令和 4 年度茨城大学大学院理工学研究科情報工学専攻

修士学位論文

BERT を用いた生成文の常識評価

所属　　情報工学専攻

著者　杉本康剛 (21NM731X)

指導教員　　新納浩幸教授

令和 4 年 2 月 3 日 (金)

令和 4 年度茨城大学大学院理工学研究科情報工学専攻　修士学位論文

BERT を用いた生成文の常識評価

著者

杉本康剛 (21NM731X)

指導教員

新納浩幸教授

論文要旨

本稿では、人間が考える常識的な文と非常識な文について BERT を用いることで 2 値分類を行い，システムに与えられた文が常識的な文か非常識な文かの判断を行う．

学習データに常識的な文のデータとして STAIR Captions を 500 件，非常識な文のデータとして自身で作成した非常識な文を 500 件用いた．

推論結果として，全体の正答率は 81.3% であり、常識的な文を正しく判別を行うことができたのが 82.5%．非常識な文を正しく判別を行うことができたのが 80.0% であった．非常識な文に対する分類がわずかに上手く行うことができなかった．また、正しく分類することができなかった非常識な文について４つのカテゴリー分けを行った結果、特に一般常識についての分類が困難であった．作成したモデルは一定の有効性を示す一方で，BERT による常識的な文と非常識な文の判別は単に 2 つのデータを用意して学習を行うだけでは上手く行かず，単語間の概念自体を学習させていく必要があるように思われる．

Master's Thesis in Scholastic 2022, Major in Computer and Information Sciences,
Granduate School of Science and Engineering, Ibaraki University

Common sense evaluation of generated sentences using BERT

Author : Yasutaka Sugimoto (21NM731X)

Adviser : Prof. Hiroyuki Shinnou

Abstract

In this paper, we use BERT to perform a binary classification of what humans consider to be common-sense sentences and what humans consider to be insane sentences , and determine whether a given sentence is sensible or insane for the system.

500 STAIR Captions were used as training data for common-sense sentences, 500 insane sentences were created by the participants as training data for insane sentences.

As a result of inference, the overall correct answer rate was 81.3%, and 82.5% of the respondents were able to correctly identify common-sense sentences. 80.0% of the respondents were able to correctly identify insane sentences. The classification of insane sentences could not be done slightly better. In addition, four categorizations were made for the insane sentences that could not be classified correctly, and the result showed difficulty in classifying the common sense category. The created model showed a certain degree of effectiveness, While it seems that the discrimination between sensible and insane sentences by BERT does not work by simply preparing two sets of data and learning them. so correct classification may require learning the concepts between the words themselves. .

目次

表目次

図目次

第 1 章

序論

1.1 本研究の背景及び目的

近年，Bidirectional Encoder Representations from Transformers(BERT) 等の事前学習済みモデルの発達により文を自動生成する研究が増えてきている. 文の自動生成には様々な種類があり，ニュース記事のような文章やメールをキャッチフレーズのような文を生成するものもある. さらに，対話システムと呼ばれるユーザの問いかけに対して発話を行い，システムとユーザが相互に会話をするシステムも発達している.

文の自動生成の研究が発達するとともに，その文が正しい文なのか，意図した内容通りの文なのかなどを判断する必要性が増してきている. 自動生成された文に大きな矛盾や文法の誤りなどがあると人間が読んだときに違和感のある文となってしまう. そのため，自動生成された文が常識的なのか非常識なのかを人が見る前にシステムが判断をする必要がある. 一般的にニュース記事や小説のジャンルの分類などのカテゴリー分類の研究では，特定単語に着目するものが多かった. また，ニュース記事や小説は数行から数十行までのまとまった文をもとに分類を行うため，カテゴリー分類を行う際の文章の特徴を捉えやすかった. しかし，カテゴリー分類ではなく常識的か非常識な文なのかを判断することはコンピュータには難しく，単語間での素性関係をコンピュータが理解をする必要がある. さらに，写真などに一言文を添えるキャプション文や会話文などの比較的短い文は常識的な文か非常識な文かを判断するには情報量が乏しく，人間のように一見して判断をすることが難しい.

システムに単語の意味理解を学習させるために、従来の研究では単語間の素性関係をまとめたコーパスを使用することが多かった. しかし，単語間の素性関係を詳しくまとめ

たコーパスは英語で作成されているものが多く，日本語で作られているものは確認できなかった. また，非常識な文とはどのような文なのかが定義がなされていないため，分類を行うことが難しい.

そこで，本研究では自動生成された文の中から非常識な文の種類に着目し，カテゴリー分けを行うことで非常識な文を明確にし，常識的な文と非常識な文を判別するシステムを作成する. また，文を自動生成するシステムを実現するために，Text-to-Text Transfer Transformer(T5) を用いて，キャプション文を学習させることで短い文を生成した. 非常識な文のカテゴリー分けは生成された文の中から「構文的誤り」「冗長性」「意味的誤り」「一般常識の欠如」の 4 つのカテゴリーに分類を行う. これにより，この非常識な文とキャプション文から得られた常識的な文から，与えられた文が常識的か非常識的かを判別することを試みる.

1.2　本稿の構成

本稿は第 1 章の序論で本研究の対象となる現在の課題を示し，それに対する本研究で行う解決案を提案した. 第 2 章の関連研究では筆者が参考にしたモデルや類似した研究を紹介し，研究分野について理解を深める. 第 3 章及び第 4 章では筆者の提案するシステムとその目的，概要について説明を行い，課題に対するアプローチを具体的に説明する. 第 5 章，第 6 章では提案したシステムを実際に構築し，その性能を検証する. 実際に得られた実験結果をもとに，なぜその結果が得られたのかを考察する. 第 7 章では本研究で行った内容をまとめて，今後の展望ついての考えを示す.

第 2 章

関連研究

本章では，文章分類において精度の高い分類が可能な BERT, 文生成の分野において広く使われている深層学習のモデルの一種である T5，そしてテキスト分類に関する研究の説明を行う.

2.1 BERT [1]

BERT とは Bidirectional Encoder Representations from Transformers の略称であり,Attention を用いた Encoder-Decoder モデルの Transformer モデル [2] から Encoder のみを使用し, 双方向 Attention を導入したモデルである (図 2.1). 従来のニューラルネットワークのモデルにおいては, 一方の方向からのみ文章を理解していたため, 処理を経るつれ, 初めのトークン情報が薄れていまっていた. しかし,BERT はトークン情報を処理する際に他のトークンを直接参照する. そのため, 離れた位置に存在する情報も適切な形で解釈できるため, 同じ単語であっても文脈を考慮した分散表現を獲得することができる.

