日本語の Math Word Problems に対する深層学習モデルの適用とデータ拡張の検証

村田夏樹　　柴田千尋
法政大学　理工学部　創生科学科
natsuki.murata.4d@stu.hosei.ac.jp，chihiro@hosei.ac.jp

概要

Transformer などの近年の深層学習モデルは，多くの自然言語理解を必要とするタスクで高い性能を記録している．しかし，数学や物理の文章題などの定量的な推論が必要なタスクでは，近年の言語モデルでも十分な結果を残せていない [1]．本稿では，Math23K というデータセットを利用して，日本語 MWP タスクに対して，既存の深層ニューラルネットモデルを組み合わせたモデルを適用し，検証を行う．なお，Math23K データセットの問題文は，中国語で書かれているため，機械翻訳を用いて問題文を日本語に翻訳しデータとして用いる．モデルの組み合わせとしては，Goal-driven Tree Structured (GTS) ネットワーク，BERT, および Graph2Tree の 3 つを組み合わせたものを用いる．特に，BERT - GTS の組み合わせと，BERT - Graph2Tree - GTS の組み合わせの比較を行う．後者は，問題文中のトークンの BERT の埋め込み表現をノードとして持つ特定のグラフを，中間表現として用いて，その後グラフ畳み込みを行う手法であり，実際に後者のほうが精度が高くなることを示す．さらに，解答となる数式を表現する木構造の可換ノードの置換や，問題文中の単語の置き換えによるデータ拡張により，精度の向上が見込めるかについて，検証を行う．

1　はじめに

Transformer に代表される最近のニューラル言語モデルは，一般的に言って，自然言語理解を必要とするような多くのタスクにおいて，高い精度を達成しているものの，例えば，数学や物理の文章問題などの定量的な推論が必要なタスクでは，近年の言語モデルを用いても解答が難しいケースがおおく，未だ改善の余地が多く残されていることが知られている [1]．その中でも，Math Word Problems (MWP) は，文章問題から適切な数式を生成して，数式の演算を通して解答を求めるタスクである．このタスクでは，文章中にはないが，数式の木構造の中に必要となる 1 や π，e といった推論が必要な問題などがあるため，現在も発展中のタスクである．既存の手法の例として，事前学習済みモデルを使用した手法や，木構造を取り入れた手法などがある．このタスクが抱える問題の一つに，データ数が十分でなく，推論が必要な問題について，十分に学習できない．Zhang ら [2] は，Math23K という数学に関する文章題のデータセットで，77.4% の正解率 (出力された計算式の計算結果の一致率) を得ている．しかし，Math23K は，中国語の問題文と解答となる式からなるデータセットであり，問題文が日本語の場合において，同様の精度が得られるかについては，検証が求められる．

そこで，本研究では，Math23K を機械翻訳サービスを用いて，中国語を日本語に翻訳してデータとして用い，BERT 及び Graph-to-Tree (Graph2Tree) を用いて，日本語の数学の問題を解答するタスクを学習させる．さらに，Graph2Tree に含まれる Graph Convolution Network (GCN) [3] がどのようなグラフを生成するのかを検証するため，中間表現として構築されるグラフを可視化して検討を行う．最後に，データ拡張を行うことで，本タスクにおける式の精度と答えの精度に対し，効果があるのかを検証する．

2　関連研究

2.1　BERT

BERT [4] は，Transformer モデルであり，事前学習済みの BERT モデルを活用することにより，多くの自然言語処理のタスクで高い精度を期待することができる．BERT の事前学習では，Masked Language

Model と Next Sentence Prediction の 2 つの事前学習を行う．それぞれ行うタスクは，入力したトークンの一部にマスクをし，前後の情報から，マスクした箇所の内容を予想するタスクと，2 つの文章を入力して，連続した文章かどうかを判定させるタスクである．事前学習を行なった BERT は，入力の文章やトークンに対する，文脈を考慮した一般的な知識が何らかの形で内部に獲得されていると考えられるため， fine-tune を行うことで，特定のタスクに対して，高い精度を出すことが可能である．

