日本語の時間的常識を理解する言語モデルの構築を目的としたマルチタスク学習における検証

船曳 日佳里 [1]　Lis Kanashiro Pereira[1]　木村 麻友子 [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
{g1820534,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，日本語での時間的常識の推論に対する言語モデルの開発を目的として，マルチタスク学習における検証を行った．複数の時間的常識を問うタスクを用いて，テキストエンコーダを変更して実験を行なった．実験の結果，対象タスクの組み合わせによって，時間的常識を問うタスクにおいて，3 %〜10 %程度の精度の向上を確認できた．また，使用するテキストエンコーダも大きく影響を与えることが確認できた．

1　はじめに

自然言語の文章を理解するうえで，イベントが意味する時間の理解は重要である．しかし，イベントの時間を意味する直接的な表現は文章内では省略されやすく，時間の理解のためには，自然言語で表現されるイベントのさまざまな時間的側面について常識的な知識を持っている必要がある．例えば，「いつの間にか眠っていて夢を見たんだ」という文章を読んで，我々は「眠る」と「夢を見る」というイベントが過去のことで，「眠る」の方が「夢を見る」よりも長い時間がかかることやこの二つのイベントが同時に起こることを理解できる．このような常識を踏まえた理解や推論をコンピュータにさせることは挑戦的な課題となっている．近年，BERT[1] などの事前学習済み言語モデルが幅広い自然言語処理タスクで大きな成果を上げているが，これらのモデルは時間推論においては未だ性能が低いと言われており [2]，汎用言語モデルを改善し時間的な常識におけるタスクの精度を上げる試みがなされている [3][4]．しかし，日本語に関する時間的な常識を捉えた研究は未だ少ない．

そこで，我々は日本語における時間的常識に基づく理解に焦点を当てて研究を進めており，本研究では時間的常識に関するタスクを用いたマルチタスク学習を行なう．具体的には，文章のイベントの時制・時間幅・時間順序・事実性を推定するタスクを組み合わせて学習させ，より効果的な学習方法を検討する．

2　提案手法

本研究では，時間的常識に関するタスクを用いたマルチタスク学習を行なって，時間的常識のタスクに関して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．

時間的常識のタスクとして，イベントの時制・時間幅・時間順序・事実性の四つを設定する．時制の推定は，イベントが発話時と比べて過去・現在・未来のいずれであるかを推定する．時間幅の推定は，イベントがどのくらいの時間を要するのかを推定する．時間順序の推定は，二つのイベントの時間的順序関係もしくは重なりについて推定する．事実性の推定は，表現が実際に起きたイベントなのか，条件節などの仮想的なイベントなのかを推定する．

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ことで，モデルの汎化性能を向上させることを目的としている．関連するタスクの共通性と差異を利用することで性能を向上させることができるため，自然言語処理において普及が進んでいる [5]．本研究では，対象データセットのデータサイズが限られているため、学習時に補助的なデータセットを利用することによって、この問題を緩和できるマルチタスク学習が有効なアプローチであると考え，この手法を採用した．

また，複数の事前学習済み日本語言語モデルを使

表 1　DVD データセットの例

文章：悪いやつらに追われてるって話すんだ			
イベント		**時制**	**時間幅**
追わ		現在	DATE
話す		未来	TIME
イベント A	**イベント B**	**時間順序**	
追わ	話す	A<=B	

表 2　日本語話し言葉コーパスの例

文章：一番今まで人生の中でつらかったことについてお話ししたいと思います			
イベント	**時制**	**時間幅**	**事実性**
つらかっ	過去	1 年以上	現実
話し	未来	1 分以上 1 時間未満	仮想
思い	現在	1 秒以上 1 分未満	現実

用して実験を行ない，日本語の時間的常識の推論タスクに最も適したものについて検証する．

3　実験

マルチタスク学習を行なった際の各タスクにおける精度を求める．さらに，タスクの組み合わせを変更した際の違いを分析する．また，言語モデルにおける差異に関しても検証する．

