時間的常識を理解する言語モデルの構築へ向けて

木村 麻友子　　Lis Kanashiro Pereira　　小林 一郎
お茶の水女子大学
{g1720512,koba}@is.ocha.ac.jp,kanashiro.pereira@ocha.ac.jp

1 はじめに

テキスト内に記述された事象の時間関係を理解するためには，その事象の時間に関する常識的背景知識が必要となる．しかし，常識がテキスト内に明示的に表現されることはほとんどなく，コンピュータにそのような知識を踏まえた理解や推論をさせることは未だ挑戦的な課題となっている．自然言語処理における質問応答に関する研究は，自然言語理解一般を中心に手法の開発が進められ，近年，顕著な進歩を遂げているが，時間的常識を用いた特定の推論能力などを対象にしているものは少ない．そこで，本研究では時間的常識に基づく理解に焦点を当て，Multiple Choice TemporAl COmmon-sense (MCTACO) [1] という自然言語で表現された事象の時間的常識を理解する課題を取り上げ，時間的常識に対する理解の精度向上を目的として手法の開発を行う．本研究でのアプローチとして，時間や常識に関する複数のコーパスを事前学習に用いた multi-step fine-tuning や，BERT 使用時において潜在トークンに対する Masked Language Modeling をする際に使用するデータを変更した場合の出力精度との関係を調査し考察を行った．また，これらの作成したモデルのアンサンブル学習を行い，単一モデルよりも良い結果が得られることを確認した．

2 関連研究

文章内のイベントに関する時間関係を正しく理解することは，自然言語理解の実現において本質的な課題であり，近年，時間的常識を踏まえた自然言語理解の研究が盛んになっている．Zhou ら [1] は，クラウドソーシングを使い時間に関する 5 つの特徴量（duration, temporal ordering, typical time, frequency, stationarity）の時間的常識を集め MCTACO という新しいデータセットを作り，時間的常識問題を解く深層学習モデルを構築したが，人の識別能力よりもおよそ 20%精度が低い結果となった．さらに彼らは，TacoLM という時間的常識を捉えた言語モデルを構築し [2]，通常の BERT [3] よりも時間関係を捉えるタスクの予測精度が高い結果を得ている．

Gururangan ら [4] は，すでに一般的なコーパスで言語モデルを事前学習しているモデルに対して，さらに対象となるドメインやタスクのデータセットで 2 段階目の事前学習を行うことによって，対象タスクの出力の精度が向上されることを示している．

また，Phang ら [5] は，データ量の多い中間的な教師付きタスクを用いて 2 段階目の事前学習を行った．このアプローチを，Supplementary Training on Intermediate Labeled-data Tasks（STILTs）とし，結果として得られるターゲットタスクモデルが改善すること，特にデータの少ないタスクで BERT を使用する場合や，データに制約のある場合に学習が大幅に安定することを示している．

3 提案手法

3.1 multi-step fine-tuning

本研究では，MCTACO を用いて時間的常識を推定する課題を解決するモデルを構築するが，モデルの精度を向上させるため，多段階のファインチューニング（multi-step fine tuning）を行う．事前学習済みの BERT を対象にして，MCTACO ではないが時間的常識に関係がありそうなタスクを採用し，それらを用いて多段階のファインチューニングを行った後に MCTACO のタスクを用いてファインチューニングすることにより，MCTACO における回答の精度向上を目指す．

3.2 Masked Language Modeling

本研究では，BERT の事前学習として採用されている Masked Language modeling（以下，MLM）と Next Sentence Prediction の内のひとつである MLM に関して，MCTACO では訓練データが提供されてい

ないため，検証データを用いてマスクする語彙の比率などを変えて潜在トークンを構築する．これにより，評価に用いるデータ（MCTACO）に更に適応した言語モデルを構築し，モデルの精度向上を目指す．

4 実験

multi-step fine-tuning による影響を調査するため，MCTACO のみ 1 段階でファインチューニングした場合と，他のデータセットを使用して 2 段階でファインチューニングした場合を比較し，精度がどのように変化するかを分析する．また，MLM に使用するデータセットを変更することによる影響を調査するため，MCTACO を使用した場合の精度を求める．さらに，マスクする割合[1)]をいくつか変更した場合の違いを分析する．