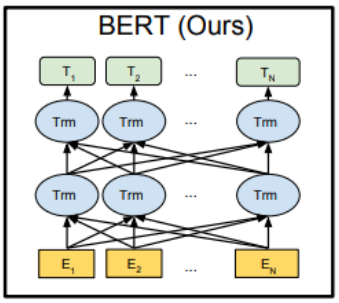

図 2.1: BERT の構造. 図は [1] より引用.

また，BERT の学習や事前学習とファインチューニングの過程に分かれており，事前学習では大量のラベルなしデータを学習させる.BERT ではマスク付き言語モデル (図 2.2) と言われる方法と次文予測 (図 2.3) の二つの方法を組み合わせ学習を行う. マスク付き言語モデルではランダムに選んだ 15% のトークンの内 80% を特殊トークン [MASK] に，10% を他の単語に，残りの 10% はそのままにし，[MASK] で隠された単語を双方向に周りの単語から予測する学習を行う. 次文予測では 2 つの文を用意し，その 1 つの文がもう一つの文に続くかを 2 値分類を予測するトークン [CLS] と終端を表すトークン [SEP] を用いて，予測する学習を行う. この入力はトークン埋め込み，文埋め込み，位置エンコーディング埋め込みをまとめて得られる.

ファインチューニングではラベル付きのデータとタスク内容に応じた分類器を用いて学習させることで特定のタスクに特化する学習を行う. 学習の際，モデルのパラメータの初期値は事前学習で得られたパラメータ，新しく追加した分類器のパラメータをランダムとして，ラベル付きデータを双方のパラメータを学習する. これらにより，ラベル付きデータは従来の方法よりも少ないデータ数でありながら，高い精度を示す.

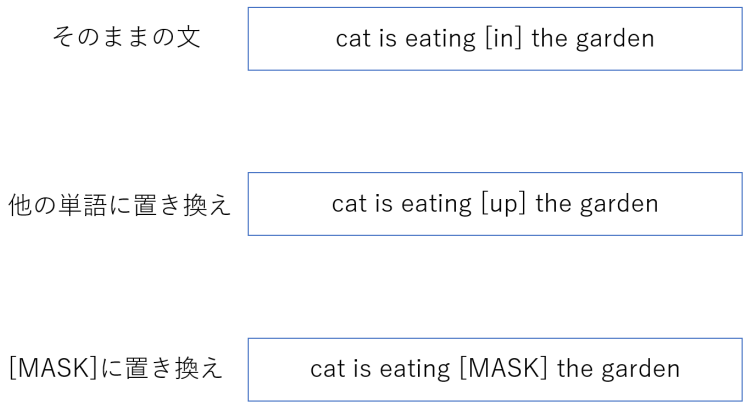

図 2.2: マスク付き言語モデルの例.

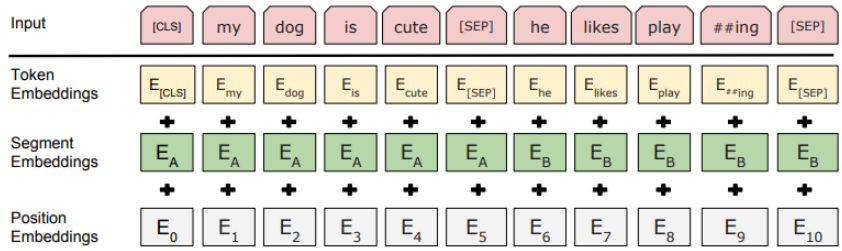

図 2.3: 次文予測の例. 図は [1] より引用.

2.2　T5 [3]

Text-to-Text Transfer Transformer は事前学習モデルの一種で T5 と呼ばれている.T5 も BERT と同様に学習は事前学習とファインチューニングの 2 つに分けられており，似た学習方法を行うが，T5 では事前学習の目的関数に Denoising Objective が用いられており，事前学習データも C4 と呼ばれる世界中の web サーバから得られたページデータをクリーニングした巨大データセットを用いている.T5 では Encoder-Decoder 層を持っており,BERT モデルに Decoder 層が加わった形になっている. 入力系列に「要約」や「翻訳」等の接頭辞を追加することで, すべてのタスクに対して, 一つの入力形式で処理を行うことができる (図 2.4). そのため,Text-to-Text の名前の通り, 入力も出力もテキストがベースとなっており, 文生成や要約といった様々な言語分野に用いられる.

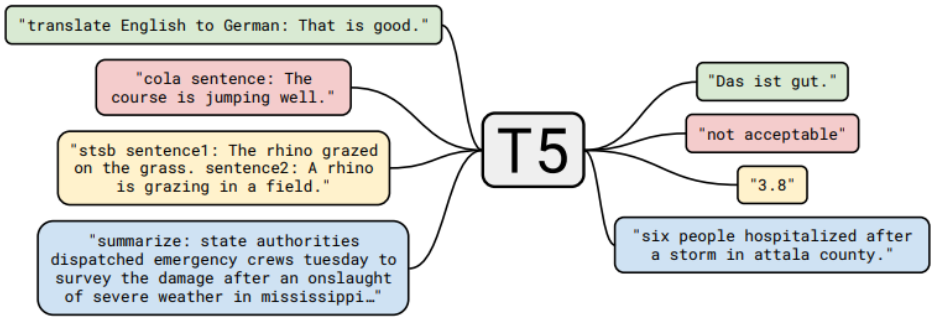

図 2.4: T5 のテキスト変換フレーム. 図は [3] より引用.

2.3　テキスト分類に関する研究

2.3.1　単語拡張によるテキスト分類

鳥山ら [4] は Twitter などのごく短い文書に対してテキストの分類精度を向上させるにあたり，単語拡張と特徴選択の二つの方法を組み合わせた手法を提案している. 単語拡張は短い文章などで文脈情報を読み取ることが困難であるとき，有効な手法であり，単語文書行列 T(TDM) に対して単語間類似行列 S を生成することで，その積で拡張された単語文書行列 TS を生成し，単語情報を補っている. 特徴選択では文書内の有益な単語を厳選し，分類の効率化を図る手法であり，従来の特徴選択法では EBNS に対して，ショートテキストに対応した S-EBNS の定義を行っている.

$$S - EBNS(t_i) = max \mid F^{-1}(tpr_{ck}(t_i)) - F^{-1}(fpr_{ck}(t_i)) \mid . \tag{2.1}$$

Reuters-21578 の R8 データセットと Newsgroups のデータセットを対象に SVM で分類精度を検証した結果，単語拡張の分類精度の向上とともに，データセットの圧縮効果も見られ，また，特徴選択の後に単語拡張を行った実験でも特徴選択で圧縮された入力に対して単語拡張の分類精度の向上を確認していた.

