時間的常識を理解する日本語汎用言語モデルの構築へ向けて

船曳 日佳里　(指導教員：小林 一郎)

1　はじめに

自然言語で表現された出来事を理解するには，時間を理解することが重要である．しかし，それらは記載が省略されることが多々あるため，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．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，本研究では日本語における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，日本語文章中の時間的常識を理解するための日本語汎用言語モデルを開発し，英語でのモデルとの識別精度の比較を行う．

本研究での具体的なアプローチとして，BERT の事前学習において潜在トークンに対する Masked Language Modeling [2] をする際に，ランダムでマスクした場合とマスクする単語を工夫した場合の出力精度との関係を日英の汎用言語モデルにおいて調査し，考察を行う．

2　時間的常識データセット

2.1　MC-TACO

MC-TACO は，時間的常識に関する 5 つの特徴量 (du-ration, 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．それら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，その答えに対して正解には yes，不正解には no とラベル付けされたものの 5 つ組で構成されている (表 1 参照)．また，一つの文章に対して複数の質問が用意され，その質問に対する答えも複数個用意されている．総数 13,225 組のデータセットで，検証データ 3,783，評価データ 9,442 と分けている．

表 1: MC-TACO の例

Sentence	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?
Answer	2 minutes
Label	yes
Type	Event Duration

2.2　翻訳手法

MC-TACO は英語による時間的常識データセットであるため，本研究ではそれを日本語に翻訳して使用した．翻訳には，機械翻訳システム DeepL Pro [1] を用いて翻訳した．次に，一つの文章に二人のアノテーターを割り振り，翻訳の誤りを修正させた．そして，二人の修正が食い違う部分は著者が確認し，修正した．今回，DeepL で翻訳した際に，英語のデータセットとのパラレルコーパスとして利用することにおいて二つの問題が発生した．一つは，データセットの都合上複数出てくる同じ英文が違う日本語文に翻訳されてしまうことがあった．翻訳したデータセットは，英語版の約 4 倍の種類の文章のデータセットと認識されてしまった．これに対して，著者が最も適した翻訳文を選んで統一することで対応した．もう一つは，MC-TACO を作成する際に同義語や類義語を使用して増やされた回答が全て同じ日本語文に翻訳されてしまい，同じデータが作成されてしまったことである．これは，意味が変わらない程度の微小な差異を加えることで対応した．

[1]https://www.deepl.com/

3　提案手法

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.

MLM とは，入力の一部のトークンをターゲットとしてマスクし、元のトークンを左から右と右から左の両方向から見て予測する穴埋め問題である．

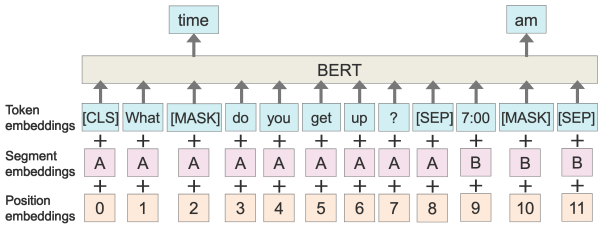

図 1: Masked Language Modeling

4　実験

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，マスクする割合をいくつか変更した場合の違いを分析する．

4.1　実験設定

MLM を行う際，及び，MLM 後に MC-TACO を用いて学習，評価する際のパラメータの設定を表 2 に示す．また，比較対象である英語での実験設定は木村ら [3] に準ずる．

表 2: 実験設定

	max seq_len	train batch_size	num train_epoc	learning rate
MLM	128	32	3	3e-5
評価	128	16	9	1e-5

モデルには東北大学の乾研究室が公開している日

本語 BERT モデル bert-base-japanese-whole-word-masking, Optimizer には Adam [4] 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．

4.2 実験結果

日英 MC-TACO データセットの評価

実験結果を表 3 に示す.

表 3: 英語と日本語の比較

fine-tuned on	EM [%]	F1[%]
MC-TACO	40.9 (42.1)	69.9 (68.2)
日本語 MC-TACO	33.9 (41.0)	61.2 (65.3)

() 内は 5 分割交差検証の結果を記載.

