潜在的正規分布によるイベントの時間関係の推定

船曳 日佳里 [1]　持橋 大地 [2]　浅原 正幸 [3]　小林 一郎 [1]
[1] お茶の水女子大学　[2] 統計数理研究所　[3] 国立国語研究所
{funabiki.hikari,koba}@is.ocha.ac.jp　daichi@ism.ac.jp　masayu-a@ninjal.ac.jp

概要

本研究では，自然言語で表現されるイベントにおける時間的常識の推定タスクの精度向上を目的として，イベントの時間を Allen の区間代数 [1] を参考に，潜在的な正規分布として捉えてイベント間の時間関係を確率的に推定する．実験の結果，ベースラインと比較して 10%以上の精度の向上を確認できた．また，正解に至らずとも正しい分布に近い分布を推定できたことも確認でき，これは時間関係認識において，確率分布を使用することの有用性を示している．

1　はじめに

自然言語処理タスクの多くで，イベントが持つ時間的常識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い．イベントに含まれる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，文脈的な知識を持っている必要がある．例えば，我々は「眠る」と「夢を見る」というイベントでは「眠る」の方が「夢を見る」よりも期間が長いことや，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．これまで，時間幅や時間の順序関係をモデル化したタイムラインの構築 [2] や潜在的なイベントに関する時間関係にも注目する [3] など，様々なアプローチが模索されている．

近年，BERT [4] 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 [5]，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 [6][7]．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．先行研究 [8] では単純なラベル分類タスクとしていたが，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る．本研究では，イベントの時間分布を相対的な対数時間における潜在的な正規分布で表現し，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す．

表 1: Allen の区間代数 [1] における時間順序の定義.

Allen の時区間関係	時間順序ラベル
A before B A meets B	A < B
A after B A met by B	B < A
A overlaps B A starts B A started by B	A ≤ B
A overlapped by B A finishes B A finished by B	B ≤ A
A equals B A during B A contains B	A = B

2　提案手法

2.1　時間関係の定義とその確率化

Allen の区間代数 [1] は，区間の重なりについての代数である．Pustejovsky ら [9] は，範囲代数で定義される 13 の関係をテキスト中のイベント間の時間順序関係を表現する TimeML を提案した．このうちいくつかの関係は言語によって識別して表出することが難しく，より単純化された時間順序ラベルが自然言語処理では用いられる．本研究も先行研究にならい、区間代数に定義される 13 の関係を表 1 のように 5 つの時間順序ラベルに縮退したデータセット (3.1 節) を用いる．

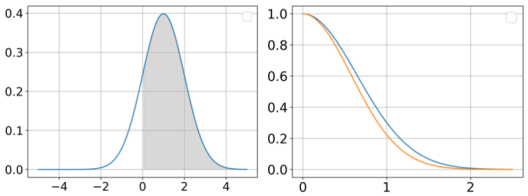

図 1: 式 (1) の積分が表す確率（左図影領域），および式 (3) の指数分布（右図青）と式 (4) の指数分布（右図オレンジ）．

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 で表すことにする．現在を原点 (0) とする相対的な対数時間軸において、正の方向を未来、負の方向を過去としよう．A, B の正確な値を文章から確定することは不可能なため，それぞれ、この軸上の正規分布 $A \sim N(\mu_1, \sigma_1^2)$，$B \sim N(\mu_2, \sigma_2^2)$ として推定することを考える．

このとき、表 1 の時間順序において $A > B$ となる確率は，$A-B$ は正規分布 $N(x|\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ に従うから，

$$P(A > B) = P(A - B > 0) = \int_0^\infty N(x|\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)dx \quad (1)$$

と表すことができる．式 (1) が表す部分を図 1 左に示した．この確率が大きいほど，二つの分布が離れていることになる (図 2 を参照)．この積分は，Python では `math.erf()` 関数で簡単に計算することができる．$B > A$ の場合も同様である．

