文 1 : Cisco economized cost to stop the weakness of sales. 文 2 : In response to sluggish sales Cisco pared spending .

表 3.17: MRPC,QQP の抽出件数

データセット	データ内容	データ難易度	問題数	正解ラベル (意味が同じ)	正解ラベル (意味が違う)
MRPC	C データ (新たなデータ)	Easy	25	13	12
		Hard	25	13	12
QQP	W データ (新たなデータ)	Easy	25	12	12
		Hard	25	12	12

表 3.18: QNLI の各データの傾向

データ内容	正解ラベル	傾向	例
C データ	含意	含意だと 言える言い換え	質問 : Where was war fought? 文章 : The war was fought primarily along the frontiers between New France and the British colonies, from Virginia in the South to Nova Scotia in the North.
	含意でない	各品詞の意味の間違いが 顕著である	質問 : Where did the Exhibition take place? 文章 : This World's Fair devoted a building to electrical exhibits.
W データ	含意	含意とは 厳密には言えない 言い換え	質問 : What does civil noncompliance protest against? 文章 : Civil disobedience is one of the many ways people have rebelled against what they deem to be unfair laws.
	含意でない	フレーズの意味が 異なることを 読み取れていない	質問 : Which rail company provides local and regional services? 文章 : Train operator Virgin Trains East Coast provides a half-hourly frequency of trains to London King's Cross, with a journey time of about three hours, these services call at Durham, Darlington, York, Doncaster, Newark North Gate and Peterborough and north to Scotland with all trains calling at Edinburgh and a small number of trains extended to Glasgow, Aberdeen and Inverness.

表 3.19: QNLI の Easy,Hard の変更方法

データ内容	変更方法	
Easy	質問文を選択し、名詞 1 単語を同義語に変換する	
	変更前	質問 : Where was war fought? 文章 : The war was fought primarily along the frontiers between New France and the British colonies, from Virginia in the South to Nova Scotia in the North.
	変更後	質問 : Where was warfare fought? 文章 : The war was fought primarily along the frontiers between New France and the British colonies, from Virginia in the South to Nova Scotia in the North.
Hard	Easy で作成した内容を元に、Easy で変更した箇所以外の品詞を 少なくとも 1 つ以上できるだけ多く同義語に変更する	
	変更前	質問 : Where was warfare fought? 文章 : The war was fought primarily along the frontiers between New France and the British colonies, from Virginia in the South to Nova Scotia in the North..
	変更後	質問 : Where was warfare battled? 文章 : The war was fought primarily along the frontiers between New France and the British colonies, from Virginia in the South to Nova Scotia in the North.

表 3.20: QNLI の抽出件数

データセット	データ内容	データ難易度	問題数	正解ラベル (含意)	正解ラベル (含意でない)
QNLI	C データ (新たなデータ)	Easy	25	13	12
		Hard	25	13	12
	W データ (新たなデータ)	Easy	25	12	12
		Hard	25	12	12

表 3.21: 抽出する際の MNLI の C データの傾向

データセット名	C データの特徴	
MNLI	正解ラベル (含意)	・2 つの文章のペアにおいて、各品詞の言い換えが簡単な単語 (文法) が多い。
	正解ラベル (中立)	・2 つの文章のペアにおいて、主語が同じでも、動詞または目的語が明らかに違う時は、 「中立」と判断できている。 ・2 つの文章のペアにおいて、全ての品詞の単語が明らかに違う。
	正解ラベル (矛盾)	・2 つの文章のペアにおいて、動詞の肯定文と動詞の否定文のペアは「矛盾」と判断できている。

表 3.22: 抽出する際の MNLI の W データの傾向

データセット名	W データの特徴	
MNLI	正解 : 含意 モデル : 中立	・2 つの文章ペアにおいて、具体的な文章と抽象的な文章を比較するときに 単語の意味を理解していないため「中立」と判断していた
	正解 : 含意 モデル : 矛盾	・2 つの文章ペアにおいて、各品詞の同義語を読み取れていないために「矛盾」と判断した。
	正解 : 中立 モデル : 含意	・2 つの文章ペアにおいて、各品詞の一部のペアが同じ、または同義語のときに「含意」と判断した。
	正解 : 中立 モデル : 矛盾	・2 つの文章ペアにおいて、各名詞のみが一致しているため「矛盾」と判断した。。
	正解 : 矛盾 モデル : 含意	・2 つの文章ペアにおいて、各動詞の対義語を読み取れていないため「含意」と判断した。 ・2 つの文章ペアにおいて、前置詞のニュアンスの違いを読み取れていないため「含意」と判断した。
	正解 : 矛盾 モデル : 中立	・2 つの文章ペアにおいて、「主語」は同じだが、動詞と目的語が違うので「中立」と判断した。

表 3.23: MNLI-easy の新たなデータの作成方法と具体例

データセット名	データ難易度	正解ラベル	新たなデータの作成方法と具体例 (C データと W データ共通)
MNLI	easy	含意と中立	・2 つの文章ペアのうち、1 つの文章を選択する。 ・その文章の「名詞」一単語を同義語に変換する。 (例) Tuppence rose.、Tuppence floated into the air.(変更前) Tuppence rose.、Twopence floated into the air.(変更後)
		矛盾	・2 つの文章ペアのうち、1 つの文章を選択する。 ・その文章の「動詞」一単語を同義語に変換する。 (例) yeah i know and i did that all through college and it worked too　I did that all through college but it never worked.(変更前) yeah i know and i did that all through college and it went well too　I did that all through college but it never worked.(変更後)

表 3.24: MNLI-hard の新たなデータの作成方法と具体例

データセット名	データ難易度	正解ラベル	新たなデータの作成方法と具体例 (C データと W データ共通)
MNLI	hard	含意と中立	・easy の内容を複製する。 ・easy で選択した質問の他の品詞を同義語に変換する。 (例) Tuppence rose. Twopence floated into the air.(easy) Tuppence rose. Twopence floated in the air.(変更後)
		矛盾	・easy の内容を複製する。 ・easy で選択した質問の他の品詞を同義語に変換する。 (例) yeah i know and i did that all through college and it went well too. I did that all through college but it never worked.(easy) yeah i know and i did that all through college and that also went well. I did that all through college but it never worked.(変更後)

表 3.25: MNLI の抽出件数

データセット	データ内容	データ難易度	問題数	正解ラベル (含意)	正解ラベル (中立)	正解ラベル (矛盾)
MNLI	C データ (新たなデータ)	easy	25	9	8	8
		hard	25	9	8	8
	W データ (新たなデータ)	easy	25	9	8	8
		hard	25	9	8	8

3.6　GLUE の検証手順 2

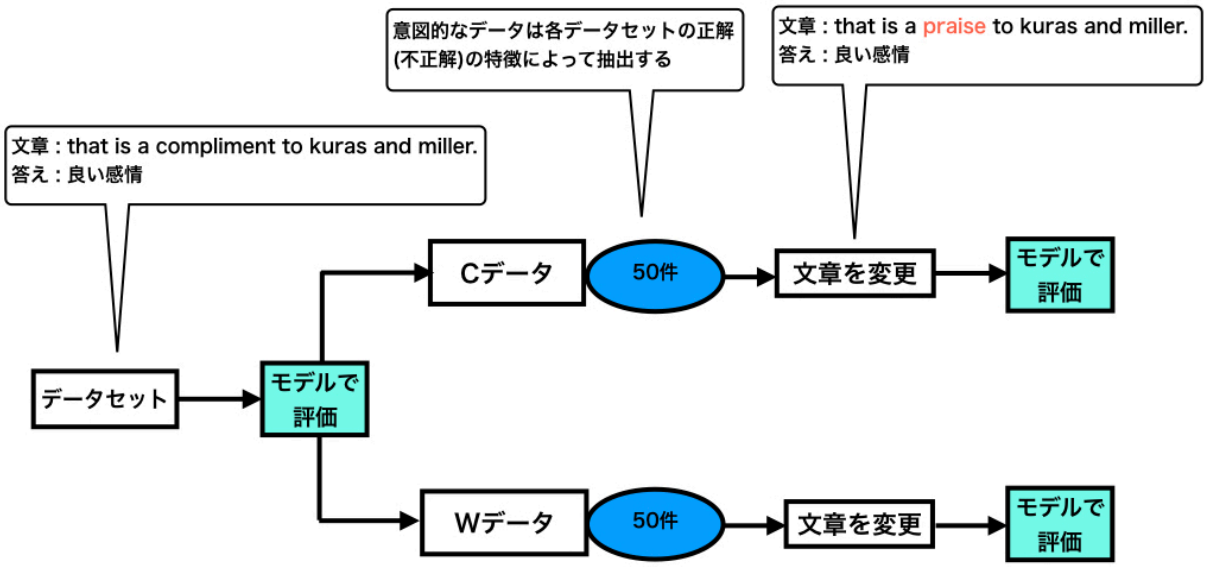

図 3.4: GLUE(SST-2 と CoLA) の検証手順 (全体図)