2.2 Goal-driven Tree Structured Network

Goal-driven Tree Structured Netowrk (GTS) [5] は，表現木を生成するためのエンド to エンドの手法で，数式を表現する木構造の生成するニューラルネットワークである．生成する数式の木は，2 分木であり，ルートノードから順に，演算記号が入り，左右の子ノードには，数字あるいは記号が入る構造が再帰的に取られている．即ち，数字または変数が入るノードは，葉ノードに対応し，演算記号が入るノードは，節ノードに対応する．

GTS のネットワーク構造について概要を説明する．まず，入力問題文を単語トークンに変換し，各トークンごとの埋め込み表現にしたのち，双方向 Gated Recurrent Unit (GRU) [6] に入力する．その後，GRU の出力結果から，数式の木を再帰型ニューラルネットワークを用いて生成する．あるノードに対して，アテンションを用いて文脈ベクトルを得たのち，そのノードのラベルの確率，つまり，変数または演算子（+ や × 等）の確率を，文脈ベクトルを用いて計算する．その後，子ノードの文脈ベクトルを，順伝播型ネットワーク (FCN) に通した結果および GRU の出力結果から再帰的に計算することを繰り返すことで，最終的に，正解となる数式全体を生成するように学習する．

2.3 Graph-to-Tree

Graph2Tree [2] は， GCN [3] を含むモデルであり，2 種類のグラフから構成され，数量間の関係や，文章の情報をグラフ構造で捉えることができる．

グラフ内のノードの内，ある量に関連する部分集合を数量セルと定義する．数量セルが作るグラフは，数量セルグラフと数量比較グラフを作る．数量セルグラフは，情報量の多い単語を数と関連付け，表現を豊かにすることができる．また，数量比較グラフは，数の数値的性質を保持し，数量間の関係を保持し，数量間の関係の表現を改善することができる．2 つのグラフを隣接行列で表し，この隣接行列と特徴を表す行列（事前に LSTM などで得られた表現）を GCN に入力し，その後，得られた表現を連結した結果に対して，正規化と残差接続，また FCN 等を行うことで，問題文が最終的にエンコードされた表現を得る．

3 提案手法

3.1 BERT と Graph2Tree を用いたモデル

本研究で用いるモデルの概要を図 1 に示す．まず，事前学習済みの BERT を使用して，文章題を入力し，得られた埋め込み表現を得る．次に，その埋め込み表現から，前章で述べた Graph2Tree を使って，中間の埋め込み表現を得る．最後に，得られた中間表現から数式の木を生成する．

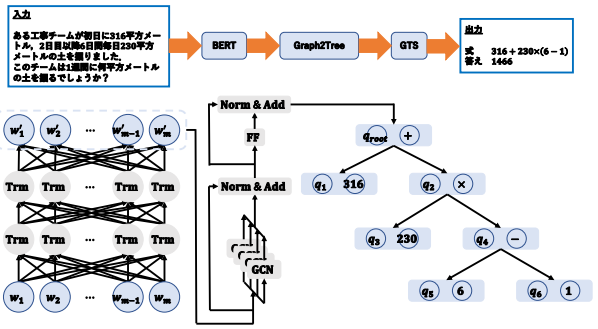


図 1: 使用するモデルの概要図

3.2 データセットの日本語化

3.2.1 Math23K

本研究では，Wang らの研究 [7] で提案されたデータセット， Math23K[1] を日本語に翻訳したものを使用する．この Math23K は，中国語で書かれた問題文，式及び答えから構成され，データ数は，23,162 件ある．

3.2.2 前処理

まず，日本語のデータセットを作成するため，DeepL [2] という機械翻訳サービスを使用して，日本語に翻訳した．翻訳前（中国語）の文章と翻訳後（日本語）の文章に含まれる数の集合（以後、「数集

1） https://ai.tencent.com/ailab/nlp/dialogue/
2） https://www.deepl.com/translator.html