2.3.2　テキスト分類のためのデータセット水増し

テキストの分類手法とは直接関係しないが，テキスト分類のデータセットの拡張に関する研究には以下のようなものがある. テキストデータの水増しに関して，従来手法ではテキスト中の単語を同義語に置換したテキストを生成する [5] というものがあった. それ対して，有田ら [6] は日本語テキストの係り受け先と文節に着目をしている. 例えば，「私は今朝学校に行った」という文を文節で区切り「行った」に係る文節「私は」などの順番を入れ替えた文「今朝私は学校に行った」を生成しても文意が変化しない. これをもとに chABSA データセットを用いてオリジナルの学習データと提案手法で水増しを行ったデータから 2 クラス分類と多クラス分類を行った結果，2 クラス分類ではオリジナルの学習データを用いたベースライン手法の精度が 87.2% であったのに対して，提案手法は 87.1% と分類精度に変化はなく，多クラス分類ではベースラインが 77.1% に対して，提案手法が 81.5% と 4.4 ポイント精度向上が見られている (表 2.1, 表 2.2).

表 2.1: 2 クラス分類の精度 (%)

ベースライン	87.2
提案手法	87.1

表 2.2: 多クラス分類の精度 (%)

ベースライン	77.1
提案手法	81.5

第 3 章

T5 による文生成システム

3.1 背景と目的

本稿の序論でも述べたように，近年，文の自動生成などの技術が発達してきており，様々な場面や状況に合わせた文の自動生成が可能になっている. 例えば，図 3.1 に示すような ELYZA Pencil [7] では，2〜8 個までのキーワードを入力することでニュース記事，メール文，職務履歴書の文章を自動生成することができる.Catchy [8] では，商品の名前や商品の説明文を入力することで，商品のキャッチコピーやユーザレビューなどを生成することができる.

しかし，生成される文が意図した文になるとは限らない. 生成された文の中には人が意味を読み取れないものや矛盾が生じている文なども存在している. そのような非常識な文をシステム側で自動的に判別できることが望ましい. しかし，非常識な文を集めたデータセットの存在は確認できなかったため，自身で作成する必要がある. そのため，実際に機械学習を用いて自動的に生成される文の中から非常識な文を抽出し，それにはどのような特徴があるのかを調査し，カテゴリー分けを行う.

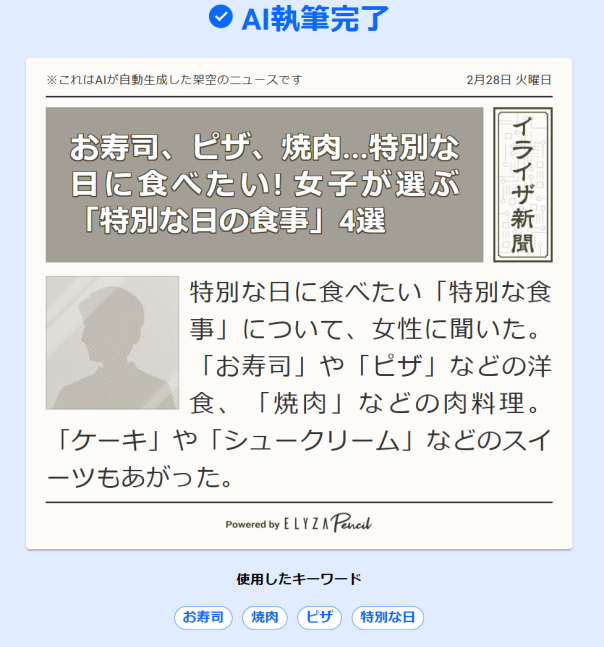

図 3.1: ELTZA Pencil による文章生成の結果

3.2　文生成システムの概要

提案する文の自動生成システムの概要の一例を図 3.2 に示す. 本研究では一文に対して常識的な文であるのか非常識な文であるのかを分類することを目的としているため，生成される文は短文で，人が一見した際に判断することできるのが好ましい. そのため，本システムはキーワードを入力することで，そのキーワードを含む文の生成を行う. キーワードは名詞及び動詞のみを対象とすることで，キーワードから人が推測しやすく，直観的にも常識的か非常識かを判断しやすい文が生成されること目指す.

また, 本研究では日鉄ソリューションズ株式会社が開発し, 公開している大規模な日本語コーパスで事前学習された T5 を生成モデルとして利用した.

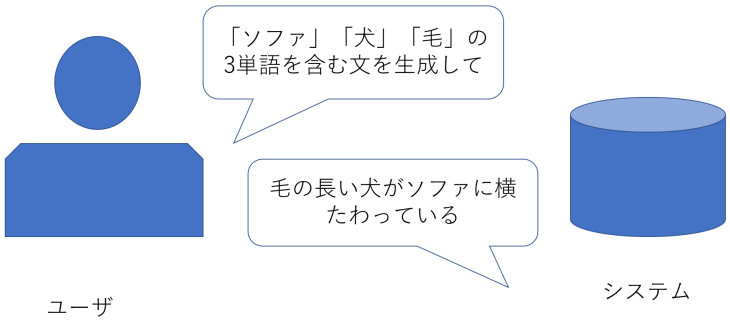

図 3.2: 提案する文生成システム

第 4 章

BERT による 2 値分類

4.1 提案モデル作成の流れ

3 章で提案した手法により生成した文から非常識な文を抽出し，分類器を作成する．
以下に提案する分類器を作成する大まかな流れを示す．

1. 生成された文の中から非常識な文を抽出し，カテゴリー分けを行う．
2. カテゴリーをもとに非常識な文を手作業で作成する．
3. キャプション文から得た常識的な文と作成した非常識な文でデータセットを作成する．
4. 作成したデータセットを用いて 2 値分類を行う BERT を作成する．

4.2　常識的な文と非常識な文

T5 を用いて自動生成される文の多くは常識的な文になると考えられる. そのため，生成された文から 200 件を抽出し，常識的な文か非常識な文かを判断する. 非常識な文かを判断する際には，文法的な誤りや語彙の意味などを考慮し，いくつかのカテゴリーに分類を行う. 誤りが明らかでない時は常識的な文とする. 例えば,「男性が海に泳いでいる」という文は正しくは,「男性が海で泳いでいる」もしくは「男性が海を泳いでいる」であるが,「男性が海に泳いでいる」でも十分に人は文意を読み取ることができる. このような文は常識的な文とする.