3.6.1　第 1 段階の概要

はじめに、GLUE の評価用データを github[7] 上に公開されている RoBERTa の評価用プログラムを用いて上図 3.4 のように C データと W データに分類する。分類した結果が表 3.26 である。

3.6.2　GLUE の第 1 段階の結果

表 3.26: SST-2,CoLA の第 1 段階におけるモデルの結果

データセット	問題数	正解ラベル	正解数 / ラベル数	Accuracy
SST-2	872	良い感情	427/444	0.96(838/872)
		悪い感情	411/428	

データセット	問題数	正解ラベル	正解数 / ラベル数	Matthews corr
CoLA	1043	文法的に正しい	667/721	0.67
		文法的に正しくない	232/322	

[7]https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.glue.md

3.6.3　第 2 段階の概要

第 1 段階で分類した正解 (C) と不正解 (W) の各データにおいて、それぞれのデータの特徴をもとに文章を手動で 50 件抽出する。抽出した文章の単語の類義語をできるだけ変更する。なお、下記の表 3.27、表 3.30 は各データの傾向である。これらの傾向を元に、各データの変更を次表 3.28、3.31、3.32 のように変更する。変更を完了したら、再び評価用プログラムにかける。表 3.29、3.33 は各データの抽出件数である。

表 3.27: SST-2 の各データの傾向

データ内容	正解ラベル	傾向	例
C データ	良い感情	・良い感情のみが書かれている。	a fast , funny , highly enjoyable movie .
	悪い感情	・悪い感情のみが書かれている。	a sometimes tedious film .
W データ	良い感情	・良い感情と悪い感情が混ざっている	as unseemly as its title suggests .
	悪い感情		the lower your expectations , the more you ’ll enjoy it .

表 3.28: SST-2 の変更方法

データ内容	変更方法 (C データと W データ共通)	
良い感情	抽出した文章の単語をできるだけ同義語に変換する。	
	変更前	a gorgeous,witty,seductive movie.
	変更後	Very beautiful,witty,attractive movie.
悪い感情	抽出した文章の単語をできるだけ同義語に変換する。	
	変更前	i had to look away this was god awful.
	変更後	i had to cast my eyes aside,this was god terrible.

表 3.29: SST-2 の抽出件数

データセット	データ内容	問題数	正解ラベル (良い感情)	正解ラベル (悪い感情)
SST-2	C データ (新たなデータ)	50	25	25
	W データ (新たなデータ)	30	15	15

表 3.30: CoLA の各データの傾向

データ内容	正解ラベル	傾向	例
C データ	正しい文法	・文章が「主語」「動詞」「目的語」の順番にきちんと並べられている。	Bill sent a package to Tom.
	正しくない文法	・同じ品詞が連続または、繰り返している。 ・自動詞しかとらない動詞なのに、他動詞になっている文は「正しくない」と判断した。	Lora buttered at the toast.
W データ	正しい文法	・自動詞と他動詞、両方の意味をとる動詞を間違える傾向があった。	Carla slid the book.
	正しくない文法	・自動詞か他動詞かを判断する時に、前置詞がいる/いらないで間違えた。	Which house does your friend live?

表 3.31: CoLA の C データの変更方法

データ内容	変更方法 (C データ)	
正しい文法	ある文章において、間違えやすくするために、単語を変換し、複雑な文を作る。	
	変更前	Bill sent a package to Tom.
	変更後	Bill gave a strange package to Tom
正しくない文法	ある文章において、動詞の後の前置詞のみを似たようなものに変更する。	
	変更前	Lora buttered at the toast.
	変更後	Lora buttered in the toast.

表 3.32: CoLA の W データの変更方法

データ内容	変更方法 (W データ)	
正しい文法	ある文章において、動詞を同義語に変換する。	
	変更前	Carla slid the book
	変更後	Carla moved the book.
正しくない文法	ある文章において、単語をひとつ変換して意図的に間違った文を作成する。	
	変更前	Which house does your friend live?
	変更後	Which house does your friend lives?

表 3.33: CoLA の抽出件数

データセット	データ内容	問題数	正解ラベル (正しい文法)	正解ラベル (正しくない文法)
CoLA	C データ (新たなデータ)	50	25	25
	W データ (新たなデータ)	50	25	25

第4章　実験

4.1　目的

本実験では、3 章で作成した新たなデータを Roberta で再評価する。各データセットの Accuracy または Matthews corr の結果から、モデルの頑健性を検証する。

4.2　検証結果

4.2.1　CommonsenseQA の第 2 段階の結果

表 4.1: CommonsenseQA の第 2 段階におけるモデルの結果

データセット名	対象データ	データ内容	正答数	What	Where	How	Why	Which	Who	When	その他	Accuracy
CommonsenseQA	C データ (再評価後)	Question	21	4/5	4/5	3/3	3/3	2/2	2/2	1/2	2/3	0.84(21/25)
		Choise	23	5/5	4/5	3/3	3/3	2/2	2/2	1/2	3/3	0.92(23/25)
	W データ (再評価後)	Question	6	2/7	1/6	2/3	0/3	0/1	0/1	0/1	1/3	0.24(6/25)
		Choise	10	4/7	2/6	2/3	1/3	0/1	0/1	0/1	1/3	0.40(10/25)

第 2 段階におけるモデルの結果としては、表 4.1 の通りである。C データの「質問」のみを同義語に変更した時と「選択肢」のみを質問に従って変更した時の Accuracy はそれぞれ 0.84、0.92 であり、ある程度モデルの頑健性があると言える。また、第 1 段階で正答率が高かった「How」、「Which」、「Who」型の質問も全て正解できていた。W データでは、「質問」のみを同義語に変更すると Accuracy は 0.24 であり、ほぼランダムで答えているのと同じ結果となってしまった。

4.2.2 GLUE の第 2 段階の結果

表 4.2: MRPC の第 2 段階におけるモデルの結果

データセット名	データ内容	データ難易度	問題数	正解ラベル (意味的に同じ)	正解ラベル (意味的に違う)	Accuracy
MRPC	C データ (再評価後)	easy	25	12/12	13/13	1.0(25/25)
		hard	25	9/12	11/13	0.80(20/25)
	W データ (再評価後)	easy	25	2/13	2/12	0.16(4/25)
		hard	25	3/13	3/12	0.24(6/25)

表 4.3: QNLI の第 2 段階におけるモデルの結果

データセット名	データ内容	データ難易度	問題数	正解ラベル (含意)	正解ラベル (含意でない)	Accuracy
QNLI	C データ (再評価後)	easy	25	13/13	12/12	1.0(25/25)
		hard	25	10/13	11/12	0.84(21/25)
	W データ (再評価後)	easy	25	2/12	2/13	0.16(4/25)
		hard	25	0/12	5/13	0.20(5/25)

表 4.4: QQP の第 2 段階におけるモデルの結果

データセット名	データ内容	データ難易度	問題数	正解ラベル (意味的に同じ)	正解ラベル (意味的に違う)	Accuracy
QQP	C データ (再評価後)	easy	25	10/13	12/12	0.88(22/25)
		hard	25	10/13	12/12	0.88(22/25)
	W データ (再評価後)	easy	25	6/12	2/13	0.32(8/25)
		hard	25	7/12	6/13	0.52(13/25)

表 4.5: MNLI の第 2 段階におけるモデルの結果

データセット名	データ内容	データ難易度	問題数	正解ラベル (含意)	正解ラベル (中立)	正解ラベル (矛盾)	Accuracy
MNLI	C データ (再評価後)	easy	25	7/9	7/8	7/8	0.84(21/25)
		hard	25	6/9	7/8	6/8	0.76(19/25)
	W データ (再評価後)	easy	25	2/9	1/8	1/8	0.16(4/25)
		hard	25	3/9	4/8	1/8	0.32(8/25)