3.1　使用データ

本研究では，DVD データセットを対象データセットとして使用し，日本語話し言葉コーパスをマルチタスク学習の補助データセットとして使用している．二つのデータセットからはそれぞれ三つずつの時間に関連する分類タスクを採用している．DVD データセットと日本語話し言葉コーパスはどちらも書き言葉ではないという点で，似ているデータセットとして本研究で採用した．

DVD データセット　DVD データセットは，DVD の音声データの書き起こし文に対して時間に関するラベルを付与したデータセットである．DVD は海外の映画やドラマの日本語吹替版や日本のアニメなどを使用している．タスクは，時制と時間幅と時間順序の三つの分類タスクを使用した．例文を表 1 に示す．それぞれのタスクの分布を付録の表 6 に示す．

日本語話し言葉コーパス　日本語話し言葉コーパスは，本人の体験談を語っている様子を録音したデータの書き起こし文に対して時間に関するラベルを付与したデータセットである．

表 3　実験に使用したテキストエンコーダの概要

Text Encoder	# Parameters	# Layers
$\text{BERT}_{\text{BASE}}$	110M	12
$\text{ALBERT}_{\text{BASE}}$	11M	12
$\text{RoBERTa}_{\text{BASE}}$	111M	12
$\text{XLM-R}_{\text{BASE}}$	270M	12
$\text{XLM-R}_{\text{LARGE}}$	550M	24

表 4　学習の際のハイパーパラメータ

batch size	learning rate	# epochs
16	1e-5	10

タスクは，時制と時間幅と事実性の三つの分類タスクを使用している．例文を表 2 に示す．それぞれのタスクの分布を付録の表 6 に示す．

3.2　テキストエンコーダ

我々は，四つの事前学習済み日本語対応言語モデルのテキストエンコーダを用いて実験を行なった．日本語 BERT モデル cl-tohoku/bert-base-japanese ($\text{BERT}_{\text{BASE}}$)，日本語 ALBERT モデル ALINEAR/albert-japanese-v2 ($\text{ALBERT}_{\text{BASE}}$)，日本語 RoBERTa モデル megagonlabs/roberta-long-japanese ($\text{RoBERTa}_{\text{BASE}}$)，多言語 RoBERTa モデル xlm-roberta-base ($\text{XLM-R}_{\text{BASE}}$)，xlm-roberta-large ($\text{XLM-R}_{\text{LARGE}}$) を使用した．これらは全て Hugging Face で公開されている [1)]．

$\text{BERT}_{\text{BASE}}$　2.6GB の日本語 Wikipedia コーパスを用いて訓練された BERT モデルである．事前学習には Masked Language Modeling (MLM) と Next Sentence Prediction (NSP) が採用されている [1]．

$\text{ALBERT}_{\text{BASE}}$　BERT よりもパラメータが大幅に削減されて軽量化されているが，事前学習で NSP の代わりに Sentence Order Prediction (SOP) を用いることで性能を上げている [6]．

$\text{RoBERTa}_{\text{BASE}}$　Common Crawl 多言語コーパスから抽出された日本語テキストの約 200M 文を用いて訓練された RoBERTa モデルである．BERT の事前学習を見直して NSP を排除し，訓練データのサイズを大幅に大きくすることで性能を上げている [7]．

$\text{XLM-R}_{\text{BASE}}$　RoBERTa の多言語版である．日本語を含む 100 言語を 2.5TB の Common Crawl 多言語コーパスで事前学習されている [8]．

$\text{XLM-R}_{\text{LARGE}}$　$\text{XLM-R}_{\text{BASE}}$ のレイヤー数を 24 に増やしたモデルである．

