B8TB2091

卒業論文

深層学習モデルの数量推論能力の評価とメタ学習活用の試み

工藤慧音

2022年3月31日

東北大学
工学部 電気情報物理工学科

本論文は東北大学 工学部 電気情報物理工学科に
学士 (工学) 授与の要件として提出した卒業論文である。

工藤慧音

審査委員：
乾 健太郎 教授 （主指導教員）

深層学習モデルの数量推論能力の評価とメタ学習活用の試み*

工藤慧音

内容梗概

今日, 自然言語処理の様々なタスクにおいて Transformer を初めとする深層学習モデルが大きな成果を上げており, 数量推論等の推論タスクにおいても高い性能に到達できることが示されている. しかしながら現状の深層学習モデルが問題の構成性を捉えて, その構造に応じた推論を行ってタスクを解いているのかについては定かではない. 実際は表面的な手がかりを利用した何らかのショートカットラーニングを行なっている恐れがある. そこで本研究では, 数量推論タスクを抽象化した形式言語を用いて既存の深層学習モデルの推論能力の評価を行なう. また, 調査を通して明らかとなった通常の教師あり学習では獲得が困難な数量推論能力を, 近年モデルに構成性を理解させる上での有効性が示されているメタ学習によって獲得することができるのかについても調査を行なう. 実験の結果, (1) 自然言語テキスト上での事前学習が形式言語上のタスクにおける性能向上につながること (2) 多段の推論を要する問題は既存の深層学習モデルで解くことが難しい問題であることがわかった.

キーワード

自然言語処理, 深層学習, 数量推論, 記号推論

*東北大学 工学部 電気情報物理工学科 卒業論文, B8TB2091, 2022 年 3 月 31 日.

Towards Improved Numerical Reasoning Ability With Meta-Learning*

Keito Kudo

Abstract

Deep learning models such as Transformers have been successful in various natural language processing tasks, including inference tasks such as numerical reasoning. However, it is not evident whether current deep learning models capture compositionality and perform inference according to the structure of the problem. They may be learning shortcuts from superficial cues. This study evaluates the performance of existing deep learning models using a formal language that abstracts numerical reasoning. We also investigate whether meta-learning, which has recently been shown to be effective in making models understand compositionality, can be used to acquire numerical reasoning skills that are difficult to obtain via conventional supervised learning. Our experimental results show that (1) pre-training on natural language texts leads to improved performance on tasks on formal languages, and (2) problems that require multi-hop inference are difficult to solve with existing deep learning models.

Keywords:

Natural Language Processing, Deep Learning, Numerical Reasoning, Symbolic Reasoning

*Graduation Thesis, Department of Erectrical, Information and Physics Engineering, Tohoku University, B8TB2091, March 31, 2022.

Contents

1　はじめに

文章読解のような言語理解では，しばしば論理推論や数量推論のような記号計算を行う能力が求められる．例えば,「太郎がクッキーを 4 枚持っていた．花子はクッキーを 2 枚持っており，さらに太郎からもクッキーを全てもらった．花子は何枚クッキーを持っているか？」という質問に答えるには,「`t = 4，h = 2 + t，h=?`」という計算をする必要がある．近年，深層学習モデルが自然言語処理分野で華々しい成功を収めているものの，単語や文をベクトルとして表現する深層学習モデルが，記号計算能力をどれほど有するのか定かではない．本研究では，数量推論問題を通して深層学習モデルの推論能力を分析し，深層学習という道具立ての性質の理解に迫る．深層学習モデルの数量推論能力については，例えば文章読解問題の DROP [1] を通して評価が行われてきた．しかしながら，語彙の偏りといった数量推論以外の手がかりを用いた解答の可能性も指摘されており [2]，モデルの数量推論能力を精緻に切り分けて評価できているとは言い難い．

本研究では，統制されたデータを用いて深層学習モデルの数量推論能力を精緻に評価・分析する．具体的には，形式言語 (数式) で記述された計算問題 (例えば，`A=1，B=A+4，C=1，B=?`) を設計し，問題の性質を制御しながら深層学習モデルの性能を観察することで，変数への代入・参照，関数適用 (四則演算) といった手続きのどこで計算が失敗しやすいのか，またどれほど複雑な問題まで計算が可能なのか理解を深める．さらに，自然言語処理分野における自然言語を用いた事前学習 (言語モデリング) の成功を踏まえ，事前学習とモデルの数量推論能力の関係についても分析を行う．

