時間関係タスクを対象にしたマルチタスク学習におけるデータの親和性の解析

木村 麻友子 [1]　Lis Kanashiro Pereira[1]　浅原 正幸 [2]
Fei Cheng[3]　越智 綾子 [2]　小林 一郎 [1]
[1] お茶の水女子大学　[2] 国立国語研究所　[3] 京都大学
{g1720512,koba}@is.ocha.ac.jp, kanashiro.pereira@ocha.ac.jp
{masayu-a,a.ochi}@ninjal.ac.jp, feicheng@i.kyoto-u.ac.jp

概要

本研究では，時間的常識推論を中心とした時間に関係する複数のタスクにおける言語モデルの汎用性に焦点を当て，マルチタスク学習を行うことにより時間的知識の理解に適した言語モデルの構築を行った．マルチタスク学習を行った結果，使用するデータセットにより精度が向上する場合もあるものの，実験結果にばらつきがあることがわかった．その結果を踏まえ，マルチタスク学習において共有する潜在空間での各データセットの分布を可視化することによる分析を行い，分布が近いデータセット同士を使用した場合に，マルチタスク学習の精度の向上が見られることを確認した．

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．しかしながら，近年，幅広い自然言語処理（NLP）タスクで大きな成果を上げている BERT [1] などの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 [2]．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．例えば，「旅行に行く」と「散歩に行く」という 2 つのイベントが与えられたとき，多くの人間は「休暇は散歩よりも長く，発生頻度も少ない」という時間的常識を持っているが，コンピュータにはこのような時間的常識を用いて推論することが困難である．

先行研究 [3] 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，対象タスクを解くのに必要な常識的知識を付加したモデルを提案した．本研究では，汎用性という点にも目を向け，マルチタスク学習を導入し，時間に関する複数のタスクの精度を同時に向上させつつそれらのタスクに汎用な言語モデルの構築を目指す．また，マルチタスク学習を通じて得られた複数のタスクに共有する潜在空間において，使用した各データセットの文章ベクトルの分布を可視化することによって分析し，マルチタスク学習に使用するデータセットの親和性と精度の関係について考察を行う．

2　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 [4]．

本研究では，MT-DNN [5] を用いてマルチタスク学習を実行し，複数の時間関連タスクに対するモデルの性能を評価する．MT-DNN は，BERT や RoBERTa などのモデルを共有テキストエンコード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．エンコーダ層では全てのタスクで重みを共有し (Shared layers)，その後，それぞれのタスクの学習を行う (Task specific layers)．タスクに特化した層ではタスクごとに重みが更新され，共有されない．図 1 に MT-DNN の構造の概要を示す．

3　実験

3.1　使用データセット

データセットの概要を以下に記す．また，表 1 にそれぞれのデータセットの統計情報を示す．

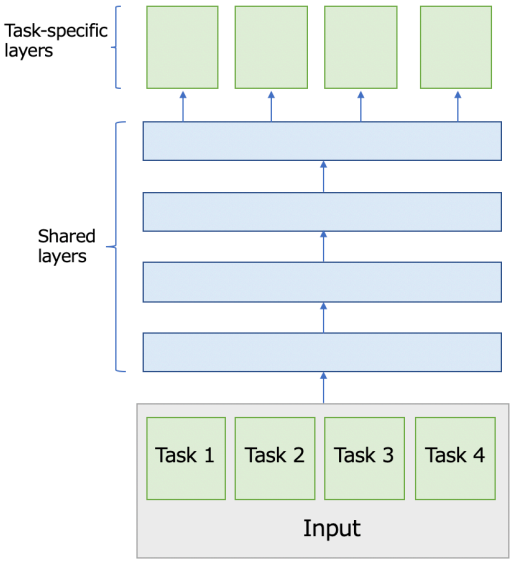

図 1　MT-DNN の概要

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
MC-TACO	-	3,783	9,442
TimeML	1,248	-	1,003
MATRES	12,716	-	838
CosmosQA	25,588	3,000	7,000