1）　https://huggingface.co/models

表 5　マルチタスク学習による実験結果

Text Encoder	Model	時制				時間幅				時間順序		事実性	
		話し言葉		DVD		話し言葉		DVD		DVD		話し言葉	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
BERT BASE	STD	74.16	61.11	69.26	55.55	41.12	21.69	47.44	32.32	41.84	19.73	86.98	60.21
	ALL	70.90	63.34	70.95	50.30	41.71	21.90	50.30	36.51	45.71	27.93	82.98	**70.18**
	DVD(時制, 順序), 話し言葉 (時制)	74.20	59.79	72.50	58.11	-	-	-	-	47.23	27.29	-	-
	DVD (ALL), 話し言葉 (時間幅)	-	-	71.84	58.31	43.13	24.25	50.65	35.10	46.39	26.82	-	-
	DVD(ALL)	-	-	69.73	53.42	-	-	51.41	33.34	45.58	29.22	-	-
ALBERT BASE	STD	71.28	55.48	69.88	56.12	39.35	20.25	47.86	34.18	43.34	23.78	87.36	60.96
	ALL	71.78	57.62	69.89	52.90	37.94	20.57	48.78	36.15	45.21	27.26	87.37	63.37
	DVD(時制, 順序), 話し言葉 (時制)	71.57	56.59	69.49	56.03	-	-	-	-	47.70	29.42	-	-
	DVD (ALL), 話し言葉 (時間幅)	-	-	69.87	52.65	39.95	22.71	51.14	39.60	46.45	29.93	-	-
	DVD (ALL), 話し言葉 (時制)	71.34	56.04	69.85	52.79	-	-	50.23	37.87	48.05	30.57	-	-
RoBERTa BASE	STD	71.51	56.46	68.49	55.07	39.29	22.56	46.67	34.49	39.50	22.38	87.89	63.40
	ALL	70.70	56.77	70.78	53.37	38.65	23.47	51.67	40.07	42.92	27.65	87.38	66.00
	DVD(時制, 順序), 話し言葉（時制）	72.48	58.94	69.66	52.57	-	-	-	-	43.43	28.85	-	-
	DVD(ALL), 話し言葉 (時間幅)	-	-	69.82	52.80	38.43	22.65	51.46	40.14	41.20	26.36	-	-
	DVD(ALL), 話し言葉（時制）	71.02	56.11	69.83	52.74	-	-	50.16	38.17	42.88	27.91	-	-
XLM-R BASE	STD	73.13	55.51	74.87	60.17	37.24	12.35	51.69	34.66	45.15	13.88	87.18	54.03
	ALL	73.67	62.11	73.78	59.42	42.12	22.16	53.04	36.51	45.49	22.91	88.02	67.78
	DVD(時制, 順序), 話し言葉（時制）	74.30	62.36	75.26	60.58	-	-	-	-	46.71	22.91	-	-
	DVD(ALL), 話し言葉 (時間幅)	-	-	75.74	60.84	43.00	25.94	54.87	41.46	45.03	22.64	-	-
	DVD (ALL), 話し言葉（時制）	74.08	60.90	75.49	60.84	-	-	53.75	37.52	44.81	23.79	-	-
XLM-R LARGE	STD	**76.64**	**65.00**	**78.26**	62.84	37.98	21.75	50.46	28.97	44.21	18.04	87.38	54.62
	ALL	73.15	57.48	69.13	48.79	40.71	23.06	52.36	37.06	48.51	27.46	**88.77**	67.39
	DVD(時制, 順序), 話し言葉（時制）	76.08	64.97	76.60	**63.23**	-	-	-	-	**50.14**	**32.54**	-	-
	DVD(ALL), 話し言葉 (時間幅)	-	-	76.42	57.87	**43.52**	**28.63**	**56.45**	**44.33**	46.91	29.28	-	-

3.3　実験設定

本研究では，Multi-Task Deep Neural Networks (MT-DNN)[9] 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した．MT-DNN はマルチタスク学習フレームワークであり，BERT などの言語モデルのテキストエンコーディング層をすべてのタスクで共有して組み込むことができ，上位層はタスクに特化したものとなっている．我々は，事前学習済みの言語モデルを用いて共有する層を初期化し，マルチタスク学習によって複数の時間関連タスクで改良した．ハイパーパラメータの設定は表 4 に示す．Optimizer には Adam [10] を使用し，評価指標としては Accuracy (ACC) と F1 スコアを採用した．F1 スコアは適合率と再現率の調和平均である．