合」と呼ぶ）を比較して，翻訳後の数集合が不足しているデータは，削除した．使用したデータと削除したデータの例をそれぞれ図 2 と図 3 に示す．

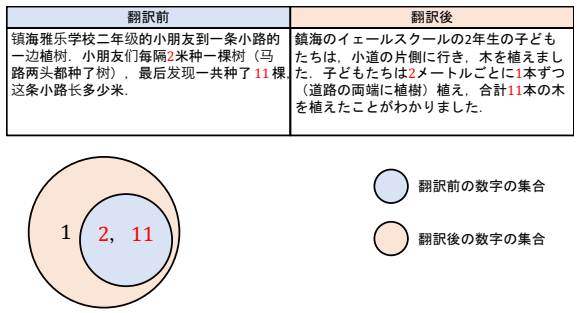


図 2: 使用したデータの例

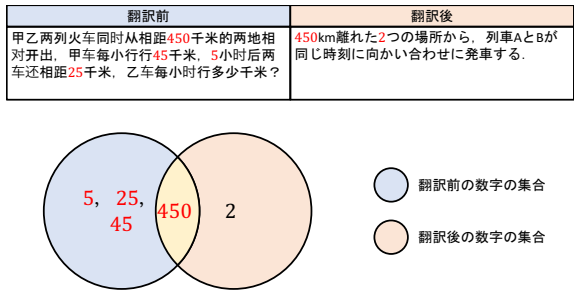


図 3: 削除したデータの例

図 2 では，翻訳前では，2 及び 11 があり，翻訳後でもこれらの数字があるため，このデータセットは使用する．一方，図 3 では，翻訳前では，5，25，45 及び 450 があるが，翻訳後では，5，25 及び 45 がないため，削除する．削除したことにより，データセットは，20,260 件になった．

3.3 データ拡張

本研究では，以下の二つのデータ拡張を行い，結果にどのような影響を与えるかを検討する．

一つ目は，正解の数式木に対するデータ拡張である．Math23K で与えられる式は，各問題分に対し，1 つの式しか用意されていない．そこで，加法交換法則と乗法交換法則に着目し，交換可能な部分木をランダムに交換することでデータ拡張を行う．拡張前後で，別の木構造について学習させる．ただし，− や ÷ だけで構成された式は，データ拡張できない．

表 1: 数式のデータ拡張の例

データ拡張前	データ拡張後
$(11-2)\times 2$	$2\times(11-2)$
$316+230\times(6-1)$	$(6-1)\times 230+316$

二つ目は，文章題の単語に一定の割合でマスクをかけるデータ拡張を用いて実験を行う．ただし，マスクをかける際，数字にはマスクがかからないようにする．

表 2: マスクの例

マスク前	マスク後
小明はある絵本を	小明は <mask> を
15 人の子どもが	15 人の <mask> が

4 実験

4.1 実装

本実験では，BERT の事前学習済みモデルとして，東北大学のもの[3] を用いた．また，節 3.1 で述べたモデルを学習する際に用いたハイパーパラメータを，表 3 に示す．

表 3: 使用した関数及びハイパーパラメータ

種類	名称
バッチサイズ	32
エポック数	85
ビームサーチ	5
隠れ層の次元	1024
学習率	3×10^{-5}（30 エポック毎に半減）
最適化関数	RAdam [8]

4.2 Graph2Tree の有無による精度の比較

まず，Graph2Tree が与える影響を検証するために，Graph2Tree を用いた場合と用いていない場合の精度を比較を行う．式の精度については，生成した式 (木構造) と模範解答が一致していない限り不正解として取り扱う．

表 4: Graph2Tree 有無による精度の比較

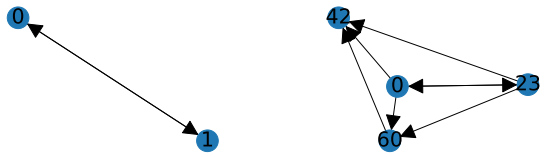
モデル	式の精度	答えの精度
BERT - GTS	44.3%	44.3%
BERT - Graph2Tree - GTS	45.7%	52.9%

表 4 より，BERT - Graph2Tree - GTS では，式の精度は向上しないが，答えの精度については，8% ほど上昇することがわかる．Graph2Tree を用いることで，式の形が違っていても答えに辿り着く可能性が上がることが分かる．