実験の結果，(i) 事前学習を施さないモデルでは，数量計算部分 (四則演算) よりも，数式の構文を理解し変数への代入・参照を行う部分でエラーが生じる傾向にあること，(ii) 自然言語テキストを用いた事前学習により，モデルが変数の代入や値参照を行う能力を獲得しやすくなること，(iii) 多段の推論を要する問題は既存の深層学習モデルにとって難しい問題であることが示唆された．事前学習の語彙と数量推論タスクの語彙の重複は小さく，事前学習時には言語の形式的な側面を明示的に教示していないにも関わらず，自然言語における事前学習が数量推論能力の獲得を助長するような帰納バイアスを導入している点は興味深い．

最後に，多段推論問題におけるモデルの性能改善を試みる．多段推論で計算が失敗する要因としては，モデルが問題の構成性を捉えきれていないことが考えられる．問題の構成性を学習させる手段としてメタ学習の有効性が示されていることを踏まえ [3]，メタ学習による多段数量推論能力の改善を試みる．本研究の実

験の範囲内では，メタ学習を用いても多段数量推論の達成が困難であることが示唆され，今後の改良方針を整理した．

2　関連研究

2.1　数量推論

近年の事前学習済み言語モデルに基づく文章読解モデル (GenBERT 等) が DROP 上で高い性能を示す他には，このようなモデルが，限られた数の範囲であれば，順序や足し算など基本的な数の構造を把握している [4] ことや，四則演算タスクにおいて有効な数のエンコーディング形式に関する知見 [5] が得られている (近年の数量推論研究のサーベイは [6] を参照)．しかし，推論タスク一般に見られる表層的な手がかりによる解法の問題 [2] のみならず，実応用的なタスク [7] では多段な数量推論を要する一方で，DROP の問題事例の大部分は二数の差に関するものであったり [8]，言語モデル一般に未知の大きさの数に対して頑健でない [4] (外挿問題) などの課題が残る。本研究では人工的なタスクから初めてこのような知見のギャップに言及することを目指すものである．

2.2　構成性を捉える能力を評価するためのデータセット

モデルの構成性を捉える能力を測るため SCAN [9] や COGS [10] などの意味解析データセットが提案されている．これらのデータセットを用いて，モデルの構成性を捉える能力の改善に向けた研究が進められている．モデルの構成性を捉える能力の改善に向けた試みの 1 つとして，メタ学習の一種である DG-MAML [11] を用いる方法が提案されている [3]．DG-MAML は標準的な Transformer などの系列変換モデルを構成性タスクに対して強力にする手法であり，本研究で用いる．(手法の詳細は 5 節詳述) ．

3　データセット・実験設定

数量推論問題の複雑さを変化させてモデルの性能を調査し，数量推論能力の限界を調査する．また，事前学習の影響も調査する．

3.1 数量推論能力評価のためのデータセット

問題設定のコントロールがしやすいよう，形式言語で記述されたデータセットを用いる．モデル評価に用いた形式言語の問題形式と，それらの形式により評価したい推論能力を表 1 に記す．学習データを暗記するだけでは解けない設定にするため，先行研究 [12] に従い，学習データ，検証データ，評価データにおいて数式を構成する数量が異なるようにする．具体的には，0 から 99 までの自然数のうち無作為に選ばれた 80 種類の自然数が学習データの数式に，残りの 20 種類のうち 10 種類ずつが検証データ，評価データに出現する．モデルの学習は，問題の形式と式の長さごとに個別に行う．各設定において，学習データは 10 万件，検証，評価データは 2,000 件とする．