4.3　分類器のデータセット作成手順

また，自動生成される文には何文の非常識な文が存在するか分からず，入力文に使用するキーワードによっては生成される非常識な文の種類に大きな差が出ることが考えられる. そのため，200 件の中からカテゴリー分けをした結果から，非常識な文のデータセット作成は筆者自身が手作業で行う. 非常識な文はカテゴリーごとに作成し，データセットとしては非常識な文として一つのラベルにまとめる. その後，常識的な文とともに BERT をファインチューニングし，2 値分類を行うシステムを開発する (図 4.1).

また本研究では東北大学の研究チームによって作成された BERT の日本語事前学習モデルを使用する.

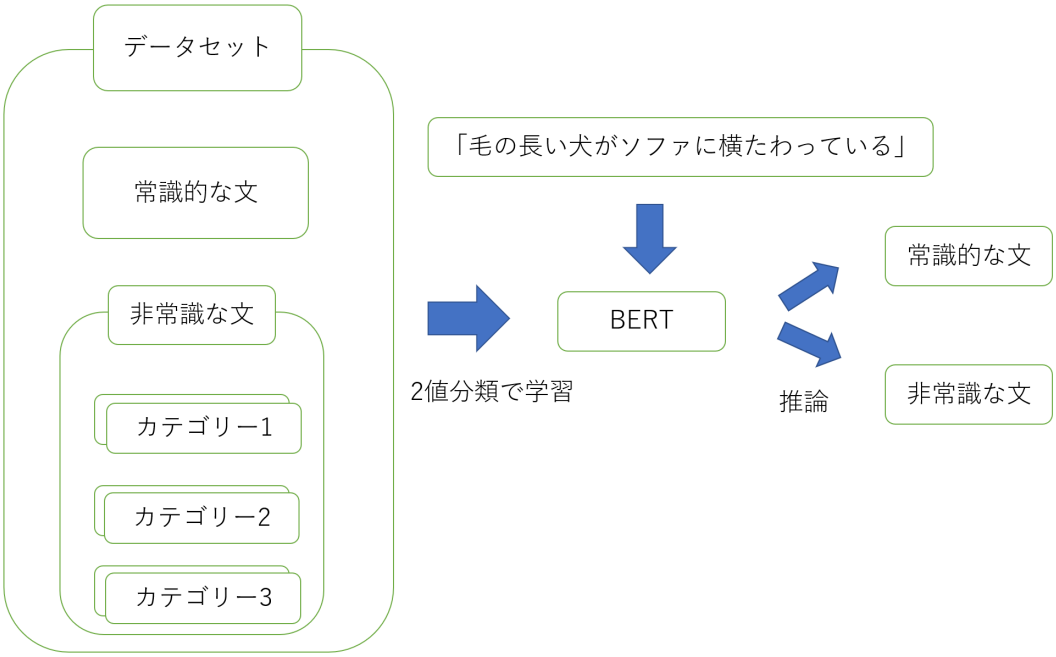

図 4.1: 提案する分類システム

第 5 章

実験

第 3 章及び第 4 章で提案した手法を用いて実験を行う.

5.1 実験の目的

実験の目的は，以下の 2 点を検証することである.

１．提案システムにより常識的な文と非常識な文をどの程度分類できているか
２．分類ができなかった文についてどのような特徴があるか

前者については，キャプション文を利用した常識的な文と自身で作成した非常識な文から 2 値分類で分類を行う. 後者については，2 値分類を行った中で，誤答をした非常識な文のカテゴリーを判断する

5.2 実験の手順

まず T5 で生成された文の中から非常識な文を抽出し，その特徴ごとにカテゴリー分けを行う. その後，筆者自らが手作業でカテゴリー分けした非常識な文を生成し，常識的な文と非常識な文から 2 値分類を行う BERT の分類器を作成する. また，比較用として非常識な文のカテゴリーを考慮せずに学習をさせた同様の分類器を作成し，両モデルの性能を検証する. 最後に，非常識な文に対してどのカテゴリーの分類がよいスコアであったのか，もしくは悪いスコアであったのかを確認することで提案モデルの考察を行う.

5.3　T5 による文生成

5.3.1　文生成データセットの作成

第 3 章で述べたように非常識な文を集めたデータセットの存在が確認できなかったため，自動生成された文の中でどのような文が存在するのかを確かめ，自身で作成をする必要がある. そのため，初めに T5 で文を自動生成するためのデータセットを作成する．文生成を行うために STAIR Captions [9] からキャプション文を用いた. 大まかな手順を図 5.1 に示す. 入力文となるキーワードは 2〜5 件の名詞もしくは動詞とするため，2〜5 件の名詞もしくは動詞を含むキャプション文を 10 万件抽出した. キーワードを抽出するためには辞書内包の形態素解析器 janome を用いた. 抽出したキーワードはランダムに並び替えることで，生成される文に含まれる指定した単語が指定した順番に含まれることを防止する.

