

# 生成データに基づいた日本語の時間関係推定

小國怜美<sup>1</sup> 持橋大地<sup>2</sup> 小林一郎<sup>1</sup>

<sup>1</sup> お茶の水女子大学 <sup>2</sup> 統計数理研究所・国立国語研究所  
{g2120511,koba}@is.ocha.ac.jp daichi@ism.ac.jp

## 概要

時間関係認識は、自然言語理解において必要となる正確な文脈理解のための重要なタスクである。人が用いる常識的な知識を学習させることで、時間関係を捉えることが可能な言語モデルを構築する試みが行われている。本研究では、先行研究 [1] の手法に倣って、イベントの生起状態を正規分布で表現し、その位置の相対関係により時間関係識別を行った。その際に大規模言語モデルを用いて、Allen の区間代数 [2] の定義に従ってデータセットを新たに生成して学習を行ったところ、モデルの精度向上を確認できた。

## 1 はじめに

時間関係認識に焦点を当てた研究は多数存在する [3] [4]。日本語を対象とした時間的順序関係推定の研究は、英語や中国語などの話者の多い言語に比べて少なかったが、BCCWJ-TimeBank [5] が構築されたことで、機械学習による統計的解析や、定量的な性能評価が可能になった。日本語の接続表現に注目して、日本語文を AFTER, BEFORE, OVERLAP, VAGUE の 4 つの時間関係に識別する取り組み [6] や、SVM や決定木などの機械学習アルゴリズムを用いて、文法的、語彙的情報を持つ特徴量から時間関係ラベルを推定する取り組み [7]、日本語に比べて豊富に存在する英語のリソースを、日本語の時間関係識別に利用する取り組み [8] など、様々なアプローチが行われてきた。

中でも先行研究 [1] では、Allen の区間代数 [2] に従った時間関係ラベルを用いて、文章中のイベントの生起状態が正規分布に従うとみなして時間関係を識別する手法が提案された。その手法に従ってモデルを構築し、同様のデータセット [9] を用いて実験を行ったところ、学習に伴って損失は減少するも精度が上昇せず、モデルの精度は約 2 割であった。そこで、よりデータ数に偏りがなく、ラベル付けの根

表 1: Allen の区間代数 [2] における時間順序の定義

Allen の時区間関係	時間関係ラベル
A before B A meets B	$A < B$
A overlaps B A starts B A finished by B	$A \leq B$
A equals B A during B A contains B	$A = B$
A overlapped by B A finishes B A started by B	$A \geq B$
A after B A met by B	$A > B$

拠が正確なデータセットを用いて学習を行いたいと考え、大規模言語モデル GPT-4o に対して Allen の区間代数 [2] の定義を与え、時間関係ラベルごとに新たにデータセットを生成した。これを用いて、2 つのイベント間の時間的順序関係を正確に認識できるモデルの構築を目指す。

## 2 時間関係の識別

### 2.1 Allen の区間代数

Allen の区間代数 [2] は、時間区間の重なりについての代数である。自然言語処理では、単純化された時間関係ラベルが用いられることが多く、本研究も先行研究 [1] にならい、時間関係ラベルを表 1 のように 5 つにまとめて識別を行う。

### 2.2 データの作成

表 1 の時間順序の定義に従い、GPT-4o を用いて 2 つのイベントの時間関係が表現されている日本語文データセットを生成した。GPT のモデルは、4o mini と比較して 4o の方が正確にデータを生成できたためこちらを用いた。

### 2.2.1 パターン別生成

生成の際プロンプトでは、時間関係の定義を与え、そのラベルは Allen の区間代数 [2] を簡易にした 5 つではなく、定義されている 13 個を用いた。それぞれの場合について、考えられる文法的パターンを提示させた。与えたプロンプトと GPT-4o の回答は、Appendix に示す。10 から 20 のパターンが示されたため、各パターンごとに生成を繰り返した。

文法的パターンには、重複、曖昧なもの、ラベルが間違っただけのもので生成されることもあるため、一度すべてのラベルについて、考えられるパターンを生成したのち、どれをどのラベルのデータとするかを明確にしてから、データの生成を進めた。1 つのラベル内でも、言い回しのパターンにデータ数の偏りが生じないように、それぞれ 100 程度生成したいと考え、600 を文法的パターン数で割った数だけ含むようにした。“A before B”と“A meets B”のように、“すぐに”、“や否や”などの時間間隔を表す表現を用いることで、その違いが表出するものもある一方で、“A equals B”、“A during B”、“A contains B”のように、定義文が言葉として表現されず、13 ラベル間の違いが言語に表出しないものが多く存在した。

### 2.2.2 $A \geq B$ , $A > B$ である場合の生成

また、 $A \geq B$ ,  $A > B$  である場合も、同様にして GPT-4o に定義を与えたが、文中で述べられる順番と、実際の時間順序が逆である文を生成することができなかったため、定義に加えて、Few-shot 学習として具体例を与えて、同じくパターンごとにデータを生成した。具体例は、先行研究 [1] で使用されていたデータセット [9] から用いた。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ったため、違う人物のセリフの間のイベントの前後関係に注目しているものがほとんどで、時間順序が曖昧なものも多く、ランダムに選択すると学習を混乱させることが考えられた。それぞれのラベルの特徴が現れている文法的パターンが存在したため、それらを具体例として ChatGPT に与えた。