4.1 使用データ

本研究では，学習用および評価用データセットとして MCTACO を使用する．また，3.1 における 1 段階目用のデータセットとして，TimeBank[6], MATRES[7], CosmosQA[8], SWAG[9] を使用する．MCTACO，TimeBank，MATRES は時間に関するデータセットであり，CosmosQA と SWAG は時間に限らず一般的な常識全般に関するデータセットである．表 1 にそれぞれのデータセットの統計情報を示す．

表 1 各データセットについて

	訓練データ	検証データ	評価データ
MCTACO	-	3,783	9,442
TimeBank	1,248	-	1,003
MATRES	12,716	-	838
CosmosQA	25,588	3,000	7,000
SWAG	73,546	20,006	20,005

以下に各データセットの概要を示す．

MCTACO [1]

MCTACO では，時間特性に関する 5 つの特徴量 (duration, temporal ordering, typical time, frequency, stationarity) を定義しており，自然言語で表現された事象の時間的常識を理解する課題から構成されるデータセットである．5 つの特徴量のいずれかの特性について記述された文章とその文章に関する質問，それに対する答えを表す複数の選択肢，その選択肢に対して正解には yes，不正解には no とラベル付けされたものから構成されている（表 2 参照）．

1) ディフォルトだと 15%である．

表 2 MCTACO の例

S1:He layed down on the chair and pawed at her as she ran in a circle under it.	
Q1:How long did he paw at her?	
A1:2 minutes [yes]	A2:2 days [no]
A3:90 minutes [no]	A4:7 seconds [yes]
Reasoning Type:Event Duration	

MCTACO では，訓練データは提供されておらず，検証データと評価データのみ提供されている．MCTACO に対しては，huggingface の transformers で提供されている分類問題用のモデルである，BertForSequenceClassification モデルを使用する．また，入力データをエンコードする際に，文章＋質問文と答えは [SEP] で分けられている．

TimeBank [6]

TimeBank は，時間的常識の中の，特に持続時間に関するデータセットである．文章内に含まれるイベントの持続時間が 1 日より長いか短いかによって yes, no のいずれかがラベル付けされている．TimeBank に対しても，BertForSequenceClassification モデルを使用する．また，入力データをエンコードする際に，文章とイベントは [SEP] で分けられている．

MATRES [7]

MATRES は，文章内に含まれる二つの動詞の時間関係に関するデータセットである．時間関係によって，AFTER, BEFORE, EQUAL, VAGUE のいずれかがラベル付けされている．MATRES に対しても，BertForSequenceClassification モデルを使用する．また，入力データをエンコードする際に，着目する文章中の二つの動詞の前後にタグマーカーを挿入している．

MATRES は TimeBank のデータ拡張用としても用いる．その際，MATRES に含まれる時間情報を含むデータに対して TimeBank でファインチューニング済みのモデルを用いてラベル付けを行い，最終層における出力の確率値が 0.9 を超えるもののみをデータ拡張に用いて実験を行う．

CosmosQA [8]

CosmosQA は，出来事の原因や影響など，明示的に言及されていない物語の行間を読むことに焦点を当てている．時間に限らず一般的な常識全般に関するデータセットであり，四択一の多肢選択問題である．CosmosQA に対しては，huggingface の transformers で提供されている多肢選択問題用のモデルである BertForMultipleChoice モデルを使用す

る．また，入力データをエンコードする際に，文章＋質問文と選択肢は [SEP] で分けられている．

SWAG [9]

SWAG は，フレーズを与えられたときに次のフレーズを常識に基づいて推測する問題のデータセットである．こちらも時間に限らず一般的な常識全般に関するデータセットであり，四択一の多肢選択問題である．SWAG に対しても，BertFormultipleChoice モデルを使用する．また，入力データをエンコードする際に，文章と選択肢は [SEP] で分けられている．

4.2 実験設定

multi-step fine-tuning に関して，パラメータの設定を表 3 に示す．それぞれのデータセットについて予備実験を通して最も精度が良くなったパラメータ（表内太字）を使用する．

また，MLM を行う際のパラメータの設定を表 4 に示す．MLM 後に MCTACO を用いて学習，評価する際のパラメータは表 3 の 1 行目に太字で記載のものを使用する．MLM には，huggingface の transformers[2] で提供されている BertForPreTraining モデルを使用する．

両者ともにモデルには bert-base-uncased を使用し，評価指標としては Exact Match (EM) と F1 スコアを採用した．EM は各質問に対する全ての答えを正しくラベル付けすることができる確率であり，F1 スコアは適合率と再現率の調和平均である．

