Transformer モデルに関する順序数の的確な把握と活用能力の調査

小川　志龍† 　林　寛治† 　宮森　恒†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g1953319,g2054056,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では、Transformer に基づくモデルが、順序数の概念を的確に把握し活用する能力をどの程度有するのかについて調査する。Transformer に基づく事前学習済みモデルは、四則演算等の単純な算術問題では比較的高い正答率を出す一方、数の概念をどのように捉え、どのように活用しているのかについては不明な点も多い。本研究では、数の概念の一つである順序数に着目し、Transformer モデルが順序数の概念を把握し活用する能力について調査する。具体的には、順序数を用いた的確な数え上げが必須となるタスクをプロービングタスクとして課すことで、順序数に対するモデルの把握・活用能力を分析する。本タスクにより、Transformer モデルの数の理解度に対する知見を深めることができると期待される。

キーワード　順序数, 概念理解, Transformer, 数え上げ, 推論

1　はじめに

Transformer に基づく言語モデルでは、四則演算等の単純な算術問題において高い正答率を出している。そのため現在はより複雑な算術問題の解決に向けて取り組まれている [1]。しかし、数の概念をどのように捉えているのかについては不明であり、高い正答率を出している単純な算術問題においても数の概念を理解した上で解いているかどうかは不明である [2]。平井らの論文 [5] では、基本的な数としての自然数の概念の捉え方について述べられており、自然数の概念は「順序数・基数・数量」の 3 つの概念に分けて捉えることができるとされている。順序数は文字通り順序を表す数であり、1 番目,2 番目, 3 番目... のような数え上げにより”何番目か”を表す。基数は個数として捉える数であり、ものの集まりの大きさを表す集合数とも呼ばれる。数量はある測定単位のいくつ分かとして捉える数であり、単位を伴った測定結果としての量を表す数である。これら基本的な数の捉え方にはそれぞれ違った特徴を持っているが、順序数と基数は小学校低学年では混同されることも珍しくない。

例えば、1 つの列に人が何名か並んでいた場合、順序数は「前から 3 番目の人」、基数は「この列には 3 人並んでいる」というように用いられる。順序数である「前から 3 番目」は 3 番目の人のみが対象となるため、順序を考慮した適切な数え方を行わないと特定することができない。一方で基数である「この列には 3 人並んでいる」は列という集合の大きさを表すため、どのような数え方をしても 3 人という数に変わりはない。

質問応答に関する従来研究では、基数や数量に関するものが多く、順序数に関する問題はほとんど取り組まれていない。基数は問題文中に出現する数値をそのまま使用することで計算可能である場合がほとんどであるが、順序数はそうではなく、数え上げや基数に変換するという処理が必要とされることが多く、順序情報を的確に把握していなければ順序数に関する問題は解くことができない。そのため、順序数に関する問題を解くことができれば、自然数を何らかの実体と対応づけながら数え上げることができている可能性があり、順序数の概念を人間と近い形で理解しているのかどうかに迫るための新たな知見が得られるのではないかと考えた。

上記を踏まえ、本研究では Transformer に基づくモデルが、順序数の概念を的確に把握し活用する能力がどの程度あるのかを明らかにすることを目的としたプロービングタスクを提案する。具体的には、順序数を用いており、正答するために数え上げや基数への変換が必要とされるような質問を与え、その正答を出力させるようなタスクを提案する。これにより言語モデルは数値推論問題において数概念を理解して解いているのかが明らかになり、今後の数値推論問題解決のアプローチや Transformer モデルの数の理解度に関する考えの手がかりとなることが期待される。

実験では、四則演算などの単純な問題から数列や比較、確率、微分積分などの複雑な問題まで幅広い分野の問題が含まれているデータセットである mathematics dataset[8] で学習した Transformer モデル [10] に対し、新たに作成した順序数に関する問題をを与え、その正解率を確認した。