また，A finishes B のように $A \geq B$ となる関係の場合，これは二つの分布の右端が揃っていることを意味している．そこで，分布の右裾を累積密度が (たとえば) 0.95 になる点と考え、この点を A について x_1, B について x_2 とおくと，$x_1 = x_2$ のときこの関係の確率は 1 で，x_1 と x_2 が離れるほど確率が下がると考えられる．よって，この確率を二乗距離 $(x_1 - x_2)^2$ についての指数分布

$$P(A \geq B) = \mathrm{Exp}((x_1 - x_2)^2; \beta) \quad (2)$$

で表現する．この様子を図 1 右に示した．正規分布の性質から、累積密度 0.95 の点は $x = \mu + 1.64\sigma$ となるから，式 (2) は

$$P(A \geq B) = \beta \exp\left(-\beta((\mu_1 + 1.64\sigma_1) - (\mu_2 + 1.64\sigma_2))^2\right) = \beta \exp\left(-\beta(\mu_1 - \mu_2 + 1.64(\sigma_1 - \sigma_2))^2\right) \quad (3)$$

で計算することができる．ここで β はどれくらい過敏に反応するかを表す係数で，本研究では実験の結果から β=1.2 を採用した．$B \geq A$ も同様である．

また，A=B の関係は発生時点 (の期待値) が近いことを意味するので，$(\mu_1 - \mu_2)^2$ についての指数分布を用いて，次のように表現できる．

$$P(A = B) = \beta \exp\left(-\beta(\mu_1 - \mu_2)^2\right) \quad (4)$$

本研究では，β=1.5 を採用している．こちらの β の方が大きいということは，図 1 からもわかるように，二つの μ の差が A=B の関係においてより重要であることを表している．

2.2 イベントの定義

自然言語におけるイベントをどのように定義するかも，結果に大きく関わっている．先行研究では，アノテーションを付与された動詞のみをイベントとして扱った．本研究では助動詞が時間順序の推定に重要であると考え，形態素解析を行って，アノテーションを付与された動詞とその後の助動詞までをイベントとして扱うこととした．

文末にアノテーションが付与された動詞がある場合，一文をイベントとして扱うことも有効であった．しかし，本研究でのデータセットは，二文以上で二つの文末の動詞の時間順序をアノテーションしている例が少なかったため，採用しなかった．また，動詞だけでなく事態性名詞もイベントの定義に関わると言われているが [10]，日本語では事態性名詞が必ずしも特定のイベントを指さないことが多く，動詞述語のみに時間の情報が付与されている．本研究でも，動詞述語のみをイベントとして扱う．

2.3 学習

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．これは，一つのイベントに対して全結合層と非線形活性化関数から構成されるニューラルネットワーク構造となっている．モデルの出力から

表 2: 時間順序ラベルの分布．

ラベル	データ数	計
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	

表 3: DVD データセットに含まれる文章の例.

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**	**時間間隔**
追わ	話す	A≤B	UNKNOWN

2.1 節で説明した時間関係確率の対数をとって損失とし，これに基づいたバックプロパゲーションを行って，モデルのパラメータを更新していく．

3 実験

3.1 使用データ

本研究で使用している DVD データセット [11] とは，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)，二つのイベントの時間順序 (表 1)，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 のラベルが付与されている．今回は時間順序の推定をタスクとし，1,508 個のデータのうち「UNKNOWN」のラベルを除いた 1,395 個のデータを訓練データ 1,274 個，評価データ 121 個に分割して使用した．時間順序ラベルの分布を表 2 に，例文を表 3 に示す．

3.2 実験設定

言語モデルは東北大学の乾・鈴木研究室が公開している日本語 BERT モデル `cl-tohoku/bert-base-japanese` を採用し，最適化には Adam [12] を使用した．学習の際のハイパーパラメータの設定を表 4 に示した．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．Accuracy は，予測が正しいサンプルの割合を示す指標である．適合率は，正事例と予測したもののなかで正解データが正事例の割合を表す指標である．再現率は，正解データが正事例のもののなかで正事例と予測した割合を表す指標である．