表 4.6: MNLI の第 2 段階におけるモデルの誤答の結果

データセット名	データ内容	データ難易度	問題数	誤答数	正解ラベル：含意 (モデル：中立)	正解ラベル：含意 (モデル：矛盾)	正解ラベル：中立 (モデル：含意)	正解ラベル：中立 (モデル：矛盾)	正解ラベル：矛盾 (モデル：含意)	正解ラベル：矛盾 (モデル：中立)
MNLI	C データ (再評価後)	easy	25	4	1/2	1/2	1/1	0/1	0/1	1/1
		hard	25	6	2/3	1/3	1/1	0/1	0/2	2/2
	W データ (再評価後)	easy	25	21	5/7	2/7	4/7	3/7	3/7	4/7
		hard	25	17	4/6	2/6	3/4	1/4	3/7	4/7

表 4.7: SST-2 の第 2 段階におけるモデルの結果

データセット名	データ内容	問題数	正解ラベル (良い感情)	正解ラベル (悪い感情)	Accuracy
SST-2	C データ (再評価後)	50	25/25	22/25	0.94(47/50)
	W データ (再評価後)	30	5/15	5/15	0.33(10/30)

表 4.8: CoLA の第 2 段階におけるモデルの結果

データセット名	データ内容	問題数	正解ラベル (正しい文法)	正解ラベル (正しくない文法)	Matthews corr
CoLA	C データ (再評価後)	50	20/25	22/25	0.69
	W データ (再評価後)	50	4/25	20/25	-0.52

第 2 段階におけるモデルの結果として、第 1 段階で正解したデータセット (Accuracy が 100%) を意図的に変更した時の Accuracy は easy で 84%～100%、hard になると 76%～88%と下降していたが、ある程度モデルの頑健性はあると考えられる。また、第 1 段階で不正解になったデータセット (Accuracy が 0%) を意図的に変更した時の Accuracy は easy で 16%～32%、hard になると 20%～52%と正解したデータセットとは逆に上昇していた。

第5章　考察

5.1　CommonsenseQA の考察

第 4 章で得た結果から考察を述べていく。まず、表 5.1～表 5.2 は第 1 段階で正解したデータの Question を変更したときの、正解した問題と不正解になった問題を一部抜粋したものである。

表 5.1: 第 1 段階で正解したデータの Question を変更したとき、正解した問題 (一部抜粋)

Question を変更したとき、正解した問題
(変更前) 第 1 段階 質問文：Why would a person like to have a large house? 選択肢：(A) have choice (B)mentally challenged (C) own house (D) obesity (E)lots of space モデルの回答： E(正解は E) (変更後) 第 2 段階 質問文：Why do people who are not so rich want to have a big house? 選択肢：(A)have choice (B)mentally challenged (C) own house (D) obesity (E)lots of space モデルの回答：E(正解は E)
(変更前) 第 1 段階 質問文：Aside from water and nourishment what does your dog need? 選択肢：(A) bone (B)charm (C) petted (D) lots of attention (E) walked モデルの回答： D(正解は D) (変更後) 第 2 段階 質問文：What is required of your dog besides water and nutrition? 選択肢：(A) bone (B)charm (C) petted (D) lots of attention (E) walked モデルの回答：D(正解は D)
(変更前) 第 1 段階 質問文：Animals make up a large part of the? 選択肢：(A) carrying cargo (B) favorite (C) ecosystem(D) nature (E) ecology モデルの回答： C(正解は C) (変更後) 第 2 段階 質問文：Animals including us are comprised the majority of the? 選択肢：(A) carrying cargo (B) favorite (C) ecosystem(D) nature (E) ecology モデルの回答：C(正解は C)

表 5.2: 第 1 段階で正解したデータの Question を変更したとき、不正解になった問題 (一部抜粋)

Question を変更したとき、不正解になった問題
(変更前) 第 1 段階 質問文：When getting in shape, this is something that does wonders? 選択肢：(A) eat more (B) starve (C) give up (D) period of recovery (E) jogging モデルの回答：E (正解は E) (変更後) 第 2 段階 質問文：When living healthy life, this is something that gives wonders? 選択肢：(A) eat more (B) starve (C) give up (D) period of recovery (E) jogging モデルの回答：A(正解は E)
(変更前) 第 1 段階 質問文：When is the worst time for having food? 選択肢：(A) digesting (B) not hungry (C) gas (D) weight gain (E) feeling of fullness モデルの回答：B (正解は B) (変更後) 第 2 段階 質問文：When is the least convenient time when you eat food? 選択肢：(A) digesting (B) not hungry (C) gas (D) weight gain (E) feeling of fullness モデルの回答：A(正解は B)
(変更前) 第 1 段階 質問文：If you're buying beer for a float trip what are you preparing to do? 選択肢：(A) get arrested (B) have fun (C) get sick (D) spend money (E) stupidity モデルの回答：B (正解は B) (変更後) 第 2 段階 質問文：If you buy alcohol for a trip, what does it make you do? 選択肢：(A) get arrested (B) have fun (C) get sick (D) spend money (E) stupidity モデルの回答：D(正解は B)

表 5.1 から分かることは、質問文の「品詞」を簡単な同義語に変換してもモデルは同じ回答を選択している。しかし、表 5.2 を見ると、質問文の品詞ではなく、「文法表現」(例えば、getting in shape → living healthy life や、are you prepareing to do → does it make you do) を同義語に変換することでニュアンスが変わり、間違いを引き起こすと考えた。※ (前者の例は、「体調を整える」→「健康的な生活をする」の同義語。後者の例は、「何を準備するか」→「あなたに何をさせるか」の同義語。)

次に、表 5.3〜表 5.4 は、第 1 段階で正解したデータの Choise を変更したときの、正解した問題と不正解になった問題を一部抜粋したものである。

表 5.3: 第 1 段階で正解したデータの Choise を変更したとき、正解した問題 (一部抜粋)

Choise を変更したとき、正解した問題
(変更前) 第 1 段階 質問文：Where would you find many varieties of plants including a rosebush? 選択肢：(A) kew gardens (B) garder (C) backyard (D) shop (E) beautiful garden モデルの回答： E(正解は E) (変更後) 第 2 段階 質問文：Where would you find many varieties of plants including a rosebush? 選択肢：(A) forest (B) Veranda (C) Countryside (D) zoo (E) beautiful garden モデルの回答：E(正解は E)
(変更前) 第 1 段階 質問文：Why would a person like to have a large house? 選択肢：(A) have choice (B) mentally challenged (C) own house (D) obesity (E) lots of space モデルの回答： E(正解は E) (変更後) 第 2 段階 質問文：Why would a person like to have a large house?? 選択肢：(A) to show off (B) challenge (C) revenge (D) Pride (E) lots of space モデルの回答：E(正解は E)
(変更前) 第 1 段階 質問文：If you want harmony, what is something you should try to do with the world? 選択肢：(A) take time (B) make noise (C) make war (D) make peace (E) make haste モデルの回答： D(正解は D) (変更後) 第 2 段階 質問文：If you want harmony, what is something you should try to do with the world? 選択肢：(A) Donation (B) sleep (C) travel (D) make peace (E) ignore モデルの回答：D(正解は D)

表 5.4: 第 1 段階で正解したデータの Choise を変更したとき、不正解になった問題 (全て抜粋)

Choise を変更したとき、不正解になった問題
(変更前) 第 1 段階 質問文：When is the worst time for having food? 選択肢：(A) digesting (B) not hungry (C) gas (D) weight gain (E) feeling of fullness モデルの回答： B(正解は B) (変更後) 第 2 段階 質問文：When is the worst time for having food? 選択肢：(A) cancer (B) not hungry (C) on a diet (D) Lose weight (E) no motivation モデルの回答：C(正解は B)
(変更前) 第 1 段階 質問文：Bob the lizard lives in a warm place with lots of water.Where does he probably live? 選択肢：(A) rock (B) tropical rainforest (C) jazz club (D) new mexico (E) rocky places モデルの回答： B(正解は B) (変更後) 第 2 段階 質問文：Bob the lizard lives in a warm place with lots of water.Where does he probably live? 選択肢：(A) near the sea (B) tropical rainforest (C) Temperate Zone (D) Hawaii (E) desert モデルの回答：A(正解は B)

表 5.3 から分かることは、質問文に従って意図的に選択肢を変更してもモデルは同じ回答を選択している。しかし、表 5.4 に注目する。誤答した 2 問に共通しているのは、間違えた選択肢の単語は一見正解に見えるがよく考えると「言い過ぎ」である。例えば、「トカゲはどこに住んでいるか」と言う質問に対してモデルは「海の近く」と回答している。確かに、海の近くは水がたくさんあるが「必ずしも」暖かい場所とは限らない。

さらに、表 5.5〜表 5.6 は、第 1 段階で不正解になったデータの Question を変更したときの、正解した問題と不正解になった問題を一部抜粋したものである。

