日本語の時間的常識の理解への取り組み

理学専攻・情報科学コース
2240668
船曳 日佳里

1 はじめに

自然言語で表現されるイベントに対して，常識的な時間関係を捉えることは，自然言語処理タスクにおいて重要な課題である．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おり，時間的常識の理解に適した汎用言語モデルを構築する試みがなされている [1]．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．

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て研究を進めており，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や，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．また，イベントの時間情報は離散的なラベルで一意に決められるものではなく，分布として捉えるべきだと考え，イベントの時間分布は正規分布で表現し，二つのイベントの時間分布からそのイベント間の正しい時間関係を推定することを目指した．

2 時間的常識に関するデータセット

2.1 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，各答えに対する正解，不正解がラベル付けされたものから構成されている．我々は，このデータセットを日本語に翻訳して実験に使用した．

2.2 DVD データセット

DVD データセット [3] 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．一つのイベントの絶対時制 (過去，現在，未来) と時間幅 (MOMENTARY，TIME，DATE，STATE)，二つのイベント A，B の時間順序 (A<B，B<A，A≤B，B≤A，A=B)，隣接イベントの時間間隔 (MOMENTARY, TIME, DATE, STATE) の四種類の情報が付与されている．いずれも文脈のみで推定できないものは「UNKNOWN」のラベルを付与された．

3 実験手法

3.1 対象タスクに対する追加学習

BERT などの事前学習済み言語モデルは対象タスクとの間にドメインの不整合がある場合，ファインチューニングを行うだけでは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．

3.2 マルチタスク学習

マルチタスク学習は，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．関連するタスクを用い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 [4]．

本研究では，MT-DNN [5] 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．

3.3 潜在的正規分布によるイベントの時間関係の推定

イベントの時間を潜在的な確率分布として捉え，イベント間の時間関係を推定する．まず，文章全文の埋め込みおよび，その文章内の比較する二つのイベントの埋め込みを入力として，イベントの潜在的な時間を表す正規分布の平均 μ（時点）と分散 σ^2（時間幅）を出力とするモデルを用意する．モデルの出力から時間関係確率の対数をとって損失とし，学習させていく．時間関係確率に関しては，Allen の区間代数 [6] にならい，区間代数に定義される 13 の関係を DVD データセットの 5 つの時間順序ラベルに縮退して，ラベルごとに計算する．文章中の二つのイベント A,B が起こった時刻を，それぞれ確率変数 A,B で表し，正規分布 $A \sim N(\mu_1, \sigma_1^2)$，$B \sim N(\mu_2, \sigma_2^2)$ として推定することを考えると，

$$P(A = B) = \exp\left(-\beta(\mu_1 - \mu_2)^2\right) \tag{1}$$

このように計算できる．他のラベルに関しては，ここでは要旨のため割愛し，論文にて記す．

4 実験

4.1 対象タスクに対する追加学習の実験

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 1 に示す．評価指標としては Exact Match (EM) と F1 スコアを採用した．

表 1: 対象タスクに対する追加学習の実験結果

	EM[%]	F1[%]
standard fine-tuning	33.9 (41.0)	61.2 (65.3)
MLM	**36.5 (42.2)**	**65.9 (66.4)**

表 2: マルチタスク学習の実験結果

	時制 [%]	時間幅 [%]	時間順序 [%]
シングルタスク	55.55	32.32	19.73
ALL	57.30	**42.70**	31.18
DVD(ALL) 話し言葉 (ALL)	50.30	36.51	27.93
DVD(ALL) MCTACO(ALL)	**57.37**	41.25	**35.47**

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．() 内は 5 分割交差検証の結果を記載する．

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．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．

4.2 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，時間関係のラベルが付与されている DVD データセットと日本語話し言葉コーパスも追加で使用した．また，言語モデルは，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．

実験結果・考察 ここでは要旨のため，対象タスクである DVD データセットを中心とした $\text{BERT}_{\text{BASE}}$ を用いてのマルチタスク学習の結果のみ表 2 に示す．評価指標も F1 スコアのみを記す．他の組み合わせの結果と Accuracy による評価は論文に記す．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，使用するタスクによって効果にばらつきがあることがわかった．マルチタスク学習は，用いるタスクどうし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．ここでは要旨のため割愛し，論文に結果とともに記す．

4.3 潜在的正規分布によるイベントの時間関係の推定実験

実験設定 本研究では，DVD データセットの時間順序ラベルの中から「UNKNOWN」のラベルを除いたデータのみを使用して実験する．同じデータでシングルタスク学習させた結果をベースラインとして分類器をしようする場合と比較する．また，言語モデルは $\text{BERT}_{\text{BASE}}$ を使用する．

表 3: 潜在的正規分布によるイベントの時間関係の推定の実験結果

	Acc [%]	Pre [%]	Rec [%]
ベースライン	39.67	46.03	39.99
時間関係確率	**50.41**	**51.2**2	**50.27**

実験結果・考察 実験結果を表 3 に示す．評価指標としては Accuracy (Acc) と 適合率 (Pre) と再現率 (Rec) を採用した．

実験の結果，本手法による精度の向上を確認できた．不正解の結果でも実際に予測した分布を図示すると，正しい分布のように感じられる例もあった．どのような推定が行われたのか伺えるという点でも，ラベル分類に確率分布を使用するということは有用であると考えられる．実際に図示した分布に関しては，ここでは要旨のため割愛し，論文に記す．

5 おわりに

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て研究を進めており，まず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．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．さらに，イベントの時間を潜在的な確率分布として捉えてイベント間の時間関係を推定し，ラベル分類に確率分布を使用することへの有用性を確認した．

参考文献

[1] 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.

[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] 浅原正幸, 越智綾子, 鈴木彩香. 時間情報アノテーションデータ. 『言語による時間生成』論文集・報告集, 2024. *to appear*.

[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] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, Vol. 26, No. 11, p. 832–843, nov 1983.