表 4: 学習の際のハイパーパラメータ.

batch size	learning rate	# epochs
32	5e-5	10

表 5: DVD データセットでの実験結果 (%)．* は他に 4 タスクのデータも使用した場合の参考値である．

	Acc [%]	Pre [%]	Rec [%]
ベースライン	39.67	46.03	39.99
マルチタスク学習 *	*52.07*	*60.49*	*51.82*
分布の重なり	42.98	63.50	43.06
時間関係確率	**50.41**	51.22	**50.27**

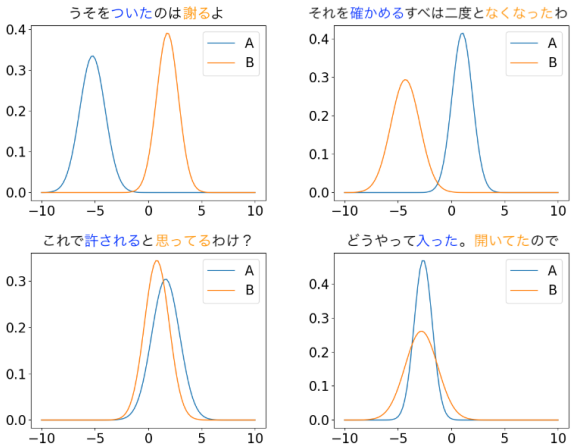

図 2: 時間関係を正しく予測した例（左上：A<B，右上：B<A，左下：B≤A，右下：A=B ）.

3.3 実験結果

実験の結果を表 5 に示す．比較対象とするベースラインは，マルチタスク学習のフレームワークである MT-DNN（Multi-Task Deep Neural Networks） [13] を使用して DVD データセットの時間順序タスクでシングルタスク学習をした結果である．また，先行研究 [8] で最も精度がよかった結果として，DVD データセットの全タスク（時制，時間幅，時間順序）と時間的常識に関する英語のデータセット MC-TACO [14] を日本語に翻訳したデータセット [15] の Duration（時間幅）と Frequency（頻度）の五つのタスクの組み合わせでマルチタスク学習をした結果を参考として記載する．さらに，モデルが出力した二つの正規分布の重なりが大きいと A = B というようにし，分布同士の重なった部分の面積を見てラベルを予測して学習させた結果も比較対象として記載した．また，正しく予測した例の一部を図 2 に示す．

結果を見ると，ベースラインや正規分布の重なりを利用した分類よりも Accuracy の向上が見られた．マルチタスク学習には少し及ばない結果であるが，

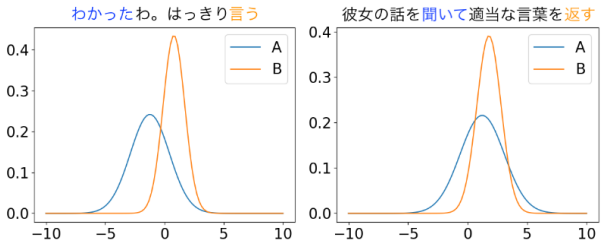

図 3: A≤B と正しく予測した例（左図）および，A=B と “誤って” 予測した例（右図）．

これは他タスクの情報を使用していないことから当然といえる．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．さらに，正解ラベルが B<A である文章が変更前と変わらず B≤A と予測していても，図示すると B<A に近づいた分布である例もあった．このように，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．また，適合率は下がってしまっているが，適合率と再現率のバランスが他の分類よりもよいことが確認できる．

4 分析

正解ラベル：A≤B に対する実際の正例と負例を図 3 に示す．左図は，「**わかった**（イベント A）わ。はっきり**言う**（イベント B）」から推定した二つの正規分布を図示した．これは，正しく A≤B と分類できた例で，分布も正しく推定できているように見受けられる．