MC-TACO [6]:　MC-TACO は，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．MC-TACO 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，これらの特徴量のいずれかの特性が記述に含まれる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が与えられ、各選択肢が答えとしてふさわしいかどうかが判断される．以下に例を示す．ふさわしい選択肢は**太字**で表記する．

Paragraph: He layed down on the chair and pawed at her as she ran in a circle under it.
Question: How long did he paw at her?
a) **2 minutes**　　b) 2 days
c) 90 minutes　　e) **7 seconds**
Reasoning Type: Duration

TimeML [7]:　MC-TACO と同様に時間的常識を問うタスクで，その中でも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それぞれ yes, no のいずれかがラベル付けされている．イベントが 1 日より短い例 (no) を以下に示す．

In Singapore, stocks **hit** a five year low.

MATRES [8]　MATRES は，文章内に含まれる二つの動詞の時間的順序関係に関するデータセットである．時間的順序関係によって，AFTER, BEFORE, EQUAL, VAGUE のいずれかがラベル付けされている．以下に二つの動詞 (e1, e2) のラベルが BEFORE の例を示す．

At one point , when it **(e1:became)** clear controllers could not contact the plane, someone **(e2:said)** a prayer.

CosmosQA [9]:　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．MC-TACO とは違い，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，四肢択一問題である．以下に例を示す．

Paragraph: Did some errands today. My prime objectives were to get textbooks, find computer lab, find career services, get some groceries, turn in payment plan application, and find out when KEES money kicks in. I think it acts as a refund at the end of the semester at Murray, but I would be quite happy if it would work now.
Question: What happens after I get the refund?
Option 1: **I can pay my bills.**
Option 2: I can relax.
Option 3: I can sleep.
Option 4: None of the above choices.

3.2　実験設定

本研究では，事前学習済み言語モデルの一つである ALBERT [10] をテキストエンコーダとして使用する．ALBERT は BERT の派生モデルであり，BERT よりも軽量でありながら GLUE などにおいて精度を改善することに成功している．また，先行研究 [3] において，ALBERT を使用すると，BERT を使用した場合よりも時間的常識タスクにおいて精度が改善することが確認されている．ここでは，ALBERT の中で最も大きい $\text{ALBERT}_{\text{xxLARGE}}$ を使用する．

マルチタスク学習時のハイパーパラメータは，文字列の最大長は 512，バッチサイズは 16，学習率は 1e-5 とし，エポック数は使用するデータセットの組み合わせによって最も良い精度が出たものを採用した．また，MT-DNN を用いた学習時には全てのパラメータを調整する．

表 2　MT-DNN を用いたマルチタスク学習の結果

Train dataset \ Evaluation dataset	MC-TACO		TimeML	MATRES
	EM [%]	F1 [%]	acc [%]	acc [%]
MC-TACO	57.6	80.6	-	-
MC-TACO, TimeML	**58.1**	79.7	81.0	-
MC-TACO, MATRES	57.3	80.1	-	**75.4**
MC-TACO, CosmosQA	**59.2**	80.4	-	-
MC-TACO, TimeML, MATRES	56.3	78.8	79.2	**76.3**
MC-TACO, TimeML, CosmosQA	53.0	76.5	79.9	-
MC-TACO, MATRES, CosmosQA	53.6	78.6	-	**76.8**
MC-TACO, TimeML, MATRES, CosmosQA	53.4	78.2	77.7	**76.8**
TimeML	-	-	81.1	-
TimeML, MATRES	-	-	79.4	**77.2**
TimeML, CosmosQA	-	-	80.4	-
TimeML, MATRES, CosmosQA	-	-	78.8	**76.2**
MATRES	-	-	-	74.6
MATRES, CosmosQA	-	-	-	**74.7**

3.3　実験結果

実験結果を表 2 に示す．MT-DNN を用いてシングルタスク学習を行った場合の結果は青字で表示し，シングルタスク学習の結果を上回った精度は**太字**で表示した．

評価指標として，MC-TACO では独自に定められた Exact Match（EM）スコアと F1 スコア，TimeML と MATRES では Accuracy 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．