その結果、単純な順序を問う問題では高い精度を示す一方で逆からの順序など少し複雑な順序を問う問題では低い精度となった。2 つの順序情報から間の数を問う問題では問題文中に出現する順序数をそのまま計算していることが明らかになり、モデルは順序数の概念を理解して推論を行うことは困難であり、課題が多いことが明らかになった。

本論文の構成は次のとおりである。第 2 章では関連する研究について記載し、現状の問題点や提案手法との違いについて述べる。第 3 章では提案手法について述べ、第 4 章では提案タスクを用いた実験を行い、その結果の考察を行う。

2 関連研究

2.1 数値推論問題解決に関する研究

言語モデルには数値推論など、数値を扱う能力があることはこれまでの研究で知られている。また、比較的難易度の低い単純な問題に関して高い精度を示しているため、現在はより難易度の高い問題の解決に向けた研究が広く取り行われている。

Hendrycks ら [1] は代数、数論、確率、幾何学、中級代数、微分積分の 7 分野からなる問題 12,500 問を含んだデータセットである”MATH データセット”を提案している。これは数学オリンピックなどで扱われる難易度の高い問題が主となっている。GPT-2 と GPT-3 に対してこれらの問題を与えた結果、最も簡単な問題 (レベル 1) で約 15%, 難しい問題 (レベル 5) で約 4%と低い結果となっていることが示されており、難易度の高い問題に関しては未だ未解決とされている。

このように難易度の高い問題の解決に向けた取り組みがなされているが、たとえ高い精度を示しているモデルであっても数値を理解して解いているのかという点については不明であり、疑問視している研究も存在する。Arkil ら [2] は現時点で言語モデルは初等数学の文章問題ですら解くことができていないのでは無いかと述べている。文章問題は世界状態を説明する文章とその状態に関する未知の情報を問う文章から構成されており、前者の情報を用いて後者の情報を求めるという動作が求められる。しかし、現在高い精度を示しているモデルはこのような文章問題において質問文が削除されていても 77.7%の問題は正解してしまうということが示された。このことから、現在数値推論文章問題において高い精度を示している言語モデルは単純なヒューリスティックに依存しており、数値を理解して解いているとは言えないのではないかと述べられている。

数値の概念理解に関する研究やモデルの内部状態を分析している研究も存在する。Naik らの研究 [3] では単語埋め込み (skipgram, FastText, GloVe) が数値の性質、中でも大きさの愛念をどの程度捉えているかを検証している。単語埋め込みが大きさを表現できるかどうかの実験を行った結果、数の大きさの概念をおおよそ捉えることができていることが示された。しかし、「3=three」などの数詞の概念については不十分であり、数を正確に扱うためには注意が必要であると述べられている。

青木らの研究 [4] では、内容に応じて層数を変化させる PonderNet モデルに対して多段の推論が必要とされるタスクを与え、モデルの動作を確認し、層数が一定な Transformer モデルとの比較を行っている。実験の結果、簡単な演算に対する正答率は 2 つのモデルで大きな差は確認できなかった。しかし、PonderNet が問題に応じて再起回数を変化させているということが示されている。

2.2 Transformer の能力を調査した研究

松本らの研究 [6] では Transformer の再起的構造を把握する能力を調査している。$(a+b) \times (c+d)$ のような四則演算を用いた問題を与えた際、Transformer はこの式における a+b や c+d のような計算の途中結果を内部的に保存して解いているのかどうかを調査した結果、Transformer は計算の途中結果を内部的に複数の成分に分散して保存していることが示されている。

Johnson らの研究 [7] では数値に関する Transformer の構成的推論能力を評価することを目的としたプロービングタスクを提案している。英語、デンマーク語、日本語、フランス語の 4 つの言語で文法性判断タスクと大小比較の合計 2 種類の分類タスクを作成し、BERT, DistilBERT, XLM に与えている。その結果、2 つのタスク両方で 95%以上の高い精度を示し、モデルは英語、デンマーク語、日本語、フランス語において数詞の生成規則をある程度理解していることが示された。一方で大小比較タスクでは全体的に数の大きさを高精度で比較するには不十分であることが示された。このことから、数詞の文法性判断の能力は有しているが、値の比較の能力は不十分であり、数詞の構成要素がどのように意味を形成しているのかといった内容の理解は難しいことが明らかになった。