表 3 multi-step fine-tuning の実験設定

	max seq_len	train batch_size	num train_epoch	learning rate
MCTACO	128	{32,**16**}	{3,4,**5**}	{**1e-5**,2e-5}
TimeBank	128	{32,**16**}	{3,**4**,5}	{1e-5,**2e-5**}
MATRES	128	{32,**16**}	{**3**,4,5}	{**1e-5**,2e-5}
TimeBank + MATRES	128	{32,**16**}	{**3**,4,5}	{1e-5,**2e-5**}
CosmosQA	256	32	{**1**,3,5}	{1e-5,**2e-5**}
SWAG	256	32	{1,**2**,3}	{1e-5,**2e-5**}

表 4 MLM の実験設定

max seq_len	train batch_size	num train_epoch	learning rate
128	32	3	3e-5

4.3 実験結果

multi-step fine-tuning 実験結果を表 5 に示す．

2） https://github.com/huggingface/transformers

表 5 multi-step fine-tuning による実験結果

fine-tuned on	EM [%]	F1 [%]
MCTACO	40.9 (42.1)	69.9 (68.2)
TimeBank → MCTACO	41.3 (40.2)	70.3 (67.1)
MATRES → MCTACO	39.6 (42.0)	69.2 (69.4)
TimeBank + MATRES → MCTACO	40.2 (40.9)	70.2 (67.7)
CosmosQA → MCTACO	42.2 (41.7)	70.4 (68.9)
SWAG → MCTACO	**43.0 (42.0)**	**71.7 (67.8)**

表 5 の 1 行目は，MCTACO を用いた 1 段階のファインチューニングの結果，2 行目以下はその他のデータセットを 1 段階目に使用し，MCTACO を 2 段階目として用いた mul を行なった結果である．また，4 行目は MATRES を TimeBank のデータ拡張用に用いた結果である．表 5 には MCTACO の評価データを使用した結果，及び（）内には 5 分割交差検証を行なった結果を記載している．

実験の結果，使用するデータセットによる差異はあるものの全体的には multi-step fine-tuning を行なったことによる精度の向上が確認された．最も良い精度となったのは最下行に記載している SWAG を 1 段階目のファインチューニングに使用した場合であった．CosmosQA と SWAG はどちらも一般的な常識全般に関するデータセットであり，二つを比較すると SWAG の方が大きなデータセットである（表 1 参照）．

Masked Language Modeling 実験結果を表 6 に示す．

表 6 MLM による実験結果

Masking Probability [%]	EM [%]	F1 [%]
15	**44.5 (45.2)**	**71.9 (72.4)**
30	43.5 (44.3)	71.9 (71.3)
60	42.8 (44.6)	71.1 (69.9)

こちらも表 5 と同様に，MCTACO の評価データを使用した結果，及び（）内には 5 分割交差検証を行なった結果を記載している．実験の結果，最も精度が良かったのは 1 行目の，ラベルが yes のデータのみを使用し 15%マスクした場合であった．

ここで，アンサンブル学習として Max Voting を行う．いくつかのパターンで 3 つのモデルを使用し，MCTACO の評価データを用いて評価した．その結果を表 7 に示す．

表 7　アンサンブル学習による実験結果

モデル	パターン 1	パターン 2	パターン 3
MCTACO			✓
TimeBank → MCTACO	✓		
CosmosQA → MCTACO	✓	✓	✓
SWAG → MCTACO	✓	✓	✓
MLM (15%)		✓	
EM [%]	45.0	**45.6**	44.4
F1 [%]	72.9	**73.2**	72.0

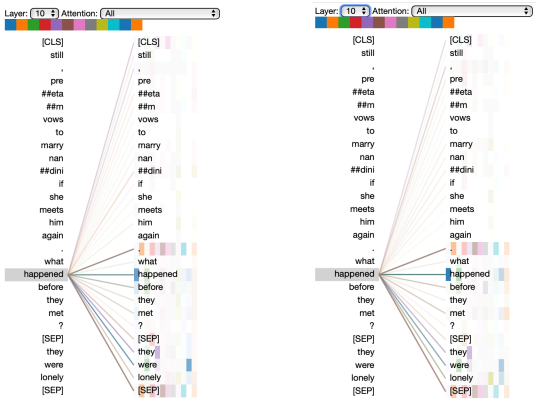
実験の結果，アンサンブル学習による精度の向上が確認された．特に，CosmosQA を用いて multi-step fine-tuning を行ったモデル，SWAG を用いて multi-step fine-tuning を行ったモデル，MLM に MCTACO を用いたモデルの 3 つを使用したパターン 2 が最も精度が良くなった．この結果は，Zhou ら [1] の実験結果（F1:69.9%，EM:42.7%）に比べてそれぞれ 3% ほど精度が向上している．