表 5.5: 第 1 段階で不正解になったデータの Question を変更したとき、正解した問題 (一部抜粋)

Question を変更したとき、正解した問題
(変更前) 第 1 段階 質問文：The player lifted his cornet and walked in rhythm, what was the player a member of? 選択肢：(A) museum (B) high school band (C) marching band (D) orchestra (E) band モデルの回答： D(正解は C) (変更後) 第 2 段階 質問文：The player hit the cymbal and walked in rhythm, what was the player a member of? 選択肢：(A) museum (B) high school band (C) marching band (D) orchestra (E) band モデルの回答：C(正解は C)
(変更前) 第 1 段階 質問文：How would you get from one side of a canal to another? 選択肢：(A) michigan (B) amsterdam (C) venice (D) bridge (E) barges to travel on モデルの回答： E(正解は D) (変更後) 第 2 段階 質問文：In daily life,how would you get from one side of a canal to another? 選択肢：(A) michigan (B) amsterdam (C) venice (D) bridge (E) barges to travel on モデルの回答：D(正解は D)
(変更前) 第 1 段階 質問文：While waiting for this appointment, people often read magazines. 選択肢：(A) doctor (B) train station (C) newsagent(D) market (E) table モデルの回答： B(正解は A) (変更後) 第 2 段階 質問文：While waiting for this appointment from 10 minutes to 20 minutes, people often read magazines or newspaper. 選択肢：(A) doctor (B) train station (C) newsagent (D) market (E) table モデルの回答：A(正解は A)

表 5.5 では、質問文の「品詞」の単語を同義語にするのではなく、詳細を具体的に追加する (In daily life や from 10 minutes to 20 minutes) ことで正解できていることが分かる。これに対して、表 5.6 では、質問文の「品詞」の単語を同義語に変更すると不正解になってしまう。このように、第 1 段階で不正解になったデータを同義語に変換するのではなく、詳細を具体的に加えることで正解に導けることが判明した。

表 5.6: 第 1 段階で不正解になったデータの Question を変更したとき、不正解になった問題 (一部抜粋)

Question を変更したとき、不正解になった問題
(変更前) 第 1 段階 質問文：If not in a stream but in a market where will you find fish? 選択肢：(A) stream (B) aquarium (C) refrigerator (D) boat ride (E) market モデルの回答： B(正解は C) (変更後) 第 2 段階 質問文：If not in a river but in a market, where will you find dead fish? 選択肢：(A) stream (B) aquarium (C) refrigerator (D) boat ride (E) market モデルの回答：E(正解は C)
(変更前) 第 1 段階 質問文：There was a toll road that meandered from Maine to New Hampshire, where was it? 選択肢：(A) massachusetts (B) new england (C) my house (D) new jersey (E) connecticut モデルの回答： E(正解は B) (変更後) 第 2 段階 質問文：There wasn't a free road from Maine to New Hampshire, where did it belong to? 選択肢：(A) massachusetts (B) new england (C) my house (D) new jersey (E) connecticut モデルの回答：A(正解は B)
(変更前) 第 1 段階 質問文：When did mammoth's live? 選択肢：(A) boscage (B) forest (C) prehistory (D) prehistoric times (E) ancient times モデルの回答： D(正解は E) (変更後) 第 2 段階 質問文：When was the mammoth extinct? 選択肢：(A) boscage (B) forest (C) prehistory (D) prehistoric times (E) ancient times モデルの回答：C(正解は E)

最後に、表 5.7〜表 5.8 は、第 1 段階で不正解になったデータの Choise を変更したときの、正解した問題と不正解になった問題を一部抜粋したものである。

表 5.7: 第 1 段階で不正解になったデータの Choise を変更したとき、正解した問題 (一部抜粋)

Choise を変更したとき、正解した問題
(変更前) 第 1 段階 質問文：What event might buy tickets for seats? 選択肢：(A) park (B) show (C) auditorium (D) movies (E) rest area モデルの回答： D(正解は B) (変更後) 第 2 段階 質問文：What event might buy tickets for seats? 選択肢：(A) subway (B) show (C) Beauty salon (D) train (E) movie's preview モデルの回答：B(正解は B)
(変更前) 第 1 段階 質問文：The dad wanted to protect his house, where did he put his gun? 選択肢：(A) police station (B) crime scene (C) restroom (D) drawer (E) holster モデルの回答： E(正解は D) (変更後) 第 2 段階 質問文：The dad wanted to protect his house, where did he put his gun? 選択肢：(A) Bedroom (B) Warehouse (C) treasury (D) drawer (E) desk モデルの回答：D(正解は D)
(変更前) 第 1 段階 質問文：Why does someone want to examine thing closely? 選択肢：(A) buy (B) learn about (C) buy (D) complex (E) interesting モデルの回答： E(正解は B) (変更後) 第 2 段階 質問文：Why does someone want to examine thing closely? 選択肢：(A) What you need to do (B) learn about (C) homework (D) Compulsion (E) challenge モデルの回答：B(正解は B)

表 5.8: 第 1 段階で不正解になったデータの Choise を変更したとき、不正解になった問題 (一部抜粋)

Choise を変更したとき、不正解になった問題
(変更前) 第 1 段階 質問文：What do audiences clap for? 選択肢：(A) cinema (B) theatre (C) movies (D) show (E) hockey game モデルの回答： E(正解は D) (変更後) 第 2 段階 質問文：What do audiences clap for? 選択肢：(A) judgement (B) game (C) gift (D) show (E) sports モデルの回答：E(正解は D)
(変更前) 第 1 段階 質問文：Where could you find hundreds of thousands of home? 選択肢：(A) field (B) neighborhood (C) star can (D) city or town (E) apartment building モデルの回答： E(正解は D) (変更後) 第 2 段階 質問文：Where could you find hundreds of thousands of home? 選択肢：(A) village (B) island (C) earth (D) city or town (E) skyscraper モデルの回答：C(正解は D)
(変更前) 第 1 段階 質問文：How would you get from one side of a canal to another? 選択肢：(A) michigan (B) amsterdam (C) venice (D) bridge (E) barges to travel on モデルの回答： E(正解は D) (変更後) 第 2 段階 質問文：How would you get from one side of a canal to another? 選択肢：(A) walking (B) jet aeroplane (C) swimming (D) bridge (E) ladder モデルの回答：C(正解は D)

表 5.7 と表 5.8 では、質問に従って選択肢を変更した。正解した問題としては、正解と不正解の選択肢がやや極端な問題ばかりだった。一方で、不正解になった問題は、例えば、「数十万世帯の家をどこで見つけることができるか」という質問に対して、モデルは「地球」と答えている。地球には数十万世帯以上の家が存在するが、実際は数十億世帯であり、答えの「都市または街」の方が妥当である。また、「どのように川の片側から向こう側へ移動するか」という質問ではモデルは「泳ぐ」と答えている。確かに泳いで川を渡ることも可能だが、このようなことは普段はしないので「橋」の方が回答である。このようにモデルは、どちらとも取れるような選択肢に対して、優劣をつけることが十分でないと考えられる。

5.2 GLUE の考察

第 4 章で得た結果から考察を述べていく。

5.2.1 MRPC の考察

まず、表 5.9 と表 5.10 は、MRPC の第 1 段階において、正解したデータを hard データに変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.9: MRPC の第 1 段階で正解したデータを hard データに変更した時、正解した問題 (一部抜粋)

MRPC の C データにおいて、hard データに変更したとき、正解した問題
(変更前) 第 1 段階 文章 1：This decision is clearly incorrect , ” FTC Chairman Timothy Muris said in a written statement . 文章 2： The decision is ” clearly incorrect , ” FTC Chairman Tim Muris said . モデルの回答： 正しい (正解は正しい) (変更後) 第 2 段階 (hard) 文章 1：This decision is clearly incorrect , ” FTC Chairman Timothy Muris said in a written statement. 文章 2： The determination is not ” incorrect at all , ” FTC Chairman Tim Muris said . モデルの回答：正しい (正解は正しい)
(変更前) 第 1 段階 文章 1：He replaces Ron Dittemore , who announced his resignation in April . 文章 2：Dittemore announced his plans to resign on April 23 . モデルの回答： 正しくない (正解は正しくない) (変更後) 第 2 段階 (hard) 文章 1：He replaces Ron Dittemore , who declared his retirement in April . 文章 2：Dittemore announced his plans to resign on April 23 . モデルの回答：正しくない (正解は正しくない)

表 5.10: MRPC の第 1 段階で正解したデータを hard データに変更した時、不正解になった問題 (一部抜粋)