Table 1: 追加学習時の学習データの形式

	長さ	入力	出力	詳細
形式 1	1	I=4, A=2, I=	4	・変数の値を参照する能力を検証する ・変数名，数字は無作為に選択される
	2	L=5, N=1, P=5, P=	5	
	3	A=7, T=1, B=3, R=9, T=	1	
	4	C=1, H=4, R=6, T=9, O=2, C=	1	
形式 2	1	L=7+2, M=2-7, L=	11	・計算結果を変数に保持し，その値を参照することができるのかを検証する ・変数名，数字，演算子は無作為に選択される
	2	G=6+2, S=2−3, D=3−5, S=	−1	
	3	E=2−9, T=3−1, A=3+8, M=4+5, M=	9	
	4	A=5+2, T=4−3, I=1+2, N=9+9, R=4+7, N=	18	
形式 3	1	D=3, V=D+1, V=	4	・計算結果を変数に保持し，次の計算でその値を参照することができるのかを検証する ・変数名，数字，演算子は無作為に選択される ・数式中の式の順番は無作為である (例：B=A+2, A=1, C=B+3)
	2	K=5, X=K−1, O=8+X, O=	12	
	3	S=9, Q=3−S, J=Q+3, W=5+J, W=	2	
	4	F=2, U=F−9, C=2−U, W=C+1, E=W−4, E=	6	
	5	Y=1, P=Y−6, N=6+P, O=N−7, H=4−O, T=5+H, T=	15	

3.2 事前学習設定

図 1 に示すように，自然言語を用いた事前学習と二項演算データを用いた事前学習を行なった後，前述の形式言語データの学習データで追加学習を行う．以降，この設定を LM+Arith と呼ぶ．事前学習の効果を分析するため，言語モデルの事前学習のみ (LM), 二項演算の事前学習のみ (Arith)，事前学習は行わない (Scratch) の 4 種類の設定での評価も行う．言語モデルの事前学習については，事前学習済みモデルのパラメータを利用することに対応する．二項演算データによる事前学習では，表 2 に示す二項の四則演算を暗記させる．本研究の主たる関心は，計算式の示す手続き (代入や参照など) を実行する能力であり，四則演算に関する知識は有している前提でモデルを評価したいため，追加学習時に出現しうる二項演

算のパターンをあらかじめ教えておく．追加学習時のデータセット内で出現しうる演算の組み合わせを全て含むようにするためにデータセット内に出現する数字は −1000 から 1000 までの範囲とし，事前学習を終えた段階での二項演算による計算については，正解率が 100%となることを確認した．

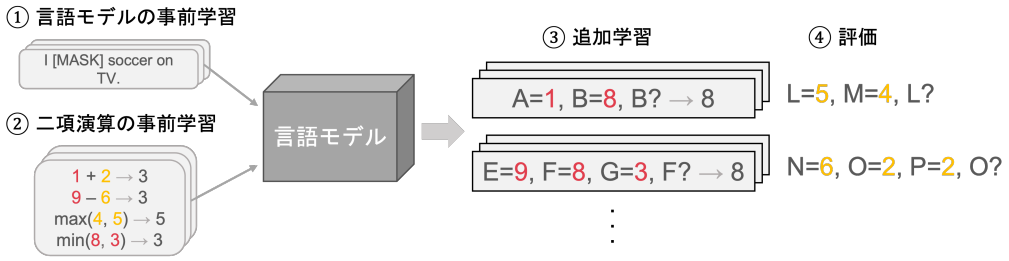

Figure 1: 深層学習モデルの推論能力の評価の流れを示した図．赤は追加学習のデータで用いる数字，黄色は評価データで用いる数字を表している．追加学習の学習データ，評価データで用いる数字の集合は互いに素である．

Table 2: 事前学習時の学習データの形式

演算子	入力	出力
加算	A=1+2, A=	3
減算	B=1−2, B=	−1
max	C=1^2, C=	2
min	D=1∗2, D=	1

3.3 モデル

BART [13] に一部変更を加えたモデルを利用する．[1] LM, LM+Arith の設定では，事前学習済みパラメータを用い，それ以外の設定ではパラメータを初期化した．

[1]BART とトークナイザの実装は huggingface の提供するライブラリである transformers (`https://huggingface.co/docs/transformers/model_doc/bart`) において公開されている’facebook/bart-base’ (`https://huggingface.co/facebook/bart-base`) を利用している．