Rodrigo らの研究 [8] では数の表面的な状態が Transformer の算術問題の学習にどのような影響を与えるかを調査している。足し算と引き算の 2 つのタスクをモデルに与え、その回答を評価しており「12」のようなアラビア数字、「1 2」のように空白で区切られた数字、「0 0 1 2」のように桁数を統一したもの「0_1_2」のようにアンダースコアで区切ったもの、「twelve」のように文字で数字を表したもの、10 進数、指数で表現したもので比較した結果、10 進数と指数表記のものが最も高精度であることが示された。しかし、どの表記も 15 桁や 20 桁など桁数が増えるにつれて精度は低下することが示され、限界があることが明らかになった。

3 提案手法

本研究では、順序数に関する問題をテキスト形式でモデルに入力し、その問題の正答をテキスト形式で出力させるようなタスクについて考える。

順序数に関する問題は、1 つ以上の数え上げスキルを用いて解くことができると考える。ここで、1 つの数え上げスキルは、数え上げ対象、数え上げ粒度、起点、向きの 4 つを把握する能力であると定義する。例えば、「studio の 2 文字目は何ですか」という問題は、1 つの数え上げスキルを必要とする問題であり、その数え上げ対象は、与えられた studio という文字列であり、数え上げ粒度は文字単位に揃っており、起点は先頭、向きは正順である。これら 4 つの項目を適切に把握することができれば、当該数え上げを正しく遂行できる可能性が高いと考えられる。

以下、数え上げスキルの各項目についてより詳細に説明する。

数え上げ対象は、その対象そのものが問題内で与えられているか (所与) か、数え上げ対象を推論する必要があるか (要推論) の 2 つに分けて考える。

「studio の 2 文字目は何か」というような問題では、数え上げを行う対象は studio という単語であり既に与えられている。「A さんは 2 番目,B さんはその 1 つ後ろに並んでいます。B さんは何番目に並んでいますか」という問題では、A さんと B さ

んが並んでいる列そのものの様子は与えられておらず、具体的な列の状況や位置関係を推論する必要がある。

起点と向きについては、問題内で具体的な指示が与えられているか否かによってそれぞれ 2 種類に分類した。

”前から 3 番目”や”後ろから 3 番目”,”A さんから数えて 3 番目”などのように、数え上げの起点が指定されている場合は、起点の指示ありに相当する。また、”studio の 2 文字目”などの場合は、起点の指示なしに相当し、常識的に左端が数え上げ起点であることを把握する必要がある。また、”右から 3 番目と 2 番目”という表記では、前の”3 番目”は”右から”という指定があるため起点・向きともに指示ありに相当するが、後の 2 番目は起点や向きの指定はなしに相当し、右から数え上げを行うことを把握する必要がある。

粒度については数え上げ対象が同じ桁数や種類で揃っているか揃っていないかで分類する。

作成した各問題をこれらの視点で難易度別に分類した。その結果、合計で 188 種類の問題となった。数え上げ回数が 1 回の問題については 20 種類、そのうち対象が所与の場合、要推論の場合でそれぞれ 10 種類、そのうち 5 種類は起点, 向きの指定あり、残りの 5 種類はなしとして作成した。数え上げ回数が 2 回の問題は 32 種類で、そのうち対象が所与の場合、要推論の場合でそれぞれ 16 種類、起点, 向きの指定の有無でそれぞれ分けて作成した。数え上げ回数が 3 回の問題は 64 種類で、そのうち対象が所与の場合、要推論の場合でそれぞれ 32 種類、起点, 向きの指定の有無でそれぞれ分けて作成した。また、順序数を用いて基数を算出させるような問題も同様の視点で作成した。基数を算出させる問題については数え上げ回数が 1 回の問題は 8 種類、2 回の問題は 16 種類、3 回の問題は 48 種類作成した。