MRPC の C データにおいて、hard データに変更したとき、不正解になった問題
(変更前) 第 1 段階 文章 1：No dates have been set for the civil or the criminal trial . 文章 2： No dates have been set for the criminal or civil cases , but Shanley has pleaded not guilty . モデルの回答： 正しくない (正解は正しくない) (変更後) 第 2 段階 (hard) 文章 1：No dates have been set for the civil or the unlawful trial,but Shanley has pleaded that he was not innocent. 文章 2： No dates have been set for the criminal or civil cases , but Shanley has pleaded not guilty . モデルの回答：正しい (正解は正しくない)
(変更前) 第 1 段階 文章 1：It will be followed in November by a third movie , ” The Matrix Revolutions . 文章 2： ” The film is the second of a trilogy , which will wrap up in November with ” The Matrix Revolutions . ” モデルの回答： 正しい (正解は正しい) (変更後) 第 2 段階 (hard) 文章 1：It will be followed in November by a third movie , ” The Matrix Revolutions . 文章 2： ” The flick is the second of three related films, which will end up with ” The Matrix Revolutions in November. ” モデルの回答：正しくない (正解は正しい)
(変更前) 第 1 段階 文章 1：PeopleSoft also said its board had officially rejected Oracle ’s offer . 文章 2：Thursday morning , PeopleSoft ’s board rejected the Oracle takeover offer . モデルの回答： 正しい (正解は正しい) (変更後) 第 2 段階 (hard) 文章 1：PeopleSoft also said its board had officially rejected Oracle ’s offer . 文章 2：Thursday morning ,the Oracle proposed to PeopleSoft’s suggestion. But PeopleSoft did not received. モデルの回答：正しくない (正解は正しい)

表 5.9 では、文章ペアの品詞を簡単な同義語に変更した場合は、正解できていた。一方で、表 5.10 では、語順を入れ替えたり、複雑な対義語を追加した時には不正解になることが分かった。これは、モデルが適切な機械読解を行えていないためである。

次に、表 5.11 と表 5.12 は、MRPC の第 1 段階において、正解したデータを hard データに変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.11: MRPC の第 1 段階で不正解になったデータを hard データに変更した時、正解した問題 (一部抜粋)

MRPC の W データを hard データに変更したとき、正解した問題
(変更前) 第 1 段階 文章 1：Blair 's Foreign Secretary Jack Straw was to take his place on Monday to give a statement to parliament on the European Union . 文章 2： Blair 's office said his Foreign Secretary Jack Straw would take his place on Monday to give a statement to parliament on the EU meeting the prime minister attended last week . モデルの回答： 正しくない (正解は正しい) (変更後) 第 2 段階 (hard) 文章 1：Blair's Foreign Secretary Jack Straw took the place of him on Monday to give a proclamation to senate on the EU . 文章 2： Blair's office said his Foreign Secretary Jack Straw would take his place on Monday to give a statement to parliament on the EU meeting the prime minister attended last week . モデルの回答：正しい (正解は正しい)
(変更前) 第 1 段階 文章 1：About two hours later , his body , wrapped in a blanket , was found dumped a few blocks away . 文章 2：Then his body was dumped a few blocks away , found in a driveway on Argyle Road . モデルの回答： 正しい (正解は正しくない) (変更後) 第 2 段階 (hard) 文章 1：About two hours later, his corpse , wrapped in a blanket was a few blocks away . 文章 2：Then his body was dumped a few blocks away , found in a driveway on Argyle Road. モデルの回答：正しくない (正解は正しくない)

表 5.12: MRPC の第 1 段階で不正解になったデータを hard データに変更した時、不正解になった問題 (一部抜粋)

MRPC の W データを hard データに変更したとき、不正解になった問題
(変更前) 第 1 段階 文章 1：” They ’ve been in the stores for over six weeks , ” says Carney . 文章 2： The quarterlies usually stay in stores for between six to eight weeks , ” Carney added. モデルの回答： 正しくない (正解は正しい) (変更後) 第 2 段階 (hard) 文章 1： ” They ’ve been in the stores for over six weeks , ” says Carney . 文章 2： They usually stay in establishment for between 42 to 56 days , ” Carney added . モデルの回答：正しくない (正解は正しい)
(変更前) 第 1 段階 文章 1：The state ’s House delegation currently consists of 17 Democrats and 15 Republicans . 文章 2： Democrats hold a 17-15 edge in the state ’s U.S. House delegation. モデルの回答： 正しくない (正解は正しい) (変更後) 第 2 段階 (hard) 文章 1： The United state’s House representative currently concludes of 17 Democrats and 15 Republicans. 文章 2： Democrats hold a 17-15 edge in the state ’s U.S. House delegation . モデルの回答：正しくない (正解は正しい)
(変更前) 第 1 段階 文章 1：It decided instead to issue them before the stock market opened Monday after the downgrade of its debt late Friday by Moody ’s, the credit rating agency. 文章 2：It decided instead to issue them before the stock market opened Monday to counteract the downgrade of its debt late Friday by Moody ’s to one step above junk status . モデルの回答： 正しい (正解は正しくない) (変更後) 第 2 段階 (hard) 文章 1： It determined taking our place to issue them before the stock market held Monday after the downgrade of its liabilities late Friday by Moody ’s , the credit rating agency . 文章 2：It decided instead to issue them before the stock market opened Monday to counteract the downgrade of its debt late Friday by Moody ’s to one step above junk status . モデルの回答：正しい (正解は正しくない)

表 5.11 では、単語ではなく、文法表現を同義語に変換することで、モデルの予測ラベルが変わり、正解することができた。これに対して表 5.12 は、単語の同義語変換のみを行っているので不正解になっていると考えた。

5.2.2 QNLI の考察

表 5.13 と表 5.14 は、QNLI の第 1 段階において、正解したデータを hard データに変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.13: QNLI の第 1 段階で正解したデータを hard データに変更した時、正解した問題 (一部抜粋)

QNLI の C データにおいて、hard データに変更したとき、正解した問題
(変更前) 第 1 段階 質問：What do these teachers NOT do? 文章： These teachers do not teach by rote but attempt to find new invigoration for the course materials on a daily basis. モデルの回答： 含意 (正解は含意) (変更後) 第 2 段階 (hard) 質問：What do these instructors never do?. 文章： These teachers do not teach by rote but attempt to find new invigoration for the course materials on a daily basis. モデルの回答：含意 (正解は含意)
(変更前) 第 1 段階 質問：What year did the the case go before the supreme court? 文章：For example, Joseph Haas was arrested for allegedly sending an email to the Lebanon, New Hampshire city councilors stating, ”Wise up or die.” モデルの回答： 含意でない (正解は含意でない) (変更後) 第 2 段階 (hard) 質問：What year did the the event go before the high court? 文章：For example, Joseph Haas was arrested for allegedly sending an email to the Lebanon, New Hampshire city councilors stating, ”Wise up or die.”. モデルの回答：含意でない (正解は含意でない)

表 5.14: QNLI の第 1 段階で正解したデータを hard データに変更した時、不正解になった問題 (一部抜粋)

QNLI の C データにおいて、hard データに変更したとき、不正解になった問題
(変更前) 第 1 段階 質問：Where did the Exposition take place? 文章： This World's Fair devoted a building to electrical exhibits. モデルの回答： 含意でない (正解は含意でない) (変更後) 第 2 段階 (hard) 質問：Where did the Exhibition take space? 文章： This World's Fair devoted a building to electrical exhibits. モデルの回答：含意 (正解は含意でない)
(変更前) 第 1 段階 質問：What came into force after the new constitution was herald? 文章：As of that day, the new constitution heralding the Second Republic came into force. モデルの回答： 含意 (正解は含意) (変更後) 第 2 段階 (hard) 質問：What enforced after the national law was declared? 文章：As of that day, the new constitution heralding the Second Republic came into force. モデルの回答：含意でない (正解は含意)

表 5.13 と表 5.14 では、前述の MRPC 同様、文章ペアの品詞を簡単な同義語に変更した場合は、正解できていた。また、単語ではなく、文法表現 (came into force → enforced) を変更すると、不正解になることが分かった。

次に、表 5.15 と表 5.16 は、QNLI の第 1 段階において、不正解したデータを hard データに変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.15: QNLI の第 1 段階で不正解になったデータを hard データに変更した時、正解した問題 (一部抜粋)