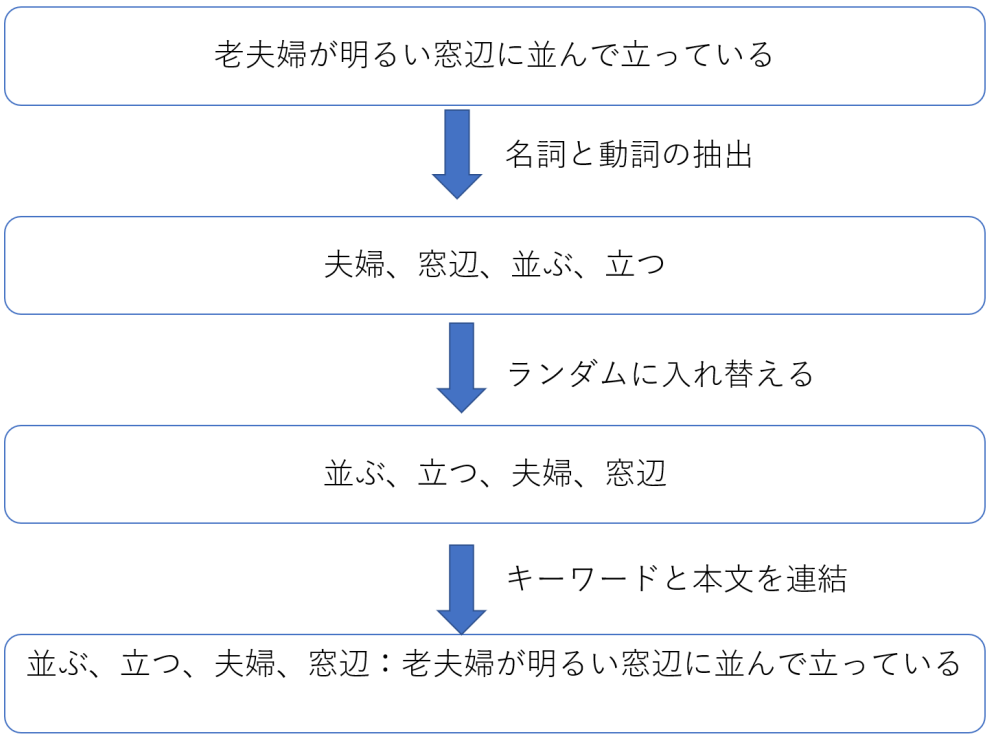

図 5.1: 文生成のデータセット作成

5.3.2　T5 の学習

作成データセットを学習データ，検証データ，訓練データに 9:0.5:0.5 で分割を行い，文の自動生成には日本語版 T5 を生成モデルとして用いて，パラメータは以下のように設定した．

- BatchSize：16.
- Epoch 数：5.
- 学習率：3e-4.
- 最適化関数：AdamW.

また，入力は 2〜5 個のキーワード，出力は短い文のため，入力最大トークン数と出力最大トークン数はそれぞれ，32 と 128 とテキスト生成のタスクとしては少ない数値とした．これらのパラメータを用いて学習を行った (図 5.2)．

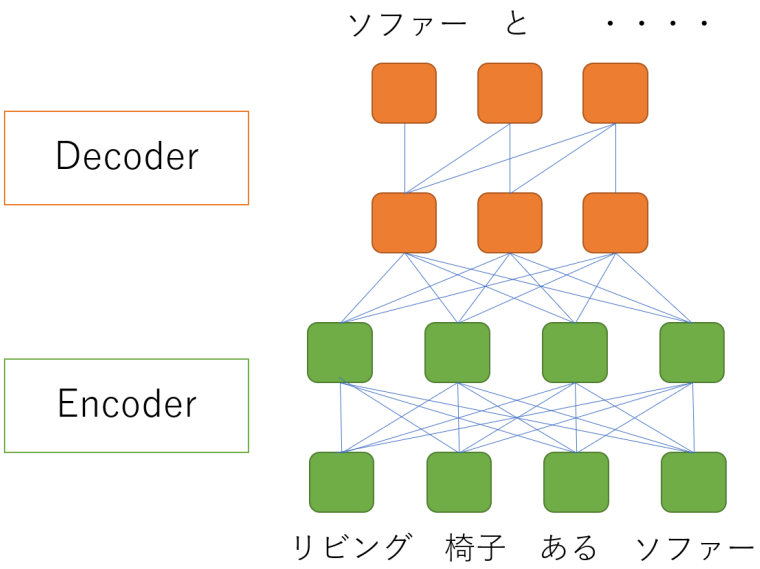

図 5.2: T5 の学習

5.3.3　文生成の結果

各入力文に対する T5 が出力した結果の一部を表 5.1 に示す. 入力文として，指定したキーワードが含まれており，短い文を生成することができた.

一方で入力文「乗る, ジャンプ, 人」から「たくさんの人がジャンプに乗っている」など, 文の意味や意図を読みとることが困難である非常識な文と考えられる文も一部生成されていた.

表 5.1: 生成文の一例

入力文	生成文
バス, 事故, 起こす	大きなバスが事故を起こしている
林, 日差し, 差し込む	林の中に日差しが差し込んでいる
テント, にらみ合う, 女性	テントの中で女性がにらみ合っている
像, 見る, 子供	小さい子供が像を見ている
乗る, ジャンプ, 人	たくさんの人がジャンプに乗っている

5.4　BERT によるカテゴリー分類

5.4.1　カテゴリー分類

5.3 節で文生成の結果とともに少数ではあるが非常識な文が生成されていることを確認できた. これら非常識な文の特徴を考え，カテゴリー分けすることで分類器の作成を試みる.

カテゴリー分けには東中ら [10] の分類を参考にし,「構文的誤り」「冗長性」「意味的誤り」「一般常識の欠如」の 4 つのカテゴリーに生成された非常識な文を分類した.

カテゴリーの意味は以下の通りである

1. 構文的誤りは主に助詞や必須格の欠如により文の意味が読み取れないもの.
2. 冗長性は同じ言葉の繰り返しなど文が意味もなく長くなっているもの.
3. 意味的誤りは単語の意味を誤って解釈しているために文の意味が読み取れないもの.
4. 一般常識の欠如は矛盾のある文や人の常識から外れているもの.

また表 5.2 には 4 つのカテゴリーの例文と表 5.3 は生成文 200 件に対して生成された 4 つの非常識な文のカテゴリーの数を示す.

表 5.2: カテゴリーと例文

カテゴリー	例文
構文的誤り	鏡に映された料理が見とれている
冗長性	自分の自慢の家宝は家宝である
意味的誤り	線路の上を黒い煙が走っている
一般常識の欠如	ゲレンデで一人の人がサーフィンをしている

表 5.3: 生成文 200 件に対する各カテゴリーの文の数.

	構文的誤り	冗長性	意味的誤り	一般常識の欠如
文の数	8	2	08	2

5.4.2　分類器の作成

これらの非常識な文のカテゴリーに対して筆者自ら文の作成を行い,「構文的誤り」を 220 件,「冗長性」を 65 件,「意味的誤り」を 220 件,「一般常識の欠如」65 件を作成した. 作成した非常識な文から各カテゴリーから 10 件ずつをテストデータとしてそれ以外を学習データと検証データとして使用した. また，常識的な文として STAIR Captions からキャプション文を 570 件抽出し，テストデータとして 40 件それ以外を学習データと検証データとした (表 5.4).