問題は単語に関する問題、数の並びの問題、列の並びの問題、上下の並びの問題、基数数え上げに関連するものとした。

単語に関する問題では、n 番目を問う問題や間の 2 つの文字の間の文字を問う問題などが含まれている。英単語の平均文字数は 4.7 文字であることから、それに近い 4 6 文字の isogram である単語を使用した。問題の例として「break の 2 文字目は何か」「break の後ろから 2 番目は何か」「アルファベットの 4 番目は何か」というものが挙げられる。1 つ目の問題は単純に順序を問う問題であり、数え上げ回数は 1 回、対象は break という文字列のため所与、数え上げ起点と向きの指定はなく、常識的把握となる。2 つ目の問題は数え上げ回数が 1 回、対象は同様に所与、数え上げ起点と向きは後ろからという指定がある。3 つ目は数え上げ回数が 1 回、対象はアルファベットとなっており、具体的な文字は指定されていない。そのため、アルファベットの情報を推論する必要がある。数え上げ起点と向きは指定はなく、常識的な把握となっている。

数の並びに関する問題では、[2,3,4,5] のような数集合が与えられ、その中の n 番目を問うような問題などが含まれている。例として「数の並び 2,3,4,5 の 2 番目は何か」「数の並び 2,3,4,5 の後ろから 2 番目は何か」「1 桁の自然数の 3 番目は何か」という問題が挙げられる。これらは単語の問題と同様で 1 つ目の問題は単純に順序を問う問題であり、数え上げ回数が 1 回、対象は 2,3,4,5 という数集合のため所与、数え上げ起点と向きの指定はなく、常識的把握となる。2 つ目の問題は数え上げ回数が 1 回、対象は同様に所与、数え上げ起点と向きは後ろからという指定がある。3 つ目の問題は数え上げ回数が 1 回、対象は 1 桁の自然数となっており、具体的な数値が指定されていないため推論が必要となる。数え上げ起点と向きの指定はなく、常識的把握となる。なお、数の種類によっての変化を確認するために数集合の数値が 1 桁,2 桁,3 桁の 3 パターン作成した。

列の並び、上下の並びの問題は数え上げ対象は明記されておらず、A さんと B さんの位置関係を考慮しながら最終結果を推論する必要がある問題となっている。例として「A さんは 3 番目,B さんはその 1 つ後ろにいます。B さんは何番目ですか」という問題が挙げられる。これは数え上げ回数が 3 回となっており、対象は順序の情報から列の情報を推論する必要があり、1 つ目の数え上げ及び最終的な結果は起点, 向きの指定はなく、2 つ目の数え上げは A さんを起点として数え上げる必要があるような問題となっている。

基数数え上げの問題では、順序情報を用いて基数を算出させるような問題となっている。例として「A さんは前から 3 番目, 後ろから 4 番目にいます。この列には何人並んでいますか」というような問題が挙げられる。これは数え上げ回数が 3 回、最終的な結果は基数である列の人数となり、1 つ目と 2 つ目の数え上げはどちらも起点, 向きの指定がある。

このような問題の答えを正しく出力するには、単純に足し算引き算の計算ではなく数え上げの概念の把握、つまり順序数の概念をを把握している必要がある。

4 実　　験

4.1 実験設定

データセット　saxton らのデータセットである mathematics dataset[9] を使用する。このデータセットは言語モデルの数値推論能力を調査するために構築されたものであり、四則演算などの単純な算数の問題や数列、連立方程式、微分積分など小学校から大学レベルまでの合計 56 分野の数値推論問題が含まれている。データはそれぞれ問題文とその答えから構成されている。これらのデータを構築した後、Transformer モデルに学習させ正答率を確認している。その結果、四則演算などの単純な問題に関しては 95%以上、数列の第 n 項を求めさせる問題は約 80%、進数変換や素因数分解などの複雑な問題では 10%にも満たない精度であることが示された。

