時間的知識の理解に適した汎用言語モデルの構築

理学専攻・情報科学コース
2140652
木村 麻友子

1 はじめに

文章中に表現される時間に関するイベントに対して，常識的な時間関係を捉えることは，自然言語理解においてとても重要な課題である．一方で，近年幅広い自然言語処理タスクで大きな成果を上げている事前学習済み言語モデルは，時間推論においてはまだ性能が低いと言われている [1]．特に困難な課題として，時間的常識を扱う推論が挙げられる．

本研究では，時間的常識推論に対するモデルの開発に焦点を当て，時間的常識を理解するための汎用言語モデルの開発を目指す．複数の事前学習済み言語モデルを対象に，複数のデータセットを用いた実験や，対象とする時間的常識を問うタスクを用いた実験を行い，対象タスクを解くために必要な共通知識を追加したモデルを提案する．また，汎用性にも目を向け，時間に関する複数のタスクにおいても同時に高い性能を発揮できるモデルの構築を目指した．

2 時間的常識タスク MC-TACO

MC-TACO [2] 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，それに対する答えの候補，各候補に対して正解には 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おり，yes か no かを予測する二値分類のタスクである．データセットは news，Wikipedia，textbooks など様々な情報源から取得されている．

3 実験手法

3.1 多段階ファインチューニング

多段階ファインチューニング (multi-step fine-tuning) は，異なるデータを用いてファインチューニングを 2 段階行うことで，モデルの頑健性や性能を向上させる手法である．特に，目的のタスクのデータ数に制限がある場合に，関連するデータセットでのファインチューニングを一段階目に行うことで精度を向上させることが示されている．

3.2 対象タスクに対する継続学習

BERT などの事前学習済み言語モデルは対象タスクに対してファインチューニングを行うだけで良い性能を発揮するが，事前学習されたモデルと対象タスクとの間にドメインの不整合がある場合，タスクの精度向上が見込めない場合がある．この問題を解決するために，対象のデータセットを用いて事前学習を行うことは，事前学習されたモデルを対象タスクに適応させるために有用であることが示されている．これに基づき，事前学習済み言語モデルに対して，通常のファインチューニングを行う前に，言語モデルの事前学習で行われているタスクを，対象タスクを用いて実施する．

表 1: 多段階ファインチューニングの結果

fine-tuned on	EM [%]	F1 [%]
BERT		
standard fine-tuning	42.6 (42.9)	70.9 (71.0)
TimeML → MC-TACO	44.8 (43.7)	72.8 (70.8)
CosmosQA → MC-TACO	**46.3** (43.6)	73.4 (70.7)
SWAG → MC-TACO	46.2 **(44.7)**	**73.6 (72.6)**
RoBERTa		
standard fine-tuning	53.8 (54.4)	75.3 **(77.6)**
TimeML → MC-TACO	51.3 (51.1)	75.7 (76.1)
CosmosQA → MC-TACO	**55.6 (55.2)**	**78.1** (77.3)
SWAG → MC-TACO	53.1 (53.9)	76.1 (77.3)
ALBERT		
standard fine-tuning	55.0 (54.6)	77.1 (77.9)
TimeML → MC-TACO	51.8 (51.3)	77.9 (75.5)
CosmosQA → MC-TACO	**59.5 (58.9)**	**80.3 (78.7)**
SWAG → MC-TACO	52.8 (51.3)	77.3 (74.6)

3.3 マルチタスク学習

マルチタスク学習は，関連する複数のタスクを同時に学習する手法で，モデルの汎化性と性能向上に効果的であることが確認されている．関連タスクの共通性と相違性を利用することで性能を向上させることができるため，自然言語処理の分野において普及が進んでいる [3].

本研究では，MT-DNN [4] を使用する．MT-DNN は，BERT や RoBERTa などのモデルを共有テキストエンコーダ層として組み込むことができるマルチタスク学習フレームワークである．

4 実験

4.1 多段階ファインチューニングの実験

実験設定 多段階ファインチューニングには，MC-TACO の他に，補助タスクとして TimeML，CosmosQA，SWAG の 3 つのデータセットを使用した．TimeML は Duration に関する時間関係のデータセットで，CosmosQA と SWAG は一般常識全般を問うデータセットである．

言語モデルは，$\text{BERT}_{\text{LARGE}}$，$\text{RoBERTa}_{\text{LARGE}}$，$\text{ALBERT}_{\text{xxLARGE}}$ を使用した．

実験結果・考察 実験結果を表 1 に示す．

評価指標には，Exact Match（EM）スコアと F1 スコアを用いた．EM スコアは，モデルが各質問に対するすべての回答候補を正しくラベル付けすることができるかを測定する評価指標である．評価には MC-TACO の評価データを使用し，() 内には 5 分割交差検証の結果を示す．