4.4　考察

multi-step fine-tuning を行うと，1 段階のみのファインチューニングを行なった場合よりも精度が良くなることが確認できた．MCTACO は時間的特徴の理解を問うタスクであるが，1 段階目に使用するデータセットに関しては，時間的な常識に関するデータセットに拘らずに一般的な常識全般に関するデータセット，特に大規模なデータセットである場合に精度が大きく改善することが確認された．

また，MLM に MCTACO を用いると，マスクする割合などによって違いはあるものの，pre-trained BERT モデルをそのまま使用する場合よりも精度が良くなることが確認できた．さらに，multi-step fine-tuning よりも精度が良くなることも確認でき，有用な手段であると考えられる．

ここで，MCTACO のみでファインチューニングしたモデルと，MLM に MCTACO を使用したモデル（表 6 のうち最も精度の良かった 1 行目の設定を使用）に関して，MCTACO の検証データに含まれる例に対し Attention の可視化を行い，観察した．結果を図 1(a)，図 1(b) に示す．Attention の可視化には BertViz[3] を使用した．BertViz では，カーソルを合わせた任意のトークンからの Attention を各レイヤー及びヘッドごとに確認することができ，Attention が大きいほど色が濃く表示される．例えば文中の"happened"という動詞について比較してみると，MLM に MCTACO を用いることによって，"met"という動詞の過去形や"lonely"という形容詞への Attention が大きくなっていることが分かる．Attention の可視化について，詳しい結果を付録 A に記載する．



(a) MCTACO でファインチューニングしたモデルを使用した場合　(b) MLM に MCTACO を用いたモデルを使用した場合

図 1　Attention 可視化の例

5　おわりに

本研究では，自然言語で表現された事象の時間的常識を理解するタスクにおいて，多段階でのファインチューニングを行うことの効果の検証，及び，事前学習 MLM において使用するデータを変更することの効果の検証を行った．実験の結果，使用するデータセットやマスクする割合による差異はあるものの，双方ともに精度の向上が確認された．さらに，MLM は精度がより良くなることが確認された．今後は，Attention の分析を進めながら MLM においてマスクするトークンの選び方を変更するなどの実験を行い，さらなる精度の向上を実現するために有益なドメイン適用手法ならびに言語モデル構築手法を開発する．

謝辞

本研究は，科研費（18H05521）の支援を受けた．ここに謝意を表す．

3）　https://github.com/jessevig/bertviz

参考文献

[1] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3363–3369, Hong Kong, China, November 2019. Association for Computational Linguistics.

[2] Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7579–7589, Online, July 2020. Association for Computational Linguistics.

[3] 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.

[4] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics.

[5] Jason Phang, Thibault Févry, and Samuel R Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018.

[6] Feng Pan, Rutu Mulkar-Mehta, and Jerry R Hobbs. Extending timeml with typical durations of events. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pp. 38–45, 2006.

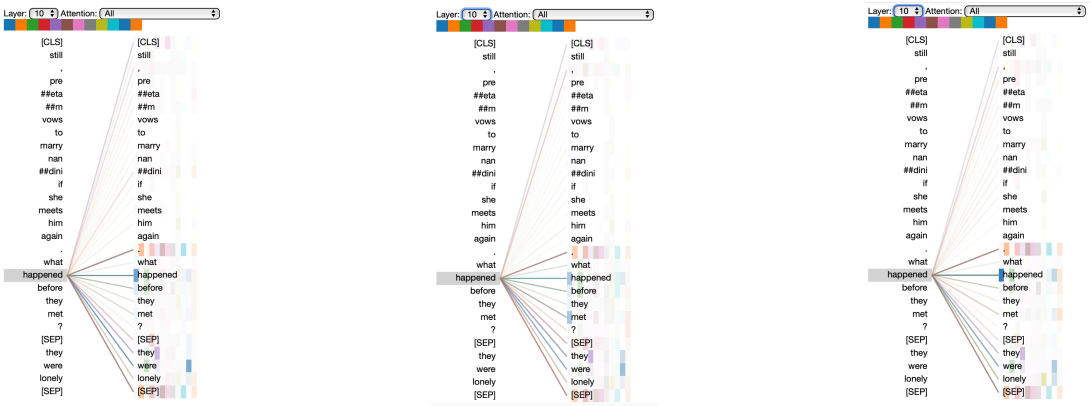
[7] Qiang Ning, Hao Wu, and Dan Roth. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1318–1328, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[8] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics.