このように、幅広い分野について触れられていることから、本研究ではこのデータセットを事前学習用データとして使用する。

モデル　モデルは上記のデータセットを用いて実験が行われている TP-Transformer[10] を使用する。これは QKV 注意機構に役割ベクトルを追加し、注意機構の出力と役割ベクトルの要素積をとる点が通常の Transformer とは異なっている。この拡張により、記号処理における各記号の役割をより際立たせることを意図したものである。

このモデルに mathematics dataset を 70 万ステップ学習させた結果、従来の Transformer モデルよりも優れた結果を発揮した。四則演算のような単純な問題に関してはほぼ 100%、数列の第 n 項を求めさせるような問題は約 85%の正答率を示している。しかし、進数変換や素因数分解のような複雑なタスクに関しては約 10%ほどという低い正答率となっている。

実験内容　本研究においても同様に mathematics dataset で 70 万ステップ学習した TP-Transformer を使用する。このモデルに対して提案タスクを与え、正解率を調査した。

4.2 実験結果

4.2.1 正解率の結果

最も正解率が高くなったのは「単語の n 文字目を問う問題」であった。これは純粋な順序を問う問題であり、数値が出現するのは 1 度のみである。一方で他の問題は A さんの順序を表す数値と B さんの順序を表す数値や数の集合と順序を表す数値から構成されているため、注視すべき数値を誤っている可能性があると考えられる。

数え上げ回数での比較を行った結果、1 回の場合は、平均正解率が 20.91%、2 回の場合は平均正解率が 33.44%となった。

数え上げ対象での比較を行った結果、対象が所与の場合は 34.37%の正解率となり、要推論の場合は 22.58%となった。

起点, 向きでの比較を行った結果、指定がない場合は 32.53%、指定がある場合は 22.74%となり、起点, 向きの指定がない場合の方が高い正解率となった。単語の n 文字目を問う最も単純な問題においては 98.83%の高い正解率となったが、起点, 向きの指定が加わった問題になると 60.07%まで正解率が低下した。これは逆順からの数え上げは、まず文字列全体を把握した後に数え上げを行う必要があり、複雑な問題となっているためであると考えられる。

数の並びの問題は単語の問題と同程度の難易度であると想定して作成したが、単語の n 文字目を問う問題は 98.83%であるのに対し、数の並びの n 番目の数値を問う問題は 22.27%と低い正解率となった。このような結果となった要因として、問題文中に基数と順序数が混在しており、順序数と基数の判別がついていないためであると考えられる。また、数の並びの問題についても同様で起点, 向きの指定が加わると精度が半減した。

数の並びの問題において、出現する数値の桁数による正解率の変化を確認した結果、桁数が増えるにつれ正解率が低下する場合が多く見られた。

今回は正解率のみを確認したが、内部状態がどのようになっているかは不明である。そのため今後は、Attention の状態を可視化し内部状態を確認する必要があると考えられる。

4.2.2 単語の長さや出現頻度による正解率の変化

単語の n 文字目を問う問題において、単語の出現頻度や長さによって変化が生じるのかを確認した。単語は単語頻度のデータ [11] から抽出した。これはブログ,WEB, テレビ (映画), 話し言葉, フィクション, 雑誌, 新聞, 学術の 8 つのジャンルで出現する 6 万の単語とその頻度がまとめられている。これを用いて高頻度語、中頻度語、低頻度語を抽出した。頻度が 100000 以上のものを高頻度とした。中頻度は 10000 以上の単語と 1000 以上の単語の 2 種類、低頻度は 100 以上のものとした。単語の長さはそれぞれ 6 文字とし、表 4.1 中の単語の n 文字目を問う問題と同様の形式の問題を作成した。また、この時 apple のように同じ文字が複数回出現している単語において、順番を誤っていても出力が正しければ正解とならないように同じ文字が出現しない単語に限定して抽出した。その結果、7 単語がヒットした。