表 5.4: データセットの各文数.

	常識的な文	非常識な文			
	キャプション文	構文的誤り	冗長性	意味的誤り	一般常識の欠如
学習データ	500	200	50	200	50
検証データ	30	10	5	10	5
テストデータ	40	10	10	10	10

5.4.3　比較モデルの作成

また，提案したモデルに対して，カテゴリーを考慮しない非常識な文を学習データとして用いた比較モデルを構築した. 図 5.3 に示すように STAIR Captions からキャプション文の名詞を入れ替えることで非常識な文を作成し，常識的な文 1300 件，非常識な文 1300 件を学習データとした比較モデルの BERT の分類器を作成した.

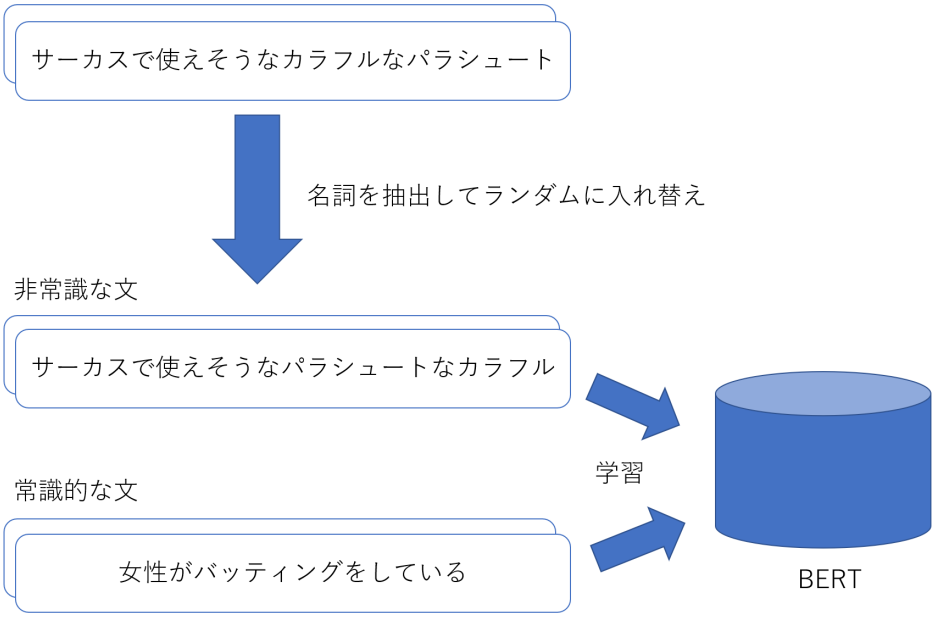

図 5.3: 比較モデルのデータセット

5.4.4　分類結果

表 5.5 は提案モデルと比較モデルのテストデータに対する BERT の分類結果の適合率，再現率，F 値を示したものである. また，提案モデルの正答率は 81.3%，比較モデルの正答率は 66.3% となった. 提案モデルの再現値から常識的な文に対する正答率が 82.5%. 非常識な文に対する正答率が 80.0% であった.

提案モデルを比較モデルと比較をすると正答率や F 値は上回っており，誤答率も低かった. 実際に生成文から非常識な文を抽出し，その特徴ごとにカテゴリー分けを行った非常識な文のデータセットの方が正確な分類を行えると言える.

また，表 5.6 は非常識な文のカテゴリーごとの正答率を表しており，「冗長性」に関する非常識な文の分類は 90% と高く，同じ単語や説明が続く文を高い精度で分別を行えた。一方で，「一般常識の欠如」が 70% と一番低い結果になっており，文内の矛盾を読み取ることができなかった.

表 5.5: テストデータに対する性能評価の結果 (%).

	提案モデル		比較モデル	
	常識的な文	非常識な文	常識的な文	非常識な文
適合率	80.4	82.0	59.7	100
再現率	82.5	80.0	100	32.5
F 値	81.4	81.0	74.8	49.1
全体の正答率	81.3		66.3	

表 5.6: 各カテゴリーに対する正答率 (%).

	構文的誤り	冗長性	意味的誤り	一般常識の欠如
正答率	80.0	90.0	80.0	70.0

第 6 章

考察

6.1 提案モデルの比較

性能評価の結果から非常識な文のカテゴリーを考慮した提案モデルの方がカテゴリーを考慮しない比較モデルよりも精度が高いことがわかる. 非常識な文のカテゴリーを考慮した提案モデルはカテゴリーを考慮しない比較モデルの非常識な文に対する適合率や常識的な文に対する再現率は 100% であったが，常識的な文に対する適合率は 59.7% で非常識な文に対する再現率は 32.5% であったことから常識な文の分類は正確に行えているが，非常識な文の分類が正確ではなく，多くの非常識な文を常識的な文と判断している.

表 6.1 に示す両モデル構築に作成した非常識な文の一部を示す. 自動された生成文に存在した非常識な文やそれを参考に筆者が作成した非常識な文は文の一部のみに違和感や誤りのある文がほとんどであった. しかし，比較モデルに使用した非常識な文は常識的な文の名詞をランダムに入れ替えているため，一文に多くの誤りが存在した. そのため，カテゴリーを考慮しない比較モデルは多くの文を常識的な文と捉えたが，カテゴリーを考慮した提案モデルは正確に判断ができたと考えられる.

表 6.1: 非常識な文比較

カテゴリーを考慮した非常識な文	カテゴリーを考慮しない非常識な文
座っている牛で男性を見ている	カーテンがおいてある猫にふせをしている
ホットドッグに調味料をつけて子供	テニス性 2 人が白い服の男をしている
お風呂の中は水を貼っている男性	少年を滑っているスケボーをしている壁
芸人が手足を法律で縛っている	豚に飛ばされてしまった風船の空

6.2　提案モデルの性能について

性能評価の結果から提案モデルは比較モデルより高い分類精度であったが，すべての指標で 90% を超えるものはなく，「一般常識の欠如」というカテゴリーの判別精度も低かった.

正確に分類を行えなかった原因は以下のことが考えられる.

1. カテゴリー分けが不十分.
2. 学習に用いたデータセットが不十分.
3. 単語間の関係性への理解不足.