[9] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 93–104, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

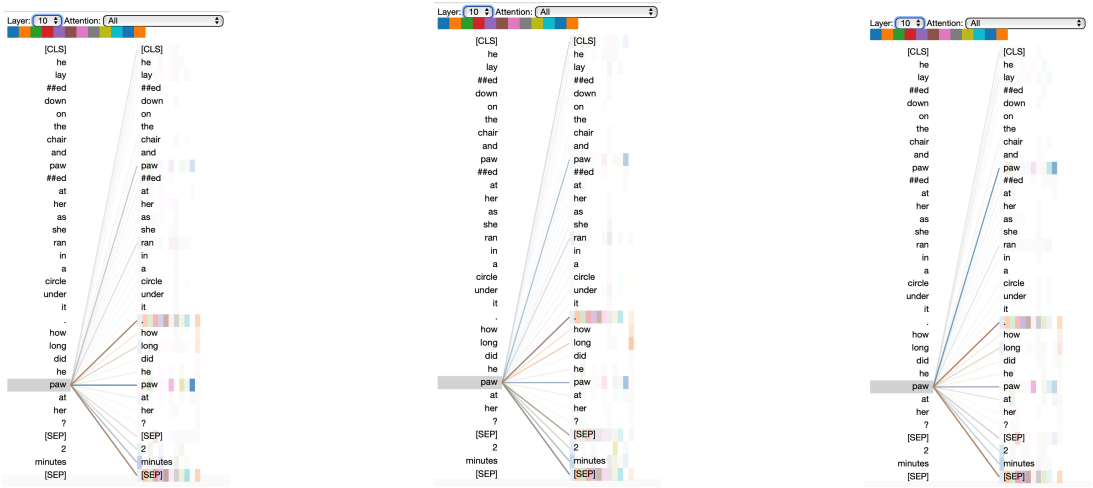
A　付録

本文中 4.4 にて行った Attention の可視化の補足として，MCTACO を用いてファインチューニング（以下，FT）したモデルと MLM に MCTACO を使用したモデルに加えて，MCTACO と SWAG を用いて multi-step fine-tuning（以下，MSFT) を行なったモデルの 3 つに関して，MCTACO の検証データに含まれる例に対して Attention の可視化を行い，どのように変化しているかを観察した．図 2 と図 3 に可視化した結果を示す．



(a) MCTACO を用いて FT したモデルを使用した場合　(b) MCTACO ＋ SWAG で MSFT を行なったモデルを使用した場合　(c) MLM に MCTACO を使用したモデルを使用した場合

図 2　Attention 可視化の例 1



(a) MCTACO を用いて FT したモデルを使用した場合　(b) MCTACO ＋ SWAG で MSFT を行なったモデルを使用した場合　(c) MLM に MCTACO を使用したモデルを使用した場合

図 3　Attention 可視化の例 2

まず，図 2 は，MCTACO を用いて FT したモデルでは正しく予測できなかったが，MCTACO ＋ SWAG で MSFT を行なったモデルと MLM に MCTACO を使用したモデルでは正しく予測できた例に対して各モデルを使用した結果である．例えば文中の"happened"という動詞について確認してみると，(a) と比較して，(b) では，"before"という時間に関する接続詞や"met"という動詞の過去形への Attention が大きくなっていることが分かる．また，(c) では，本文中でも述べた通り"met"や"lonely"という形容詞への Attention が大きくなっていることが分かる．

次に，図 3 は，MCTACO を用いて FT したモデルでは正しく予測できていたが，MCTACO ＋ SWAG で MSFT を行なったモデルと MLM に MCTACO を使用したモデルでは正しく予測できなかった例に対して各モデルを使用した結果である．例えば文中の"paw"という動詞について確認してみると，(a) と比較して，(b) と (c) の双方とも時間に関する単語や動詞等へ Attention が大きくなっていることはあまり確認できない．