QNLI の W データを hard データに変更したとき、正解した問題
(変更前) 第 1 段階 質問：What problems did the Yuan dynasty have near its end? 文章： In time, Kublai Khan's successors lost all influence on other Mongol lands across Asia, while the Mongols beyond the Middle Kingdom saw them as too Chinese. モデルの回答： 含意 (正解は含意でない) (変更後) 第 2 段階 (hard) 質問：What troubles did the Yuan dynasty get near its end? 文章： In time, Kublai Khan's successors lost all influence on other Mongol lands across Asia, while the Mongols beyond the Middle Kingdom saw them as too Chinese. モデルの回答：含意でない (正解は含意でない)
(変更前) 第 1 段階 質問：Where is corporal punishment practiced the most? 文章：This often used to take place in the classroom or hallway, but nowadays the punishment is usually given privately in the principal's office. モデルの回答： 含意 (正解は含意でない) (変更後) 第 2 段階 (hard) 質問：Where is physical punishment used the most? 文章：This often used to take place in the classroom or hallway, but nowadays the punishment is usually given privately in the principal's office. モデルの回答：含意でない (正解は含意でない)

表 5.16: QNLI の第 1 段階で不正解になったデータを hard データに変更した時、不正解になった問題 (一部抜粋)

QNLI の W データを hard データに変更したとき、不正解になった問題
(変更前) 第 1 段階 質問：Where can one find the formerly Huguenot farms in South Africa? 文章： Many of the farms in the Western Cape province in South Africa still bear French names. モデルの回答： 含意 (正解は含意でない) (変更後) 第 2 段階 (hard) 質問： Where can one discover the formerly Huguenot plantations in South Africa? 文章： Many of the farms in the Western Cape province in South Africa still bear French names. モデルの回答： 含意 (正解は含意でない)
(変更前) 第 1 段階 質問： What is the most important item for civil disobedience to follow through? 文章： The key point is that the spirit of protest should be maintained all the way, whether it is done by remaining in jail, or by evading it. モデルの回答： 含意でない (正解は含意) (変更後) 第 2 段階 (hard) 質問： What is the most significant item for civil disobedience to obey? 文章： The key point is that the spirit of protest should be maintained all the way, whether it is done by remaining in jail, or by evading it. モデルの回答：含意でない (正解は含意)
(変更前) 第 1 段階 質問：What is the name of the trophy given to anyone who plays on the winning team in a Super Bowl? 文章：Like the Lombardi Trophy, the 50 will be designed by Tiffany & Co. モデルの回答： 含意 (正解は含意でない) (変更後) 第 2 段階 (hard) 質問： What is the name of the trophy awarded to anyone who plays on the victorious squad in a Super Bowl? 文章：Like the Lombardi Trophy, the 50 will be designed by Tiffany & Co. モデルの回答：含意 (正解は含意でない)

表 5.15 では、頻繁に使われない単語 (practice=使うと言う意味) をよく使われる単語 (use) に変換すると、問題に正解している。つまり、W データの難解な単語を簡単な単語に変換すると、正解になりやすい。これに対して表 5.16 は、比較的簡単な単語を同義語に変換しているので正解することが出来なかったと考えられる。

5.2.3 QQP の考察

表 5.17 と表 5.18 は、QQP の第 1 段階において、正解したデータを hard データに変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.17: QQP の第 1 段階で正解したデータを hard データに変更した時、正解した問題 (一部抜粋)

QQP の C データにおいて、hard データに変更したとき、正解した問題
(変更前) 第 1 段階 質問 1：How can I upgrade my English Writing skills? 質問 2： How can I improve my English vocabulary and writing skills? モデルの回答： 意味的に同じ (正解は意味的に同じ) (変更後) 第 2 段階 (hard) 質問 1：How can I upgrade my English Writing skills? 質問 2：”How can I progress my English vocabulary and writing abilities?”. モデルの回答：意味的に同じ (正解は意味的に同じ)
(変更前) 第 1 段階 質問 1：How do I buy used car in India? 質問 2：Which used car should I buy in India? モデルの回答：意味的に違う (正解は意味的に違う) (変更後) 第 2 段階 (hard) 質問 1：How do I buy used car in India? 質問 2：Which used vehicle should I purchase in India? モデルの回答：意味的に違う (正解は意味的に違う)

表 5.18: QQP の第 1 段階で正解したデータを hard データに変更した時、不正解になった問題 (一部抜粋)

QQP の C データにおいて、hard データに変更したとき、不正解になった問題
(変更前) 第 1 段階 質問 1：What does Richard Muller think of philosophy? 質問 2：What does Richard Muller think about philosophy? モデルの回答： 意味的に同じ (正解は意味的に同じ) (変更後) 第 2 段階 (hard) 質問 1：What does Richard Muller think of philosophy? 質問 2： What does Richard Muller conceive about doctrine? モデルの回答：意味的に違う (正解は意味的に同じ)
(変更前) 第 1 段階 質問 1：Do I need a midical test for the visa interview to the US? 質問 2：Do I need a midical test for visa interview to US? モデルの回答： 意味的に同じ (正解は意味的に同じ) (変更後) 第 2 段階 (hard) 質問 1：Do I need a midical test for the visa interview to the US? 質問 2：Do I want a health check for visa interview to US? モデルの回答：意味的に違う (正解は意味的に同じ)

表 5.17 では、文章ペアの品詞を簡単な同義語に変更した場合は、正解できていた。一方で、表 5.18 では、前述の QNLI と同様に、よく使われる単語 (want) を頻繁に使われない単語 (want=必要があると言う意味) に変換すると、正解することが出来なかった。つまり、C データの単語を難しくすることで間違いを誘発することが出来る。

次に、表 5.19 と表 5.20 は、QQP の第 1 段階において、正解したデータを hard データに変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.19: QQP の第 1 段階で不正解になったデータを hard データに変更した時、正解した問題 (一部抜粋)

QQP の W データを hard データに変更したとき、正解した問題
(変更前) 第 1 段階 質問 1：What is the craziest thing that you ever did in your life with your best friend? 質問 2：What is the craziest thing you have ever done with your friends? モデルの回答： 意味的に違う (正解は意味的に同じ) (変更後) 第 2 段階 (hard) 質問 1：What is the craziest thing that you ever did in your life with your best friend? 質問 2： What is the craziest event you have ever done with your chum? モデルの回答：意味的に同じ (正解は意味的に同じ)
(変更前) 第 1 段階 質問 1：What are some good arguments against the existence of God? 質問 2：What are the scientific arguments against the existence of God? モデルの回答： 意味的に違う (正解は意味的に同じ) (変更後) 第 2 段階 (hard) 質問 1：What are some good arguments against the existence of God?. 質問 2： What are the scientific discussion against the being of God? モデルの回答：意味的に同じ (正解は意味的に同じ)
(変更前) 第 1 段階 質問 1：How do you make a pregnant belly costume?. 質問 2：How do you make an infant Batman costume? モデルの回答： 意味的に同じ (正解は意味的に違う) (変更後) 第 2 段階 (hard) 質問 1：How do you make a pregnant belly costume? 質問 2：How do you create an toddler Batman clothing? モデルの回答：意味的に違う (正解は意味的に違う)

表 5.20: QQP の第 1 段階で不正解になったデータを hard データに変更した時、不正解になった問題 (一部抜粋)

QQP の W データを hard データに変更したとき、不正解になった問題
(変更前) 第 1 段階 質問 1：What was your best practical joke? 質問 2：What is the best practical joke? モデルの回答： 意味的に違う (正解は意味的に同じ) (変更後) 第 2 段階 (hard) 質問 1： What was your best practical joke? 質問 2： What is the best useful pun?” モデルの回答：意味的に違う (正解は意味的に同じ)
(変更前) 第 1 段階 質問 1： Is there any way to get rid of fat soluble drugs without losing weight? 質問 2： Is there any way you can get rid of the fat soluble drugs stored in your fat without weight loss? モデルの回答： 意味的に違う (正解は意味的に同じ) (変更後) 第 2 段階 (hard) 質問 1： Is there any way to get rid of fat soluble drugs without losing weight? 質問 2： Is there any way you can remove of the blubber soluble medicine stored in your fat without lost weight? モデルの回答：意味的に違う (正解は意味的に同じ)
(変更前) 第 1 段階 質問 1：What are the best resources to learn digital signal processing for machine learning? 質問 2：What are the best resources to learn about digital signal processing? モデルの回答： 意味的に違う (正解は意味的に同じ) (変更後) 第 2 段階 (hard) 質問 1： What are the best resources to learn digital signal processing for machine learning?. 質問 2： What are the best resources to take lessons about digital traffic light processing? モデルの回答：意味的に違う (正解は意味的に同じ)

表 5.19 は、第 1 段階で「名詞」の読み取りができなかったため不正解になっている。これに対して、「名詞」を変更することで正解できている。表 5.20 は、QNLI 同様に比較的簡単な単語を同義語に変換しているので正解することが出来なかったと考えられる。