位置埋め込みについては，エンコーダにおける入力系列の各トークンに対する位置埋め込みが絶対位置に基づいた表現ではなく，相対的な位置埋め込みとなるように変更を行なう．これは，学習時の入力系列長よりも長い系列で数量推論能力を評価する際に，未知の長さの系列を処理する能力の限界から，推論が失敗することを防ぐためである．具体的には，先行研究 [14, 12] に倣い，ある定数 K を最大幅として，絶対位置埋め込みを入力の度にシフトして学習を行う．最大シフト幅 K は十分大きな値として 100 とする．また単語分割については，GenBERT [12] に従い，数量表現について桁ごとに分割が行われるよう変更を加えたものを利用する．

4 実験結果

事前学習設定，データセット形式ごとの，モデルの性能を表 3 に示す．実験の結果，(i) 事前学習の設定に応じて性能が異なること， (ii) 多段の推論を要する問題 (形式 3) は深層学習モデルにとって比較的難しいタスクであると分かった．

Table 3: 各設定ごとの評価データに対する正解率 (%)．LM は言語モデル，Arith は二項演算の事前学習を行ったことを表す (Scratch は事前学習なしの場合).「長さ」については表 1 参照.

	形式 1: `A=1, B=2, C=3, C=?`				形式 2: `A=1+2, B=2+3, C=3+4, C=?`				形式 3: `A=1, B=A+2, C=3+B, C=?`			
長さ	1	2	3	4	1	2	3	4	1	2	3	4
LM+Arith	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.95	96.35	87.90
LM	100.00	100.00	100.00	100.00	100.00	99.80	100.00	100.00	100.00	99.70	88.35	73.05
Arith	100.00	100.00	100.00	100.00	48.95	34.85	27.05	21.45	99.85	70.95	41.45	22.60
Scratch	100.00	100.00	100.00	100.00	48.60	31.05	26.15	21.20	88.60	58.90	31.25	14.35

言語モデル事前学習の効果: 事前学習設定による性能の違いについて，形式 2，3 (表 3) では，自然言語を用いた事前学習によって性能が大きく向上している．特に形式 2 の問題に対する性能改善が顕著であり，自然言語を用いた事前学習のみで 100%に近い性能を達成しており，二項演算の結果を明示的に教えることなく四則演算の実行も行えていることが分かる．言語モデルの事前学習を行わない場合について，モデルの予測の間違いを分析すると，多くの事例で質問で問われている変数とは異なる変数の値を出力していることがわかった (表 4)．形式 2 において，二項演算の事前学習のみを用いる学習設定 (Arith) での評価データ

に対する正解率は，モデルが計算は正しく行えているが，どの変数の値を問われているかの判断はできないと仮定した場合のチャンスレートとほぼ一致している．例えば，形式 2，長さが 3 の問題形式の事例として A=1+2，B=2+3，C=3+4 を考えると, モデルが正しく計算を行えていた場合，解答となりえる可能性のある数字は 3，5，7 の 3 種類である．これらの数字の中から無作為に数字が選択される場合，正解率は約 33% となる．実際，二項演算の事前学習のみを用いる学習設定 (Arith) における形式 2，長さ 3 のときの正解率は表 3 に示した通り 34.85% とほぼ一致している．このことから，言語モデルの事前学習はテキスト中の参照関係や四則演算などの計算の学習を行うためにも，適した帰納バイアスを導入できていると考えられる．

Table 4: 形式 2，二項演算の事前学習のみを用いる学習設定 (Arith) において評価データで誤答した事例

評価データで誤答した事例	正解	モデルの予測
j = 3 - 28, m = 3 + 3 m =	6	-25 (= j)
z = 31 + 13, r = 35 + 86 z =	44	121 (= r)

多段推論の難しさ: 表 3 に示した結果から，3 つの問題形式のうち形式 3 の多段の推論を要する問題が，従来の教師あり学習の枠組みで深層学習モデルが学習するには比較的難しいタスクであるということがわかる．多段推論問題が深層学習モデルにとって困難な問題形式である要因の 1 つとして，問題の構成性を理解せず表面的な問題形式のパターンマッチのみで特定の問題を解いている可能性が考えられる． 例えば，モデルは与えられた問題文の構成性を捉えて構文解析を行い，それによって得られた手続きに基づいて計算を行うのではなく，データセット内に出現する特定の演算子の並びや変数名などのパターンをテンプレートとして覚え，推論時には与えられた数値に対して，似たテンプレートに対応する処理を行なっている可能性がある．この場合，問題の長さが増加するほど，パターンの数は指数関数的に増加することとなり，それが正解率の低下につながると考えられる． そこで，次章では構成性を捉えた学習が可能になると報告されているメタ学習の手法を用いて，多段推論問題におけるモデルの性能向上を試みる．