6.2.1　カテゴリー分けが不十分

今回非常識な文をカテゴリー分けするにあたり 4 つのカテゴリー「構文的誤り」,「冗長性」,「意味的誤り」,「一般常識の欠如」に分類したが，T5 により自動生成された文にある非常識な文には複数のカテゴリーに属すと考えられる文も存在した (図 6.1). そのため，分類の難しいカテゴリーが現れたと考えられる. そのため，カテゴリー分類を用いて非常識な文を分類する場合，カテゴリーを細分化やマルチラベルでの分類を行うなどの検証も行っていく必要がある.

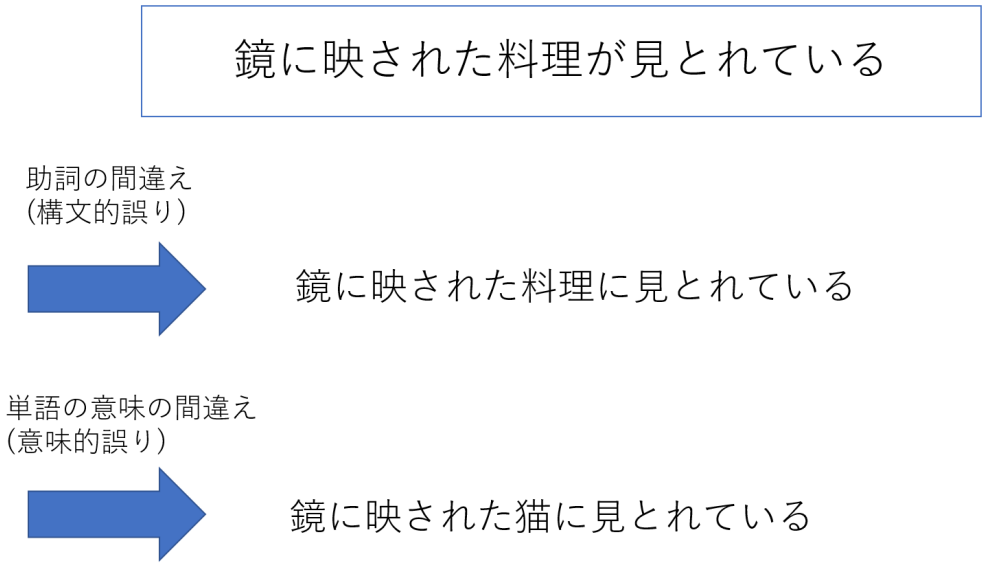

図 6.1: 複数のカテゴリーに属する例

6.2.2　学習に用いたデータセットが不十分

提案モデル用に作成したデータセットは常識的な文 570 件と非常識な文 570 件で構成されている. さらにこのデータセットから学習データ，検証データ，テストデータに分割をするため，実際に学習に用いたデータ数はそれぞれ 500 件ずつとなっている. 学習データの少なさが，正答率などの精度を低下させていると考えられる. 一般的に機械学習において，学習用のデータが多いほど精度の高い結果が得られる. そのため，精度向上のためにはデータセットの数を増やす必要がある.

常識的な文は STAIR Captions からキャプション文を利用したため，データを増やすことは容易であるが，非常識な文は筆者が手作業で作成したため，データ件数の増加は難しい. そのため，非常識な文を自動生成するというシステムを開発する必要がある.

6.2.3　単語間の関係性への理解不足

カテゴリーごとの正解率から「一般常識の欠如が」一番低く 70%,「意味的誤り」も 80% であることから，単語そのものの意味理解や単語間の関係性への理解が不足していると考えられる. 例えば「丸太を沢山積んで体力を回復しているランナー」のように人が読めば，丸太を運ぶことは体力を回復することにはつながらない. そのため ConceptNet [11]

とよばれる単語間の関係性をまとめたようなデータセット (図 6.2) とともに，「丸太を持つ=疲れる」のような，一文を更に小さく分割して常識を学習させる手法が検討される．

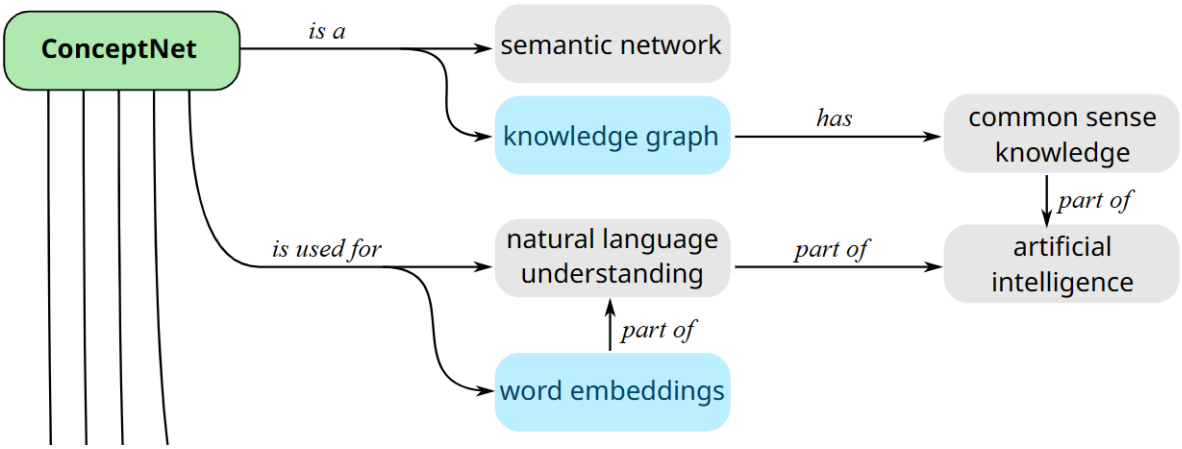

図 6.2: ConceptNet の概要一部. 図は [11] より引用

第 7 章

結論

7.1 まとめ

本稿では常識的な文と非常識な文の分類を行うことを目的として与えられた文が常識的か非常識なのかを判別するシステムを開発した. 本システムを開発するにあたって以下に示す内容に取り組んだ.

1. 文の自動生成のためのデータセット作成.
2. 自動生成された文より非常識な文の抽出.
3. 非常識な文のカテゴリー分類.
4. 常識的な文とカテゴリー分けをした非常識な文を用いた分類器の作成.
5. カテゴリーを行った分類器とカテゴリーを考慮しない分類器の比較実験.