5.2.4 MNLI の考察

表 5.21 と表 5.22 は、MNLI の第 1 段階において、正解したデータを hard データに変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.21: MNLI の第 1 段階で正解したデータを hard データに変更した時、正解した問題 (一部抜粋)

MNLI の C データにおいて、hard データに変更したとき、正解した問題
(変更前) 第 1 段階 文章 1： The red moon made her skin glow. 文章 2： Her skin was glowing from the red moon. モデルの回答： 含意 (正解は含意) (変更後) 第 2 段階 (hard) 文章 1： The crimson horned moon made her skin glow. 文章 2： Her skin was glowing from the red moon. モデルの回答： 含意 (正解は含意)
(変更前) 第 1 段階 文章 1： As a basic guide, the symbols below have been used to indicate high-season rates in Hong Kong dollars, based on double occupancy, with bath or shower. 文章 2： As you can see, the symbols are of dolphins and octopuses.. モデルの回答： 中立 (正解は中立) (変更後) 第 2 段階 (hard) 文章 1： As a basic explanation, the symbols below have been used to indicate peak season rates in Hong Kong dollars,based on a twin room, with bath or shower.. 文章 2： As you can see, the symbols are of dolphins and octopuses. モデルの回答：中立 (正解は中立)
(変更前) 第 1 段階 文章 1： He hadn't seen even pictures of such things since the few silent movies run in some of the little art theaters. 文章 2： He had recently seen pictures depicting those things. モデルの回答： 矛盾 (正解は矛盾) (変更後) 第 2 段階 (hard) 文章 1： He had never seen even pictures of such things since the few silent movies go on in air in some of the little art theaters.. 文章 2： He had recently seen pictures depicting those things.. モデルの回答：矛盾 (正解は矛盾)

表 5.22: MNLI の第 1 段階で正解したデータを hard データに変更した時、不正解になった問題 (一部抜粋)

MNLI の C データにおいて、hard データに変更したとき、不正解になった問題
(変更前) 第 1 段階 文章 1： GAO recommends that the Secretary of Defense revise policy and guidance. 文章 2： GAO recommends that the Secretary of Defense keep policy and guidance the same. モデルの回答： 含意 (正解は含意) (変更後) 第 2 段階 (hard) 文章 1： The Government Accountability Office recommends that SD revise policy and guidance. 文章 2： GAO recommends that the Secretary of Defense keep policy and guidance the same. モデルの回答：中立 (正解は含意)
(変更前) 第 1 段階 文章 1： Tuppence rose.. 文章 2： Tuppence floated into the air.. モデルの回答： 中立 (正解は中立) (変更後) 第 2 段階 (hard) 文章 1： Tuppence rose.. 文章 2： Twopence floated in the air. モデルの回答：含意 (正解は中立)
(変更前) 第 1 段階 文章 1： What's truly striking, though, is that Jobs has never really let this idea go. 文章 2： Jobs never held onto an idea for long.. モデルの回答： 矛盾 (正解は矛盾) (変更後) 第 2 段階 (hard) 文章 1： What struck me the most, though, is that Jobs has never really released this plan. 文章 2： Jobs never held onto an idea for long. モデルの回答： 中立 (正解は矛盾)

表 5.21 では、文章ペアの品詞を簡単な同義語に変更した場合は、正解できていたが、表 5.22 では、細かなニュアンスによる同義語 (前置詞や動詞や省略) により正解することが出来なかった。これは、モデルが適切な機械読解を行えていないためである。

次に、表 5.23 と表 5.24 は、MNLI の第 1 段階において、正解したデータを hard データに変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.23: MNLI の第 1 段階で不正解になったデータを hard データに変更した時、正解した問題 (一部抜粋)

MNLI の W データを hard データに変更したとき、正解した問題
(変更前) 第 1 段階 文章 1： According to a 1995 Financial Executives Research Foundation report,5 transaction processing and other routine accounting activities, such as accounts payable, payroll, and external reporting, consume about 69 percent of costs within finance. 文章 2： Almost 70% of costs within finance are for routine accounting activities.. モデルの回答： 矛盾 (正解は含意) (変更後) 第 2 段階 (hard) 文章 1： According to a 1995 Financial Executives Research Foundation report,5 transaction processing and other usual accounting activities, such as accounts payable, payroll, and external reporting, use up about 70 percent of costs within finance.. 文章 2： Almost 70% of costs within finance are for routine accounting activities. . モデルの回答： 含意 (正解は含意)
(変更前) 第 1 段階 文章 1： He pulled his cloak tighter and wished for a moment that he had not shaved his head.. 文章 2： The man pulled his super hero cape around himself to show off.. モデルの回答： 矛盾 (正解は中立) (変更後) 第 2 段階 (hard) 文章 1： He pulled his cape tighter and hoped for a moment that he had not shaved his head.. 文章 2： The man pulled his super hero cape around himself to show off. モデルの回答： 中立 (正解は中立)
(変更前) 第 1 段階 文章 1： Text Box 2.1: Gross Domestic Product and Gross National Product 48Text Box 4.1: How do the NIPA and federal unified budget concepts of. 文章 2： Text about GBP and USD.. モデルの回答： 中立 (正解は矛盾) (変更後) 第 2 段階 (hard) 文章 1： Text Box 2.1: GDP and GNP 48Text Box 4.1: How do the National Income and Product. Accounts and federal unified estimate concepts of. 文章 2： Text about GBP and USD. モデルの回答： 矛盾 (正解は矛盾)

表 5.24: MNLI の第 1 段階で不正解になったデータを hard データに変更した時、不正解になった問題 (一部抜粋)

MNLI の W データを hard データに変更したとき、不正解になった問題
(変更前) 第 1 段階 文章 1： Tracking down the tiger is a subtle affair, and requires a degree of dedication, calm, and stealth.. 文章 2： You must be very silent when tracking tigers.. モデルの回答：中立 (正解は含意) (変更後) 第 2 段階 (hard) 文章 1： Tracking down the panthera tiger is a subtle matter, and requires a degree of dedication, calmness, and stealth.. 文章 2： You must be very silent when tracking tigers.. モデルの回答：中立 (正解は含意)
(変更前) 第 1 段階 文章 1： Enter the realm of shopping malls, where everything you're looking for is available without moving your car. 文章 2： Everything can be found inside a shopping mall.. モデルの回答： 含意 (正解は中立) (変更後) 第 2 段階 (hard) 文章 1： Enter the area of shopping centers, where everything you're looking for is available without moving your vehicle. 文章 2： Everything can be found inside a shopping mall. モデルの回答： 含意 (正解は中立)
(変更前) 第 1 段階 文章 1： Sir James's presence in Manchester was not accidental.. 文章 2： Manchester was not the place that Sir James had intended to go. モデルの回答： 含意 (正解は矛盾) (変更後) 第 2 段階 (hard) 文章 1： Sir James's presence in Manchester was not accidental. 文章 2： Manchester was not the place that Sir Jamie had planned to go. モデルの回答：中立 (正解は矛盾)

表 5.23 では、第 1 段階で正解できなかった抽象型 (69%と約 70%や GDP や NIPA) の推論において、数値や名称を具体化することで正解に導くことが出来ると考察した。また、表 5.24 では、W データにおいて、単語を同義語に変換しても正解ラベルが反転することはなかった。

5.2.5 SST-2 の考察

表 5.25 と表 5.26 は、SST-2 の第 1 段階において、正解したデータを変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.25: SST-2 の第 1 段階で正解したデータを変更した時、正解した問題 (一部抜粋)

SST-2 の C データを変更したとき、正解した問題
(変更前) 第 1 段階 文章 1：a sometimes tedious film . モデルの回答： 悪い感情 (正解は悪い感情) (変更後) 第 2 段階 文章 1：a sometimes boring film . モデルの回答： 悪い感情 (正解は悪い感情)
(変更前) 第 1 段階 文章 1：a tender , heartfelt family drama モデルの回答： 良い感情 (正解は良い感情) (変更後) 第 2 段階 文章 1：a gentle and soft , sincere family drama . モデルの回答： 良い感情 (正解は良い感情)

表 5.26: SST-2 の第 1 段階で正解したデータを変更した時、不正解になった問題 (一部抜粋)

