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　(指導教員：小林 一郎)

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．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 (MC-TACO) [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 [2] [3] や，BERT 使用時において潜在トークンに対する Masked Language Modeling [4] 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　MC-TACO

MC-TACO 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 1 参照）．

表 1: MC-TACO 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	

3　提案手法

3.1　multi-step fine-tuning

本研究では，MC-TACO 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，MC-TACO 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 MC-TACO 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，MC-TACO 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，MC-TACO の検証データ[1]を用い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（MC-TACO）に更に適応した言語モデルを構築し，モデルの推定精度向上を目指す．

4　実験

multi-step fine-tuning による影響を調査するため，MC-TACO 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 において使用データを変更することによる影響を調査するため，MLM に MC-TACO の検証データを用いた場合の精度を求める．さらに，マスクする割合[2]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 MC-TACO 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 [5], MATRES [6], CosmosQA [7], SWAG [8] を使用する．MATRES は TimeBank のデータ拡張用としても用いる．表 2 にそれぞれのデータセットの統計情報を示す．

表 2: 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	種類
MC-TACO	-	3,783	9,442	時間的常識
TimeBank	1,248	-	1,003	持続時間
MATRES	12,716	-	838	時間関係識別
CosmosQA	25,588	3,000	7,000	一般常識
SWAG	73,546	20,006	20,005	一般常識

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．

表 3: multi-step fine-tuning の実験設定

	max seq_len	train batch_size	num train_epoch	learning rate
MC-TACO	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**}

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 MC-TACO 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

[1]MC-TACO では訓練データが提供されていないため．
[2]ディフォルトだと 15%である．

使用する．両者ともにモデルには bert-base-uncased, Optimizer には Adam 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 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．

表 5: multi-step fine-tuning による実験結果

fine-tuned on	EM [%]	F1 [%]
MC-TACO	40.9 (42.1)	69.9 (68.2)
TimeBank → MC-TACO	41.3 (40.2)	70.3 (67.1)
MATRES → MC-TACO	39.6 (42.0)	69.2 (69.4)
TimeBank + MATRES → MC-TACO	40.2 (40.9)	70.2 (67.7)
CosmosQA → MC-TACO	42.2 (41.7)	70.4 (68.9)
SWAG → MC-TACO	**43.0 (42.0)**	**71.7 (67.8)**
Human Performance	75.8	87.1

表 5 には MC-TACO 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の 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．なお，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 2 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，MC-TACO 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 15%マスクした場合であった．

ここで，Max Voting を行いアンサンブルモデルを得る．いくつかのパターンで 3 つのモデルを使用し，MC-TACO 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: アンサンブルモデルによる実験結果

モデル	パターン 1	パターン 2	パターン 3	パターン 4
MC-TACO			✓	
TimeBank → MC-TACO	✓			
CosmosQA → MC-TACO	✓	✓	✓	
SWAG → MC-TACO	✓	✓	✓	
MLM (15%)		✓		✓
MLM (15%) (random seed 値を変更）*2				✓
EM [%]	45.0	**45.6**	44.4	44.7
F1 [%]	72.9	**73.2**	72.0	72.4

実験の結果，アンサンブルモデルにおける精度の向上が確認された．これは，各単一モデルが異なる有用な特徴を学習していることを示している．また，この結果は，Zhou ら [1] の実験結果（EM:42.7%，F1:69.9%）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，精度が良くなることが確認できた．MC-TACO 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 MC-TACO を用いると，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．

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，及び，事前学習 MLM において使用データを変更することの効果を検証した．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．

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 *EMNLP-IJCNLP*, Hong Kong, China, November 2019.

[2] 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*, Online, July 2020.

[3] 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.

[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)*, Minneapolis, Minnesota, June 2019.

[5] 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.

[6] 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)*, Melbourne, Australia, July 2018.

[7] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *EMNLP-IJCNLP*, Hong Kong, China, November 2019.

[8] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP2018*.