3.4　実験結果

マルチタスク学習をした実験結果を，表 5 に示す．Model の列は，マルチタスク学習で使用したタスクを記している．STD は standard fine-tuning，ALL は六つの全てのタスクでマルチタスク学習を行なった結果である．今回は精度が大きく上がった組み合わせを中心に記載し，他の組み合わせの結果を省く．結果は全て 5 分割交差検証をした値である．

結果として，マルチタスク学習を行なった方が STD に比べて精度が上がる場合が多かった．特に時間幅や時間順序のタスクに関しては XLM-R_{LARGE} において STD と比較して 15 %程度 F1 スコアの値を上げている．また，$ALBERT_{BASE}$ に関しては，

$BERT_{BASE}$ や $RoBERTa_{BASE}$ よりもはるかに小さなモデルであるにもかかわらず、時間順序のタスクなどで精度を上回ることができた．$RoBERTa_{BASE}$ よりも $XLM\text{-}R_{BASE}$ の方が少し精度を上回ることも確認できた．

4 解析

マルチタスク学習を行なう前と行なった後のモデルの最終層の隠れ状態を低次元空間に可視化して解析を行なった．$BERT_{BASE}$ では [CLS] トークン，$XLM\text{-}R_{LARGE}$ では <s> トークンを使用している．可視化には UMAP[2] を使用した．本実験で一番よい精度を得られた $XLM\text{-}R_{LARGE}$ のテキストエンコーダを用いて，DVD データセットの三つの全てのタスクと日本語話し言葉コーパスの時間幅タスクでマルチタスク学習させたモデルを使用した．$XLM\text{-}R_{LARGE}$ と $BERT_{BASE}$ のマルチタスク学習前が比較できるよう可視化をした．時間順序のタスクにおける可視化の結果を図 1 に示す．

$XLM\text{-}R_{LARGE}$ のマルチタスク学習後のモデルはばらつきが少なくなっていて，時間的特徴に基づくイベントのクラスタリングをより良く行うことができることがわかる．さらに $BERT_{BASE}$ のマルチタスク学習前のモデルと比べるとその違いが顕著に確認できる．

5 考察

時間的常識に関するタスクにおいて，マルチタスク学習は通常のファインチューニングよりも効果があることが確認できた．また，単純に様々なタスクを増やすよりもタスクを組み合わせた方が効果が高かったため，補助タスクの組み合わせが重要だと考えられる．結果から，違うデータセットの同じタスクを増やすことでそのタスクの精度を上げられることが確認できた．しかし，それが最も精度を上げる組み合わせではないため，他にも要因があると考えられる．また，時制のタスクに関しては $BERT_{BASE}$ 以外のテキストエンコーダを使用した際の効果が薄く，マルチタスク学習に適したタスクではない，または他の種類のタスクを使用した実験が必要だと考えられる．テキストエンコーダに関しては，$XLM\text{-}R_{LARGE}$ が一番精度を上げる結果となった．$XLM\text{-}R_{BASE}$ も良い結果であることから，モデルのサイズが大きいだけでなく，多言語モデルであることも精度を上げる要因になると考えられる．