SST-2 の C データを変更したとき、不正解になった問題
(変更前) 第 1 段階 文章 1：forced , familiar and thoroughly condescending . モデルの回答： 悪い感情 (正解は悪い感情) (変更後) 第 2 段階 (hard) 文章 1： compel , kind and thoroughly condescending . モデルの回答：良い感情 (正解は悪い感情)
(変更前) 第 1 段階 文章 1：this movie is maddening . モデルの回答： 悪い感情 (正解は悪い感情) (変更後) 第 2 段階 (hard) 文章 1： this movie is crazy. モデルの回答：良い感情 (正解は悪い感情)

表 5.25 では、感情表現が明確な文を同義語に変更した時は、正解できているが、表 5.26 のように、良い感情と悪い感情を合わせた文や、「crazy」などの「良い」「悪い」どちらの意味も取れる単語は、間違いやすいと考えた。

表 5.27 と表 5.28 は、SST-2 の第 1 段階において、不正解になったデータを変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.27: SST-2 の第 1 段階で不正解になったデータを変更した時、正解した問題 (一部抜粋)

SST-2 の W データを変更したとき、正解した問題
(変更前) 第 1 段階 文章：you wo n't like roger , but you will quickly recognize him. モデルの回答：良い感情 (正解は悪い感情) (変更後) 第 2 段階 文章：you dislike roger , but you will not slowly recognize him. モデルの回答： 悪い感情 (正解は悪い感情)
(変更前) 第 1 段階 文章：as unseemly as its title suggests . モデルの回答： 悪い感情 (正解は良い感情) (変更後) 第 2 段階 文章：as improper as its title suggests. モデルの回答： 良い感情 (正解は良い感情)

表 5.28: SST-2 の第 1 段階で不正解になったデータを変更した時、不正解になった問題 (一部抜粋)

SST-2 の W データを変更したとき、不正解になった問題
(変更前) 第 1 段階 文章 1：rarely has leukemia looked so shimmering and benign . モデルの回答：良い感情 (正解は悪い感情) (変更後) 第 2 段階 文章 1： rarely has leukemia looked so glisten and gentle. モデルの回答：良い感情 (正解は悪い感情)
(変更前) 第 1 段階 文章 1：if steven soderbergh 's ' solaris ' is a failure it is a glorious failure. モデルの回答： 悪い感情 (正解は良い感情) (変更後) 第 2 段階 文章 1：if steven soderbergh 's ' solaris ' is a miss it is a brilliant miss. モデルの回答： 悪い感情 (正解は良い感情)

表 5.27 の「you won't〜him.」という文章は「否定→逆接 (肯定) →肯定」の構造であり、これによりモデルが最終的に「良い感情」と答えたと考えられるが、第 2 段階で同義語を作成する際の文章は、「否定→逆接 (肯定) →否定」の構造によって最終的に「悪い感情」となって正解したものだと考えられる。

5.2.6 CoLA の考察

表 5.29 と表 5.30 は、CoLA の第 1 段階において、正解したデータを変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.29: CoLA の第 1 段階で正解したデータを変更した時、正解した問題 (一部抜粋)

CoLA の C データを変更したとき、正解した問題
(変更前) 第 1 段階 文章：Bill sent a package to Tom. モデルの回答： 正しい (正解は正しい) (変更後) 第 2 段階 文章：Bill gave a strange package to Tom yestreday. モデルの回答：正しい (正解は正しい)
(変更前) 第 1 段階 文章：Kim put in the box. モデルの回答： 正しくない (正解は正しくない) (変更後) 第 2 段階 文章：Kim put at the box. モデルの回答：正しくない (正解は正しくない)

表 5.30: CoLA の第 1 段階で正解したデータを変更した時、不正解になった問題 (一部抜粋)

CoLA の C データを変更したとき、不正解になった問題
(変更前) 第 1 段階 文章：John went home. モデルの回答： 正しい (正解は正しい) (変更後) 第 2 段階 文章： John fast returned home. モデルの回答：正しくない (正解は正しい)
(変更前) 第 1 段階 文章：No one can forgive that comment to you. モデルの回答：正しくない (正解は正しくない) (変更後) 第 2 段階 文章： No one can forgive that comment for you. モデルの回答：正しい (正解は正しくない)

表 5.29 では、文章を複雑化したり前置詞を意図的に変更してもある 8 割程度の問題に正解できている。しかし、表 5.30 に注目すると、副詞 (fast) や代名詞 (you ではなくて yours)、複雑な間接疑問文の認識が不十分だと考えた。

表 5.31 と表 5.32 は、CoLA の第 1 段階において、不正解になったデータを変更した時、正解した問題、不正解になった問題を表にしたものである。

表 5.31: CoLA の第 1 段階で不正解になったデータを変更した時、正解した問題 (一部抜粋)

CoLA の W データを変更したとき、正解した問題
(変更前) 第 1 段階 文章：Carla slid the book. モデルの回答： 正しくない (正解は正しい) (変更後) 第 2 段階 文章 1：Carla moved the book. モデルの回答：正しい (正解は正しい)
(変更前) 第 1 段階 文章：If I am a rich man, I'd buy a diamond ring. モデルの回答：正しい (正解は正しくない) (変更後) 第 2 段階 文章： If I is a rich man, I'd buy a diamond ring. モデルの回答：正しくない (正解は正しくない)

表 5.32: CoLA の第 1 段階で不正解になったデータを変更した時、不正解になった問題 (一部抜粋)

CoLA の W データを変更したとき、不正解になった問題
(変更前) 第 1 段階 文章 1：Which book's, author did you meet? モデルの回答： 正しくない (正解は正しい) (変更後) 第 2 段階 文章 1： Which book's, author did you encounter?. モデルの回答：正しくない (正解は正しい)
(変更前) 第 1 段階 文章 1：Students studying English reads Conrad's Heart of Darkness while at university. モデルの回答： 正しい (正解は正しくない) (変更後) 第 2 段階 文章 1： Students studying English read Conrad's Heart of Darkness while at university. モデルの回答：正しい (正解は正しくない)

表 5.31 では、動詞を同義語に変換することによる正答率は悪かったが、単語を一つだけ意図的に変更することで間違った文法を抽出することが出来た。(前節 4 章の表 4.8 参照) これに対して表 5.32 は、動詞のみを同義語に変更しても正解ラベルは反転しない。これは、モデルが複雑な文章を読み取ることが出来なかったためであり動詞を変更しても意味はないと考えられる。

第6章　まとめと今後の展望

本稿では、自然言語処理において汎用言語モデルの BERT を派生した RoBERTa に対して、評価データセットの内容を意図的に変更し、敵対的なデータを作成することで RoBERTa の頑健性を検証した。CommonsenseQA では、質問のみを変更したパターンと選択肢のみを変更したパターンに分類して敵対的なデータを作成した。結果として、第 1 段階で正解したデータセット (Accuracy が 100%) を意図的に変更した時の Accuracy は各パターンにおいて、84%、88%とある程度モデルの頑健性があると考えられるが、第 1 段階で不正解になったデータセット (Accuracy が 0%) を意図的に変更した時の Accuracy は各パターンで 24%、40%と半分以下の問題しか正解できていないことが分かった。また、GLUE の各データセットでは、データ内の品詞を一つのみ変更するパターン (easy) と複数変更するパターン (hard) に分類して敵対的なデータを作成した。結果として、第 1 段階で正解したデータセット (Accuracy が 100%) を意図的に変更した時の Accuracy は easy で 84%～100%、hard になると 76%～88%と下降していたが、ある程度モデルの頑健性はあると考えられる。また、第 1 段階で不正解になったデータセット (Accuracy が 0%) を意図的に変更した時の Accuracy は easy で 16%～32%、hard になると 20%～52%と正解したデータセットとは逆に上昇していた。これは、第 2 段階の easy において正解できなかった文章を hard における品詞の複数変更によって正解ラベルが逆転し、Accuracy が上昇したためである。今後としては、敵対的なデータの抽出件数を増やした状態で実験を行うこと、不正解になったデータセットに対して Accuracy の値を向上させる方法を考えること、検証で用いた CommonsenseQA と GLUE 以外のデータセットを用いて実験を行うこと、さらに、RoBERTa 以外でも ALBERT や XLnet のような最新モデルを使って実験を行うことである。

謝 辞

本論文を作成するにあたり、教授の宮森 恒先生には指導教官として本研究の実施の機会を与えて戴き、その遂行から論文の執筆にあたるまで、終始ご指導を戴いた。ここに心より深謝の意を表する。また作業の際に、一人では解決できない問題が多々あり、その一つ一つに丁寧にご指導下さった、京都産業大学大学院先端情報学研究科の木村 輔氏、杉本 翔氏、宮森研究室の岡本 卓氏、冨永 陽羽氏、スウ ハクギ氏に深く感謝する。