実験の結果，MATRES で精度の向上が見られたが，MC-TACO では学習に用いたタスクによって差が生じ，TimeML では向上が見られなかった．

また，使用するデータセットによって，精度が向上する場合（例：MC-TACO と CosmosQA で学習し，MC-TACO で評価した場合の EM スコアは 59.2%）と低下する場合（例：MC-TACO と TimeML と CosmosQA で学習し，MC-TACO で評価した場合の EM スコアは 53.0%）があり，やや不安定な状態であることがわかった．精度が改善する場合とそうでない場合について，なぜこのような結果になるのかについて分析を行った内容を次節に示す．

3.4　データセットの分析

マルチタスク学習の前提条件の 1 つは，異なるタスクとそのデータ間の関連性である．多くの研究では，マルチタスク環境において，正の相関を持つタスクを学習することが望ましいとされている [11]．また，タスクの相性が悪い場合，精度が低下する場合もある [12]．

ここでは，使用した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り，各データセットの親和性と実験結果の相関を明らかにすることを目指す．

設定　マルチタスク学習をして得られた共通のベクトル空間における文章ベクトルを求めるため，MC-TACO，TimeML，MATRES，CosmosQA の 4 つを，$\text{ALBERT}_{\text{xxLARGE}}$ を用いてマルチタスク学習したモデルをエンコーダとする．各データセットからランダムに 1,000 ずつサンプルを取り，各サンプルの文章ベクトルを計算する．ここでは，最終層の隠れ層のベクトルに注目する．ベクトルはトークンごとに出力されるが，文章ベクトルを生成するには，トークンに付与されたベクトルを集約し，単一のベクトルにする必要がある．今回は，各サンプルの全トークンのベクトルを合計してトークン数で割ることで得られたベクトルを平均化し，これを各サンプルの文章ベクトルとする．得られた文書ベクトルを t-SNE[1] と UMAP[2] の 2 つの手法で 2 次元に次元圧縮を行い，可視化する．

可視化結果・考察　比較のために，マルチタスク学習を行わず，$\text{ALBERT}_{\text{xxLARGE}}$ をそのままエンコーダとして用いた場合の文章ベクトルを可視化した結果を図 2 に示す．また，マルチタスク学習したモデルをエンコーダとして用いた場合の文章ベクトルを可視化した結果を図 3 に示す．

マルチタスク学習後のベクトル空間では，青で表示された MC-TACO の文章ベクトルと，緑で表

1）　https://lvdmaaten.github.io/tsne/
2）　https://umap-learn.readthedocs.io/en/latest/