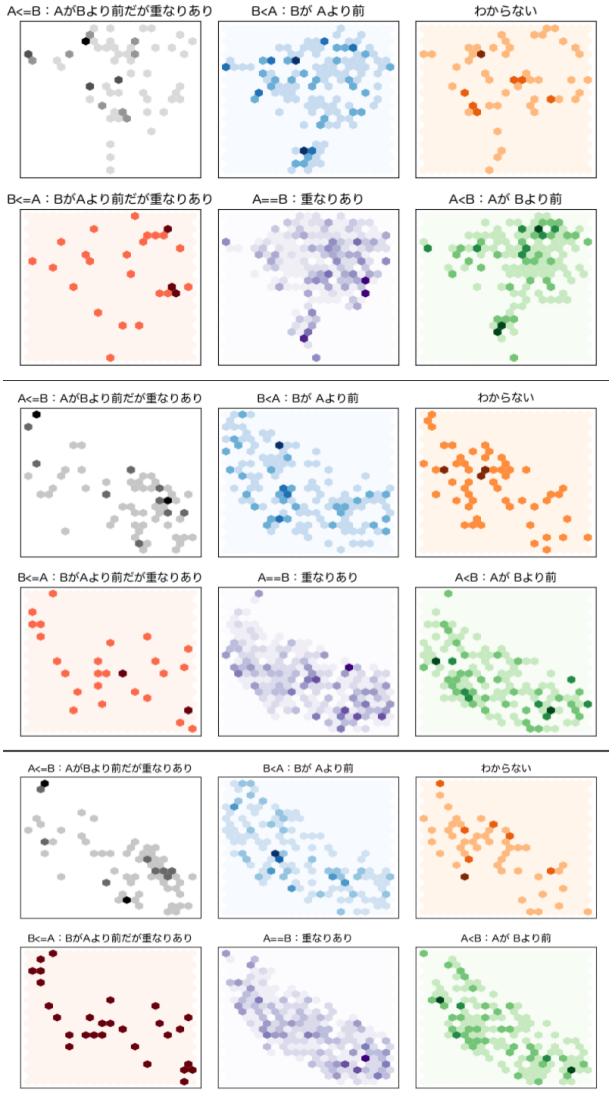

図 1 時間順序タスクにおける最終層の隠れ状態の可視化．$BERT_{BASE}$(上) と $XLM\text{-}R_{LARGE}$ のマルチタスク学習前 (中央) と後 (下)

6 おわりに

本研究では，自然言語で表現されたイベントの時間的常識を理解するタスクにおいて，マルチタスク学習を行なうことの効果を検証した．実験の結果，タスクの組み合わせや使用するテキストエンコーダによって精度の向上が確認された．今後はマルチタスク学習の補助データセットの種類を増やし，さらに効果的な組み合わせの傾向などを検証していきたいと考えている．

2） https://umap-learn.readthedocs.io/en/latest/

謝辞

本研究は，科研費（18H05521，20H05054）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 **Proc. of NAACL-HLT2019**, pp. 4171–4186, June 2019.

[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] Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. Temporal Common Sense Acquisition with Minimal Supervision. In **Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)**, 2020.

[4] 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.

[5] 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.

[6] 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.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. **arXiv preprint arXiv:1911.02116**, 2019.

[9] 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.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

A　付録

表 6　データセットの分布

Dataset	Task	Labels	# Samples	# Total Samples
DVD データセット	時制	現在	668	1,942
		過去	492	
		未来	739	
		わからない	43	
	時間幅	瞬時〜1 秒未満	809	1,942
		TIME	359	
		DATE	127	
		STATE	399	
		UNKNOWN	248	
	時間順序	A<B：A が B より前	273	1,508
		B<A：B が A より前	242	
		A<=B：A が B より前だが重なりあり	116	
		B<=A：B が A より前だが重なりあり	46	
		A==B：重なりあり	718	
		わからない	113	
日本語話し言葉コーパス	時制	現在	1,381	4,241
		過去	2,466	
		未来	394	
	時間幅	瞬時〜1 秒未満	377	4,241
		1 秒以上 1 分未満	266	
		1 分以上 1 時間未満	391	
		1 時間以上 1 日未満	200	
		1 日以上 1 年未満	179	
		1 年以上（常に成り立つは除く）	1,425	
		常に成り立つ	1,071	
		わからない	332	
	事実性	現実	3,722	4,242
		仮想	520	

データセットの各タスクにおける分布情報を表 6 に示す．

DVD データセットの時間幅の TIME は {1 秒以上 1 分未満, 1 分以上 1 時間未満, 1 時間以上 1 日未満 }，DATE は {1 日以上 1 年未満, 1 年以上（常に成り立つは除く）}，STATE は { 常に成り立つ }，UNKNOWN は { わからない } のように，ラベルをまとめたものである．