5　メタ学習による多段推論へのアプローチ

メタ学習の手法として，DG-MAML [11] を用いる．既存研究では，系列変換モデルを用いた意味解析タスクにおいて構成性を捉えた学習が可能になることが報告されている [3]．擬似コードを Algorithm 1 に示す.[2]この手法は，MAML [15] の亜種であり，ある事例セット D_{train} でモデルのパラメータを更新 ($\theta \rightarrow \theta'$) した直後に，その事例と入力の形式がわずかに異なる別の事例セット D_{dev} での汎化を促す (D_{train}，D_{dev} とも本来の学習セットの一部であることに注意)．これにより，モデルに対し，事例の構成要素の僅かな違いを認識しなければ損失 L を最小化できないというような制約をかけられる．

先行研究では，D_{dev} の事例は D_{train} に対して類似度をもとに学習セット中から選択されるが，本研究では，以下のルールを適用することによって D_{dev} の事例を直接生成する．具体的には，あらかじめ用意した D_{train} の事例に対して，長さ (推論のステップ数) を増やすことにより D_{dev} の事例を生成する．例えば，`A=1,B=A+1,B?`という長さ 1 の事例に対しては，`A=1,B=A+1,C=B+2,B?`のように 1 つ数式が追加された長さ 2 の問題を生成し D_{dev} の事例として用いる．このような設定でメタ学習を行うことで，モデルは訓練事例の構造 (式の連なりで構成されること) をより認識しやすくなり，複雑な問題にも対処できるようになると期待できる．

Algorithm 1 メタ学習のアルゴリズムの疑似コード

1: 初期値 θ をランダムに決める
2: **for** $i = 0 \rightarrow max\text{-}epoch$ **do**
3:　**for** $D_{\text{dev}}, D_{\text{train}}$ **in** Datasets **do**
4:　$\theta' \leftarrow \theta - \alpha \nabla_\theta L_{D_{\text{train}}}(\theta)$
5:　$L(\theta) = L_{D_{\text{train}}}(\theta) + L_{D_{\text{dev}}}(\theta')$
6:　$\theta \leftarrow \theta - \beta \nabla_\theta L(\theta)$
7:　**end for**
8: **end for**

[2]Algorithm 1 の 5 行目において $L_{D_{\text{train}}}(\theta)$ を足すのは誤りではなく，MAML に対する DG-MAML の特徴である．

5.1 実験設定

多段推論問題 (形式 3) に対するメタ学習の効果を検証するため，通常の教師あり学習，メタ学習のそれぞれでモデルの学習を行った際の評価データに対する正解率を比較する．学習時には，D_{train} として長さが n の問題，D_{dev} として長さが $n+1$ の問題を採用する．長さ $n+1$ の問題の作成方法は前述のとおりである．通常の教師あり学習の場合には，D_{train} と D_{dev} の和集合を学習データとして用いる．

5.2 実験結果

表 5 (上半分) に，評価データの長さを学習データの長さの最大値に揃えて評価した際の正解率を示す．メタ学習，通常の教師あり学習それぞれで学習を行った場合について報告している．実験の結果，メタ学習を活用しても評価データにおける正解率は向上しないことが観察された．

正解率が向上しなかった理由として，DG-MAML によって構成性を捉えるような学習を行った結果，モデルにとってより難しい条件で問題を解かせるように強制することとなってしまうため，学習が難しくなり評価データに対する正解率が低下してしまった可能性がある．一方で，仮に構成性を捉え，式の意味について理解した学習が行えていた場合，長さが変化しても対応できるようになっていることが期待される．そこで，学習データに含まれる問題よりも長さが 1 大きい問題におけるモデルの性能を評価した (表 5 下半分)．しかしながら，学習データよりも長さが 1 大きい問題に対する正解率についても，メタ学習を利用した場合と通常の教師あり学習を用いた場合では差が見られなかった．このため，現状のメタ学習の適用は有効とは言えず、さらなる調査は今後の課題とする．

Table 5: メタ学習の有無ごとの評価データにおける正解率 (%).