右図は，「彼女の話を**聞いて**（イベント A）適当な言葉を**返す**（イベント B）」から推定した二つの正規分布を図示した．これは不正解で A = B と予測してしまった例である．しかし，分布を確認したところ，「話を聞く」よりも「適当な言葉を返す」の時間幅が顕著に短く推定されている点や時点の前後は合っているものの，同時に起こり得ることを捉えている点から，後者の方が人間が常識的に捉えるような正しい分布のように思われる．このように正しい時間分布から予測された時間関係とアノテーションされた時間関係が一致しない例も確認された．

また，実験の予測結果の詳細を表 6，表 7 に示す．横の行に正解ラベルのデータ数，縦の列に予測ラベルのデータ数を記載した．すなわち，太字が正解ラベルを予測した数となる．() 内はデータの個数を記載する．

この結果から，時間関係式を用いた学習の方は

表 6: 分類器を用いたベースライン実験結果の詳細．

	A<B (26)	B<A (20)	A≤B (10)	B≤A (2)	A=B (63)
A<B(24)	**13**	1	0	0	10
B<A(24)	4	**10**	2	0	8
A≤B(24)	5	4	**5**	1	9
B≤A(24)	2	3	1	**1**	17
A=B(25)	2	2	2	0	**19**

表 7: 時間関係式を用いた実験結果の詳細．

	A<B (16)	B<A (17)	A≤B (29)	B≤A (22)	A=B (37)
A<B(24)	**10**	0	6	0	8
B<A(24)	0	**8**	2	8	6
A≤B(24)	6	0	**16**	0	2
B≤A(24)	0	9	1	**10**	4
A=B(25)	0	0	4	4	**17**

A<B と A≤B のように近い時間関係のものはお互いのデータからも予測されやすく，A<B と B<A のように遠い時間関係のものは予測されにくいことが確認できる．また，データセットのラベル分布に偏りがあるにもかかわらず（表 2），ほぼ満遍なく予測をできていることも確認できる．分類器を用いたベースラインの実験結果を確認すると，ラベルごとの訓練データ数にかなり依存しているように見受けられる．不正解でも近い時間関係を予測しているのではなく，データ数が多いものを予測しているようで，その違いは顕著である．

5 おわりに

本研究では，自然言語で表現されるイベントにおける時間的常識の推定タスクにおいて，イベントの潜在的な時間を正規分布として捉えてイベント間の時間関係を推定する実験を行った．実験の結果，時間関係を式で表現して学習させることでより正しい時間分布を推定して，イベント間の時間関係推定の精度向上に有用であることを確認できた．

今後の課題として，正しい時間分布を正しい時間関係と分類できるような評価方法の作成があげられる．また，本研究では使用しなかった同データセットの時制と時間幅の情報を用いたよりリッチな学習を検討しているが，一つでも「UNKNOWN」のラベルが付与されたデータを除いてしまうと，データ数がかなり少なくなってしまう懸念がある．現在はデータ数が限られているため，よりデータ数の多いデータセットの使用も検討していきたい．

謝辞

本研究は JSPS 科研費 18H05521 の助成を受けて行ったものです.

参考文献

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

[2] Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2906–2919, Florence, Italy, July 2019. Association for Computational Linguistics.

[3] Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1361–1371, Online, June 2021. Association for Computational Linguistics.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 船曳日佳里, KANASHIRO Pereira Lis, 木村麻友子, 浅原正幸, 越智綾子, CHENG Fei, 小林一郎. マルチタスク学習を用いた時間を認識する汎用言語モデルの構築. 人工知能学会全国大会論文集, Vol. JSAI2023, pp. 3A1GS603–3A1GS603, 2023.

[9] James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. ISO-TimeML: An international standard for semantic annotation. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[10] Feng Pan, Rutu Mulkar, and Jerry R. Hobbs. Learning event durations from event descriptions. In Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle, editors, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 393–400, Sydney, Australia, July 2006. Association for Computational Linguistics.

[11] 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*.

[12] 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.

[13] 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.

[14] 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.

[15] 船曳日佳里, 木村麻友子, KANASHIRO Pereira Lis, 小林一郎. 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. 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 2B4GS604–2B4GS604, 2022.