3） https://huggingface.co/cl-tohoku/bert-large-japanese

4.3 GCN のグラフの検討

前節の実験結果から，Graph2Tree を入れることで，精度が上がることが確かめられた．一方で，Math23K 中国語のデータセットに対して報告されている精度 (77.4%) [2] に比較すると精度が低いことがわかる．そこで，この節では，その原因を検証するために，Graph2Tree に含まれる GCN が，どの程度，意図したグラフを正しく生成できているのかを確認するために，数量セルグラフと数量比較グラフを可視化を行う．入力した文章は，「2 台の車 A，B が同時に A，B から反対方向に走り，2 時間半後に途中で出会う．車 A の速度は 90km/h，車 B の速度は車 A の速度の (4/5) であることが分かっている．A と B の間の距離は何キロメートルくらいか聞いてみてください.」を入力した．ただし，どちらの図についても，エッジがないノードや自己ループは可視化していない．



(a) 数量セルグラフ　　(b) 数量比較グラフ

図 4: 例文より得られる Graph2Tree によるグラフ表現

図 4a と図 4b は同一の文章から生成している．図 4a は，単語毎のノードがあり，それを結び，ノードの 0，1 は，問題文の「2」と「台」に相当する．また，表示はしていないが，図 4a では，エッジが結べていないノードが 92 個ある．つまり，数量セルグラフが，日本語における単語間の情報を結ぶエッジを十分に結べていないことが分かる．

図 4b は，文章中にある数字を使って，数の関係性について学習をし，ノードの 0，23，42 と 60 は，問題文の 2，2，90，(4/5) に相当する．図 4b では，数が小さい方から大きい方へ矢印を引いて，数量の関係性について表現している．

4.4 データ拡張

式の精度は，前述の通り， Math23K に用意されている正答の木と完全一致を見るため，データ拡張を行う際，使用するモデルは，BERT - Graph2Tree - GTS を使用して実験を行う．まず，数式をデータ拡張した実験結果を表 5 に示す．データ拡張後では，式の精度は下がっているものの，答えの精度としては同程度となっている事がわかる．

表 5: 数式のデータ拡張

データ拡張の有無	式の精度	答えの精度
無し	45.7%	52.9%
有り	40.1%	53.7%

表 6: 単語のマスク

マスクの有無	式の精度	答えの精度
無し	45.7%	52.9%
有り	46.7%	53.7%

次に，問題文にマスクをかけた実験では，50% の確率で mask をかけた．mask をかける単語については，エポック毎にランダムで選択させた．単語にマスクをかけた実験結果を表 6 に示す．単語にマスクをかけた場合も，同様に，式及び答えの精度に大きな差は見られていない．

どちらのデータ拡張の方法についても，本稿で用いたモデルに対して有効かどうかは，差が僅かであるため，さらなる検証が必要である．

5 おわりに

本研究では，他の言語のデータセットを日本語化し，BERT や Graph2Tree を用いて，実験を行った．また，日本語データ入力時における Graph2Tree で得られるグラフ表現について，検討を行った．二種類のデータ拡張を用いた実験では，少なくとも現在のネットワーク構成においては，式と答えの精度の観点から見て，効果は限定的であるという結果になった．今後の展望として，より数式生成にふさわしいと考えられるグラフ表現の抽出方法，および，グラフ構造を大きく変えるようなデータ拡張方法を探求してゆく予定である．

謝辞

法政大学大学院理工学研究科システム理工学専攻の辺見一成氏には，実験を行う際のプログラムの修正などご助言を頂きましたことを深く感謝申し上げます．本研究の一部は JSPS 科研費 JP18K11449 の助成を受けたものです．

参考文献

[1] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022.

[2] Jipeng Zhang, Lei Wang, Roy Ka-Wei Lee, Yi Bin, Yan Wang, Jie Shao, and Ee-Peng Lim. Graph-to-tree learning for solving math word problems. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3928–3937, Online, July 2020. Association for Computational Linguistics.

[3] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In **International Conference on Learning Representations**, 2017.

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

[5] Zhipeng Xie and Shichao Sun. A goal-driven tree-structured neural model for math word problems. In **Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19**, pp. 5299–5305. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[6] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[7] Yan Wang, Xiaojiang Liu, and Shuming Shi. Deep neural solver for math word problems. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 845–854, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[8] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In **International Conference on Learning Representations**, 2020.