学習時の式の長さ	評価時の式の長さ	通常学習	メタ学習
1, 2	2	99.05	98.85
2, 3	3	94.25	93.45
3, 4	4	85.65	88.60
4, 5	5	79.80	79.00
1, 2	3	50.40	52.10
2, 3	4	72.15	73.35
3, 4	5	78.80	77.25
4, 5	6	72.40	72.50

6 おわりに

本研究では，既存の深層学習モデルの推論能力の評価を行い，その結果モデルでの学習が困難であると明らかになった問題に対してメタ学習を用いた性能の改善を試みた．実験の結果，多段推論問題が既存の深層学習モデルにとって困難な問題形式であるとこと，自然言語テキストを用いた事前学習により，モデルが変数の代入や値参照を行う能力を獲得しやすくなることが確認できた．またメタ学習については，既存の実験設定に倣い適用するだけでは多段推論問題の性能向上に有効でないことがわかり，深層学習モデルによる多段数量推論の実現が挑戦的なものであることが示唆された．

今後の課題として，深層学習モデルの推論能力評価については，異なるサイズのモデルにおける推論能力の違いの評価，今回用いた形式以外の問題を用いた推論能力の評価など，様々な軸での調査が考えられる．また，言語モデルの事前学習のどのような要素が，形式言語で記述されたデータセットでの学習に効果があるのかも興味深い点である．メタ学習については，学習データにおける，D_{train}，D_{dev} の組の作成方法については工夫の余地がある．様々な実験設定におけるメタ学習の効果についても検証していきたい．

謝辞

本研究を進めるにあたり，ご指導，ご助言を頂いた乾健太郎教授に心より感謝致します．また，日頃より研究活動や論文執筆において直接ご指導，ご協力を賜りました吉川将司さん，Ana Brassard さん，栗林樹生さん，青木洋一さんに心より感謝致します．さらに，日々の議論の中で多くのご助言を頂きました Tohoku NLP の皆様に感謝致します．

References

[1] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL-HLT*, pp. 2368–2378, Minneapolis, Minnesota, June 2019. ACL.

[2] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *NAACL-HLT*, pp. 107–112, New Orleans, Louisiana, June 2018. ACL.

[3] Henry Conklin, Bailin Wang, Kenny Smith, and Ivan Titov. Meta-learning to compositionally generalize. In *ACL and IJCNLP*, pp. 3322–3335, Online, August 2021. ACL.

[4] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In *EMNLP-IJCNLP*, pp. 5307–5315, Hong Kong, China, November 2019. ACL.

[5] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Investigating the limitations of transformers with simple arithmetic tasks, 2021.

[6] Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. Representing numbers in NLP: a survey and a vision. In *NAACL-HLT*, pp. 644–656, Online, June 2021. ACL.

[7] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *EMNLP*, pp. 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. ACL.

[8] Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Text modular networks: Learning to decompose tasks in the language of existing models. In *NAACL-HLT*, pp. 1264–1279, Online, June 2021. ACL.

[9] Brenden Lake and Marco Baroni. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks, 2018.

[10] Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *EMNLP*, pp. 9087–9105, Online, November 2020. ACL.

[11] Bailin Wang, Mirella Lapata, and Ivan Titov. Meta-learning for domain generalization in semantic parsing. In *NAACL-HLT*, pp. 366–379, Online, June 2021. ACL.

[12] Mor Geva, Ankit Gupta, and Jonathan Berant. Injecting numerical reasoning skills into language models. In *ACL*, pp. 946–958, Online, July 2020. ACL.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pp. 7871–7880, Online, July 2020. ACL.

[14] Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. SHAPE: Shifted absolute position embedding for transformers. In *EMNLP*, pp. 3309–3321, Online and Punta Cana, Dominican Republic, November 2021. ACL.

[15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *ICML*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 06–11 Aug 2017.