まず，STAIR Captions のデータセットを用いることで写真に対して付与されているその情景を表すキャプション文を 10 万件抽出した. 抽出したキャプション文からそれぞれの文に対して 2〜5 個の名詞もしくは動詞を抽出しキーワードとすることで，それらの単語ともとのキャプション文をそれぞれ入力と出力としたデータセットを作成した.

次に，作成したデータセットをもとに T5 を用いた文の自動生成システムを作成した. このデータセット用いて学習した文生成システムはいくつかのキーワードを入力とすることで，そのキーワードを含む文を生成することを目的とした. 実際に生成された文は指定したキーワードを含まれていることが確認できた. また，学習データにキャプション文を用いているため，生成される文の特徴は短い文かつ情景を説明している文であった.

生成された文の中には人が読むと意味が分からないものや冗長性があり読みづらい物

が存在した. このような文を非常識な文として考えることで，生成文 200 件を調査してみたところ 20 件の非常識な文が確認できた. これらの非常識な文について特徴を考えてカテゴリー分けを行った.

カテゴリー分けを行う際の基準として，文法的な誤り，言い回しの誤り，単語の意味の誤り，常識の誤りの 4 つの観点から考え，カテゴリー分けを検討した. カテゴリー分けの結果，主に助詞の誤りや必須格の欠如により文の意味が読み取れない「構文的誤り」，同じ言葉の繰り返しなど文が意味もなく長くなっている「冗長性」，単語の意味を誤って解釈しているために文の意味が読み取れない「意味的誤り」，は矛盾のある文や人の常識から外れている「一般常識の欠如」の 4 つのカテゴリーに非常識な文を分類した.

20 件あった非常識な文を 4 つのカテゴリーに分類した結果,「構文的誤り」は 8 件,「冗長性」は 2 件,「意味的誤り」は 8 件,「一般常識の欠如」は 2 件存在していることが分かった.

常識的な文と非常識な文の分類を行うためのシステムを開発するために，4 つのカテゴリーに分けた非常識な文をもとに手作業で非常識な文を作成した. それぞれのカテゴリーは生成文 200 件あたりの存在数を考慮して「構文的誤り」を 220 件,「冗長性」を 65 件,「意味的誤り」を 220 件,「一般常識の欠如」65 件で合計して 570 件を作成した. 作成した非常識な文と STAIR Captions のキャプション文を利用した常識的な文を用いて，BERT に学習をさせ，常識的な文と非常識な文の分類を行うシステムを開発した.

非常識な文をカテゴリー分けしたデータセットで学習をさせた BERT の分類器と比較をするために，非常識な文の特徴を考慮せずに，STAIR Captions のキャプション文の名詞をランダムに入れ替えた非常識な文と常識的な文のデータセットを作成し，同じく常識的な文と非常識な文の分類を行う分類器を構築することで，非常識な文のカテゴリーを考慮した提案モデルとカテゴリーを考慮しない比較モデルの 2 つの分類器による比較実験を行った.

実験には 40 件の常識的な文と 40 件の非常識な文の合計 80 件のデータを用いた. 分類の精度を検討する指標として，正答率，適合率，再現率，F 値を求めた. 非常識な文のカテゴリーを考慮した提案モデルでは正答率が 81.3% とカテゴリーを考慮しない比較モデルの正答率 66.3% を大きく提案モデルが上回った. 提案モデルは常識的な文も非常識な文もすべての指標で 80% を超えているため，両方の文において分類が行えていることがわかった. また，比較モデルでは非常識な文に対する適合率と常識的な文に対する再現率

が 100% であるが，常識的な文に対する適合率が 0.597%，非常識な文に対する再現率が 32.5% と多くの文を常識的な文と判断してしまい，正しく分類ができていないことがわかった.

最終的に非常識な文のカテゴリーを考慮した提案モデルの方がカテゴリーを考慮しない比較モデルよりも精度が高く，カテゴリー分類の有効性を示すことができた.

7.2　今後の課題

今回 4 つのカテゴリーに非常識な文を分類し分類器をすることで，カテゴリー分類を行わなかった分類器よりも精度が高いことを示すことができた. しかし，作成した非常識な文は手作業で作成したため，学習データやテストデータに使うデータが多くない. そのため，非常識な文の作成にかかるコストを少なくするために，カテゴリーに沿った非常識な文を自動で作成システムを構築する.

また，非常識な文のカテゴリーを考慮した提案モデルは全ての指標で 80% を超えたが，90% を超えるものはなかった. そのため，分類が困難であった非常識な文のカテゴリー「意味的誤り」や「一般常識の欠如」を正確に分類するために ConceptNet のような単語の意味や単語間の関係をまとめたデータセットを作成して，分類精度の向上を目指す.

謝辞

本研究を進めるにあたり，終始丁寧なご指導とご鞭撻を賜りました，茨城大学 工学部 情報工学科 新納浩幸 教授に深く感謝致します.

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”,arXiv:1810.04805.

[2] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N.Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,”,Pro ceedings of the 31st International Confere.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P.J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” Journal of Machine Learning Research, vol.21, no.140, pp.1–67, 2020.

[4] 鳥山修平, 世木博久, “単語拡張によるテキスト分類精度の改善と評価” , 第 83 回全国大会講演論文集 2021 (1), 533-534, 2021.

[5] Jason Wei, Kai Zou, “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks”, arXiv 190111196.

[6] 有田郎人, 駒井雅之, 佐藤大輔, 丸古凌介, 大木環美, 野村侑司, 田邊豊, 平博順, “テキスト分類学習における文節入れ替えによるデータ水増し手法” , 人工知能学会全国大会論文集 JSAI2021 (0), 3J1GS6a04-3J1GS6a04, 2021.

[7] 文章生成 AI ELYZA Pencil, https://www.pencil.elyza.ai/drafts/16994779071596740647 (参照日 2023-1-26).

[8] Catchy Writing Assistant, https://lp.ai-copywriter.jp(参照日 2023-1-26) .

[9] STAIR Captions, http://captions.stair.center/ (参照日 2023-1-28).

[10] 東中竜一郎, 荒木雅弘, 塚原裕史, 水上雅博, “雑談対話システムにおける対話破綻を生じさせる発話類型化” , 自然言語処理/29 巻 (2022) 2 号 p. 443-466.

[11] ConceptNet, https://conceptnet.io/ (参照日 2023-1-29).