Masked Language Modeling

ランダムに選んだトークンをマスクした場合の結果を表 4 に示す．英語での実験では全トークンの 15 % をマスクした場合が最も精度が良く，日本語でも同様に 15 %マスクした場合が最も精度が良い結果を得られた.

表 4: ランダムマスクによる実験結果

Masking Rate [%]		EM [%]	F1[%]
English	15	44.5 (45.2)	71.9 (72.4)
Japanese	15	36.5 (42.2)	65.9 (66.4)
	30	36.1 (38.7)	65.9 (65.3)
	60	35.8 (39.6)	64.1 (63.9)

() 内は 5 分割交差検証の結果を記載.

次に，時間関係の単語を多くマスクした場合の実験結果を表 5 に示す．時間関係の単語は，日本語 MC-TACO のデータを確認し手動で定義する．主な内容としては，数字，形容詞，副詞，時間の単位で，総数 247 語である.

表 5: 時間関係の単語のマスクによる実験結果

Masking Rate [%] (Temporal Words)		Masking Rate [%] (Others)	EM [%]	F1[%]
English	80	20	45.1(44.3)	72.7(70.7)
Japanese	100	0	35.7(38.6)	64.0(63.7)
	90	10	37.1(40.3)	65.9(62.1)
	80	20	36.9(40.2)	66.1(63.3)
	70	30	35.5(36.6)	64.8(62.3)

() 内は 5 分割交差検証の結果を記載.

これは，全トークンの約 30 %をマスクした場合の結果である．英語での実験では 8:2 の割合で時間関係の単語を多くマスクした場合が最も精度が良かったが，日本語では 9:1 の割合でマスクした場合も精度が良かった.

最後に，Saliency の値が大きい単語をマスクした場合の実験結果を表 6 に示す．Saliency とは各単語の埋め込みベクトルが最終的な判断にどれだけ寄与しているかを表す指標である [5]．Saliency を求める際には損失と単語の埋め込みベクトルの勾配を求めることになるが，本研究では，今回使用するモデルである BERT の入力を構成する，Token Embeddings, Position Embeddings, Segment Embeddings の 3 種類の埋め込みベクトル全てを使用して Saliency を求めた.

表 6: Saliency に着目したマスクによる実験結果

Masking Rate [%] (Saliency Words)		Masking Rate [%] (Others)	EM [%]	F1[%]
English	90	10	43.8 (43.2)	70.7 (70.7)
Japanese	100	0	35.1 (39.9)	66.2 (64.4)
	90	10	34.9(36.5)	63.1(63.3)
	80	20	34.8(36.8)	62.4(64.9)

() 内は 5 分割交差検証の結果を記載.

これは，全トークンの約 15 %をマスクした場合の結果である．英語での実験では 9:1 の割合で Saliency の値が大きい単語を多くマスクした場合が最も精度が良かったが，日本語では Saliency の値が大きい単語のみをマスクした場合が最も精度が良かった.

4.3 考察

他のタスクと同様に，時間的常識に関するタスクでも日本語は英語と比較して難しいことが確認できた．また，日本語の場合でも 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．その際，マスクするトークンを時間関係の単語を多めに選ぶことでさらに精度が良くなることも確認できた.

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，事前学習 MLM においてマスクするトークンを工夫することの効果を検証し，英語での研究との比較を行った．実験の結果，日本語での実験は英語に比べて精度が劣るものの，ベースラインの精度と比較して精度の向上が確認された.

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. *arXiv preprint arXiv:1909.03065*, 2019.

[2] 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.

[3] 木村麻友子, Lis Kanashiro Pereira, 浅原正幸, Fei Cheng, 越智綾子, 小林一郎. 時間的常識理解へ向けた効果的なマスク言語モデルの検証. 第 28 回言語処理学会年次大会, 2022.

[4] 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.

[5] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In *Proc. of NAACL-HLT2016*, pp. 681–691, San Diego, California, June 2016. Association for Computational Linguistics.