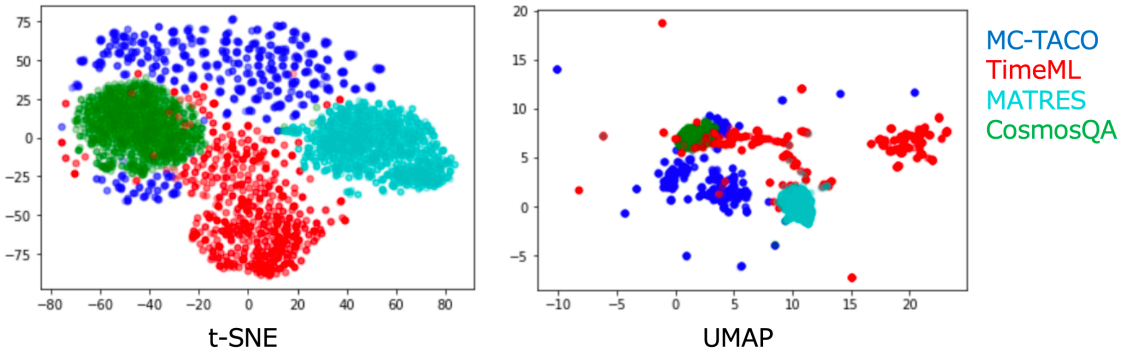

図 2　$\text{ALBERT}_{\text{xxLARGE}}$ をそのままエンコーダとして用いた場合の文章ベクトルの可視化結果

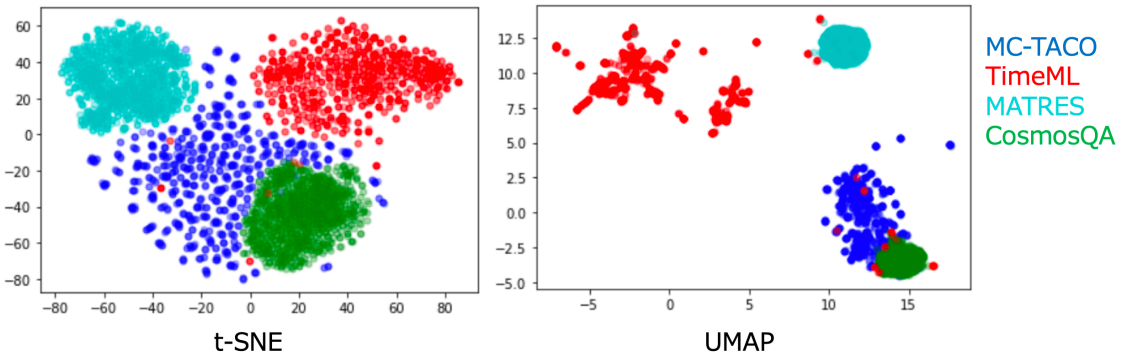

図 3　マルチタスク学習後のモデルをエンコーダとして用いた場合の文章ベクトルの可視化結果

示された CosmosQA の文章ベクトルが，ベクトル空間内で近い位置に分布していることがわかる．t-SNE では少し分かりにくいが，UMAP では顕著であることが確認できる．また，赤で表示している TimeML は，一部のサンプルに分布のばらつきがあり，MC-TACO と CosmosQA の分布に近い位置に文章ベクトルが位置するサンプルもあることがわかる．MC-TACO と CosmosQA をマルチタスク学習した場合，MC-TACO で評価した場合の EM スコアが上がっていたこと，また，MC-TACO と TimeML をマルチタスク学習した場合も，MC-TACO で評価した場合の EM スコアがわずかではあるものの上がっていたことを踏まえると，学習後の共有されたベクトル空間内において文章ベクトルの分布が近い場合に，データセットの親和性が高いと言え，そのような場合にマルチタスク学習の効果が出ると考えることができる．

一方，MATRES では，すべてのマルチタスク設定で精度が向上しているが，文章ベクトルの分布は他のデータセットと近いようには見えない．MATRES は，文章ベクトルの分布が広くなく，まとまっている．そのようなデータセットを対象にマルチタスク学習を行うことで，補助データセットを有益なノイズとして取り入れた敵対的学習が行われ，精度が改善した可能性もあると考えられる．

4　おわりに

本研究では，マルチタスク学習を行うことにより，時間的知識の理解に適した言語モデルの構築を目指した．実験の結果，使用するデータセットにより精度が向上する場合がある一方，精度が下がる場合もあり，結果にばらつきがあることがわかった．その点を踏まえ，各データセットの文章ベクトルを可視化することによる分析を行い，共有ベクトル空間における文章ベクトルの分布が近いデータセット同士を使用した場合に，マルチタスク学習の精度の向上が期待できることを確認した．今後の課題としては，文章ベクトルの分布がまとまっているデータセットにはマルチタスク学習が有効であるという考察の検証をすること，また，複数の時間関係のタスクで同時に精度を向上できるような汎用言語モデルの構築のため，使用するデータセットの数を増やし，さらなる分析と実験を重ねることが考えられる．また，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る試みも検討したい．

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．

参考文献

[1] 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.

[2] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics.

[3] Mayuko Kimura, Lis Kanashiro Pereira, and Ichiro Kobayashi. Toward building a language model for understanding temporal commonsense. In **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop**, pp. 17–24, Online, November 2022. Association for Computational Linguistics.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. **arXiv preprint arXiv:1909.11942**, 2019.

[11] 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.

[12] Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning, 2021.