これらを用いて問題を作成し、モデルに与えた際の正解率を以下の表にまとめた (表 1,2,3,4)。

この結果より、高頻度語の場合がやや高い正解率となっていることがわかる。全体的に低い正解率となっているが、これはデータ数が少ないことが関係していると考えられる。

表 1　高頻度語の結果

問題文	正解率 (%)
「method」の 2 文字目は何か?	33.33
「method」の 2 文字目の後ろの文字は何か?	44.44
「method」の 2 文字目の前の文字は何か?	44.44
「method」の右から 2 文字目は何か?	33.33
「method」の右から 2 文字目の後ろの文字は何か?	22.22
「method」の右から 2 文字目の前の文字は何か?	55.56

表 2　中頻度語 (頻度 10000 以上) の結果

問題文	正解率 (%)
「studio」の 2 文字目は何か?	11.11
「studio」の 2 文字目の後ろの文字は何か?	22.22
「studio」の 2 文字目の前の文字は何か?	11.11
「studio」の右から 2 文字目は何か?	0
「studio」の右から 2 文字目の後ろの文字は何か?	0
「studio」の右から 2 文字目の前の文字は何か?	44.44

表 3　中頻度語 (頻度 1000 以上) の結果

問題文	正解率 (%)
「cinema」の 2 文字目は何か?	11.11
「cinema」の 2 文字目の後ろの文字は何か?	22.22
「cinema」の 2 文字目の前の文字は何か?	22.22
「cinema」の右から 2 文字目は何か?	22.22
「cinema」の右から 2 文字目の後ろの文字は何か?	0
「cinema」の右から 2 文字目の前の文字は何か?	55.56

表 4　低頻度語の結果

問題文	正解率 (%)
「curate」の 2 文字目は何か?	22.22
「curate」の 2 文字目の後ろの文字は何か?	44.44
「curate」の 2 文字目の前の文字は何か?	22.22
「curate」の右から 2 文字目は何か?	22.22
「curate」の右から 2 文字目の後ろの文字は何か?	0
「curate」の右から 2 文字目の前の文字は何か?	55.56

続いて単語の長さによってどのような変化が生じるのかを確

認した。英単語の平均文字数は 4.7 文字であるため 3 文字で構成される短い単語の場合と平均値に近い 5 文字の場合、9 文字以上の長い単語の場合の問題をそれぞれ作成し、モデルに与え正解率を確認した。その結果が以下の表である (表 5,6,7)。

この結果を比較すると、5 文字の単語の正解率が最も高く、9 文字以上の長い単語では正解率が低下していることがわかる。これは文字数が増えるにつれて数え上げの範囲が増え、正確な順序を把握することが困難になっていることが原因であると考えられる。

表 5　3 文字の単語の結果

問題文	正解率 (%)
「cut」の 2 文字目は何か?	96.09
「cut」の 2 文字目の後ろの文字は何か?	97.27
「cut」の 2 文字目の前の文字は何か?	99.22
「cut」の右から 2 文字目は何か?	96.88
「cut」の右から 2 文字目の後ろの文字は何か?	99.61
「cut」の右から 2 文字目の前の文字は何か?	97.27

表 6　5 文字の単語の結果

問題文	正解率 (%)
「about」の 2 文字目は何でか?	99.22
「about」の 2 文字目の後ろの文字は何か?	100
「about」の 2 文字目の前の文字は何か?	100
「about」の右から 2 文字目は何か?	95.31
「about」の右から 2 文字目の後ろの文字は何か?	98.83
「about」の右から 2 文字目の前の文字は何か?	100

表 7　9 文字以上の単語の結果

問題文	正解率 (%)
「bricklehampton」の 2 文字目は何か?	97.27
「bricklehampton」の 2 文字目の後ろの文字は何か?	94.14
「bricklehampton」の 2 文字目の前の文字は何か?	96.48
「bricklehampton」の右から 2 文字目は何か?	52.52
「bricklehampton」の右から 2 文字目の後ろの文字は何か?	53.52
「bricklehampton」の右から 2 文字目の前の文字は何か?	61.72