実験の結果，本手法により MC-TACO における精度が向上することがわかった．特に，CosmosQA や SWAG など，一般常識のデータセットを補助タスクと

表 2: 対象タスクに対する継続学習の結果

	EM [%]	F1 [%]
BERT		
standard fine-tuning	42.6 (42.9)	70.9 (71.0)
MLM (MC-TACO)	45.2 (45.0)	72.5 (71.9)
RoBERTa		
standard fine-tuning	53.8 (54.4)	75.3 (77.6)
MLM (MC-TACO)	51.2 (54.4)	76.2 (77.5)
ALBERT		
standard fine-tuning	55.0 (54.6)	77.1 (77.9)
MLM (MC-TACO)	**59.2 (58.3)**	**79.9 (78.2)**

して使用した場合に良い結果となった．言語モデルに関しては，ALBERT を用いた場合が最も良い結果となった．全体として，ALBERT を用いて，CosmosQA と MC-TACO で多段階ファインチューニングを行った場合が最も良い精度となった．

一般常識のデータセットを補助タスクとして使用した場合に良い結果となった理由として，一般的な常識的推論タスクは時間的事象に関する推論も含んでいる点が挙げられる．例えば，現在の事象の前後にどのような事象が起こりうるか，といった時間的推論は多く含まれており，それらの恩恵を受けている可能性がある．

言語モデルに関して，最も良い結果が出たのは ALBERT を使用した場合だった．この理由として，ALBERT の性能が良いことに加え，事前学習のタスクの違いが考えられる．ALBERT の事前学習では，Masked Language Modeling に加えて，Sentence Order Prediction という入力された 2 文が正しい順番で並んでいるかどうかを判断する二値分類タスクが採用されている．順番を予測するタスクを事前学習で行うことにより，MC-TACO のタスクを解くのに必要な時間的な知識を多く得られたと考える．

4.2 対象タスクに対する継続学習の実験

実験設定 通常のファインチューニングの前に，対象タスクである MC-TACO を使用した別のタスクを行うことを考える．ここでは，BERT の事前学習として採用されている Masked Language Modeling をタスクとして採用する．マスクする単語はランダムに選ばれる．マスクする単語の選び方を任意の方法に変更する実験も行ったが，ここでは要旨のため割愛し，論文にて結果とともに記す．

実験結果・考察 実験結果を表 2 に示す．

実験の結果，本手法により MC-TACO における精度が向上することがわかった．また，多段階ファインチューニング同様，ALBERT を使用した場合に最も良い結果が得られた．これは，対象のデータセットのみで，使用するデータセットを増やすことなくモデルの性能を上げられる効果的な手法である．

4.3 マルチタスク学習の実験

実験設定 マルチタスク学習には，MATRES という Ordering に関する時間関係のデータセットも追加で使用した．言語モデルは，ここまでの実験において最も良い精度を出している $\text{ALBERT}_{\text{xxLARGE}}$ を使用した．

表 3: MT-DNN を用いたマルチタスク学習の結果
(MC: MC-TACO, Ti: TimeML,
MA: MATRES, Co: CosmosQA)

Train\Eval	MC		Ti	MA
dataset	EM [%]	F1 [%]	acc [%]	acc [%]
MC	57.6	80.6	-	-
MC, Ti	**58.1**	79.7	81.0	-
MC, MA	57.3	80.1	-	**75.4**
MC, Co	**59.2**	80.4	-	-
Ti	-	-	81.1	-
MA	-	-	-	74.6

実験結果・考察 ここでは要旨のため，MC-TACO と他のデータセットを用いたペアワイズマルチタスク学習の結果のみ表 3 に示す．他の組み合わせの結果は論文に記す．評価には，時間関係のタスクを用いる．参考に，MT-DNN を用いてシングルタスクで学習した結果も載せる．

実験の結果，使用するデータセットによって，精度が向上する場合と悪化する場合があることがわかった．マルチタスク環境においてはタスクの相性が重要であると言われているため，データセットの分析が必要であると考え，各データセットに含まれるデータの文章ベクトルを可視化することによる分析も行った．要旨のためここでは割愛し，論文に結果とともに記す．

5 おわりに

本研究では，まず時間的常識推論に対するモデルの開発に焦点を当て，時間的常識を理解するための言語モデルの開発を目指した．複数のデータセットを用いた実験や，対象とする時間的常識推論タスクを学習に用いた実験を行い，モデルを提案した．また，汎用性にも目を向け，時間に関する複数のタスクにおいても同時に高い性能を発揮できる，時間的知識の理解に適した汎用言語モデルの構築を目指した．マルチタスク学習を行い，データセットの親和性と精度の関連について分析を行った．今後の課題としては，マルチタスク学習前に，データセットの表層的な特徴を捉えて相性の良いデータセットの組み合わせを見つけられるようにしたい．また，使用するデータセットを増やし，汎用性に焦点を当てたさらなる実験を行なっていきたい．

参考文献

[1] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

[2] 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.

[3] Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods, 2022.

[4] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496, Florence, Italy, July 2019. Association for Computational Linguistics.