4. 2. 3　モデルの推論結果

最後に、モデルの出力を一部抽出して確認した。「ビルの 2 階と 37 階の間には何階あるか」のような 2 つの順序の間にある数を求めさせる問題において正解は 34 であるのに対し、モデルの回答は 35 という誤った回答を出力していた (図 1)。これは問題文中に出現している 2 と 37 を基数としてそのまま用いて計算した結果であり、順序の状態を考慮して推論を行えていないと考えられる。以上のような結果から、Transformer に基づくモデルは基数との判別が必要な問題では正しい推論を行うことが困難であるということが明らかになった。

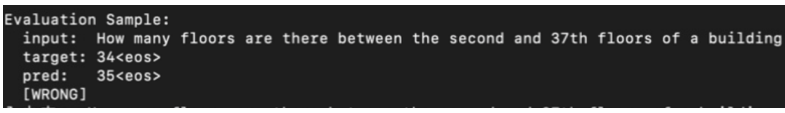

図 1　間の階数を問う問題の誤答例

(参照 20220126)

5 ま と め

本研究では順序数を用いた的確な数え上げが求められるようなタスクを課すことにより Transformer モデルの順序数の概念把握能力を有するかどうかについて調査した。その結果、全体的に低い精度となり、数え上げが必須となる問題を解くことは困難であることが明らかになった。特に順序数と基数の判別が必要な問題では正しい推論を行うことが困難であるということが明らかになった。このことから、Transformer モデルが順序数の概念を理解することは困難であると言える。

今回は 3 つの数概念のうちの順序数のみに着目して実験を行ったが、基数と数量に関しても同様の調査を行うことで数概念全体の理解能力を明らかにできると考えられる。また、Transformer 内部の挙動分析を進めていくことが今後の課題である。

文　　献

[1] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, Jacob Steinhardt, “Measuring Mathematical Problem Solving With the MATH Dataset,”” 35th Conference on Neural Information Processing Systems (NeurIPS 2021)

[2] Arkil Patel, Satwik Bhattamishra, Navin Goyal, “Are NLP Models really able to Solve Simple Math Word Problems?,”” 2021

[3] Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, Eduard Hovy, “Exploring Numeracy in Word Embeddings,”” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3374–3380 2019.

[4] 青木洋一, 工藤慧音, Ana Brassard, 栗林樹生, 吉川将司, 乾健太郎, “多段の数量推論タスクに対する適応的なモデルの振る舞いの検証,””, 言語処理学会 第 28 回年次大会 発表論文 (2022 年 3 月)

[5] 平井安久, 青山陽一, 曽布川拓也, “数の概念の捉え方について,” 数理解析研究所講究録, 第 1828 巻 2013 年 pp.86–100.

[6] 松本 悠太, 吉川 将司, Benjamin Heinzerling, 乾 健太郎, “四則演算を用いた Transformer の再帰的構造把握能力の調査,”言語処理学会 第 28 回年次大会 発表論文 (2022 年 3 月)

[7] Devin Johnson, Denise Mak, Drew Barker, Lexi Loessberg-Zahl, “Probing for Multilingual Numerical Understanding in Transformer-Based Language Models,”” Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 184–192 Online, November 20, 2020.

[8] Rodrigo Nogueira, Zhiying Jiang, Jimmy Lin, “INVESTIGATING THE LIMITATIONS OF TRANSFORM- ERS WITH SIMPLE ARITHMETIC TASKS,”” 1st Mathematical Reasoning in General Artificial Intelligence Workshop, ICLR 2021.

[9] David Saxton, Edward Grefenstette, Felix Hill, Pushmeet Kohli, “ANALYSING MATHEMATICAL REASONING ABILITIES OF NEURAL MODELS,”” Published as a conference paper at ICLR 2019

[10] Imanol Schlag, Paul Smolensky, Roland Fernandez, Nebojsa Jojic, Jurgen Schmidhuber, Jianfeng Gao, “Enhancing the Transformer With Explicit Relational Encoding for Math Problem Solving,”” 2020

[11] www.wordfrequency.info, “Word frequency_based on one billion word COCA corpus””, https://www.wordfrequency.info/samples.asp,