$A \geq B$  のラベルが付与されたデータは既存のデータセット中に少なく、このラベル特有のパターンとしては“A と B ている (B ている：継続する心理状態)”しか存在しなかったが、B にはさまざまな動詞が考えられ、そのバリエーションを増やすことでラベル付けの妥当性を保った状態で多くのデータを用

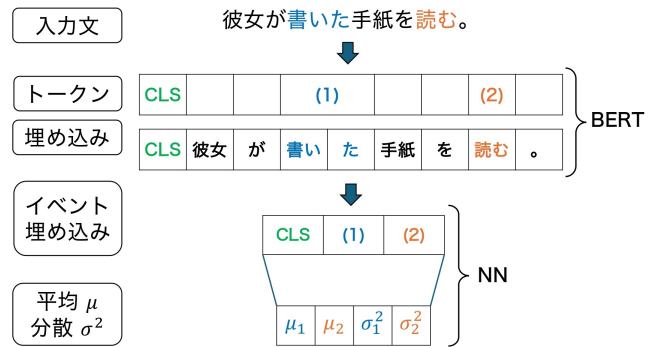


図 1: 時間関係識別モデルの概要。



図 2: イベントトークンの抽出。

意した。

### 2.2.3 データの妥当性の確認

表 2 に生成したデータの文法的パターンの一覧を示す。これらはいずれも、ChatGPT に対し Allen の区間代数 [2] のいずれの時間関係に該当するかを尋ね、正しい回答を得られている。

## 2.3 時間関係識別モデル

モデルの概要を図 1 に示す。日本語文を入力とし、自然言語処理ライブラリ GiNZA [10] を用いて形態素解析を行って、動詞と、それに続く助詞または助動詞までをイベントとした (図 2)。続いて文を自然言語処理モデル BERT [11] に入力して得られた 2 つのイベントトークン、および CLS トークンの埋め込みをモデルへの入力として、4 次元の値を出力する。モデルは、線形層とドロップアウト層で構成されている。出力 4 値はそれぞれ、2 つのイベントの生起状態が従う正規分布の平均と分散とする。ただし、分散は正であるため、exp 関数を通した値を分散として用いている。各ラベルが、理想とする分布にどれだけ近いかを表す確率  $p$  とすると、その負の対数  $-\log p$  を損失関数とし、逆伝播を行ってモデルのパラメータを更新する。

## 2.4 時間関係確率

損失関数に用いる時間関係確率は、先行研究 [1] に従って算出する。文章中の二つのイベント A, B

表 2: 生成データの文法的パターン・含まれる手がかり表現

時間関係ラベル	文法的パターン・含まれる手がかり表現
$A < B$	A てから/た後 B A たら/た結果/たことで/たせいで/たおかげで B A とすぐに/や否や B
$A \leq B$	A かけたところ に/で B A ついでに B A て B A つつ B 始める [A starts B] A ながら B 終える [A finished by B]
$A = B$	A ながら/つつ B A とともに/かたわら B A 間に B A ている [名詞] が/に/を B(現在)
$A \geq B$	A と感じている [B: 感情] A と期待して/信じている [B: 評価・判断] A と覚悟している [B: 態度・意図] A と想像して/予想して/願っている [B: 想像・推測] A と考えて/思っている [B: 認識] ※ A は現在形で, 未来の内容
$A > B$	A ように/ために B A 前に/まで/頃には B A と 言われた/決まった/知った/された A つもりなのが 分かる/見える A ための [名詞] が/に/を B

の生起確率  $A, B$  が, 現在を原点とする時間軸上でそれぞれ正規分布  $N^1(\mu_1, \sigma_1^2), N^1(\mu_2, \sigma_2^2)$  に従うと仮定し, 確率はそれぞれ以下の式で求める.

$$P^1 A > B^0 = P^1 A \quad B > 0^0 = \int_0^1 N^1(\mu_1, \sigma_1^2, \sigma_2^2) dx \quad (1)$$

$A < B$  の場合も同様にして考える.

$$P^1 A \geq B^0 = \exp^{-1} x_1 \quad x_2^{0^2}; \beta^0 = \exp^{-1} \beta^1 \mu_1 - \mu_2, 1.64^1 \sigma_1 - \sigma_2^{0020} \quad (2)$$

$x_1, x_2$  はそれぞれ  $A, B$  の累積密度 95%点を示す. 先行研究 [1] にならい,  $\beta = 1.2$  を採用した.  $A \leq B$  の場合も同様にして考える.

$$P^1 A = B^0 = \exp^{-1} \mu_1 - \mu_2^{0^2}; \beta^0 \quad (3)$$

先行研究 [1] にならい,  $\beta = 1.5$  を採用した.

### 3 実験

2つのイベントの関係が記述された自然言語文を入力とし, 表 1 で示した時間関係ラベルを推定する. ラベル推定時には, 推定した  $\mu_1, \mu_2, \sigma_1, \sigma_2$  を 5 つの確率関数にそれぞれ代入し, 正規化して, そのうち最も値の大きいものを推定ラベルとする. 生成データ, 教師データ, その 2 つを混ぜたデータの 3

つをそれぞれ用いてモデルを学習させ, 教師データを用いてその精度を算出する.

#### 3.1 データセット

先行研究で用いられたデータ [9] (教師データ), 2.2 節の通り生成したデータ (生成データ), 並びに 2 つのデータを 1:1 の割合で混ぜたデータを用いる (表 3). 教師データには様々なパターンの文が含まれるため, 教師データをなるべく多く含み, かつ生成データとのデータ数比が 1:1 になるように混ぜた.

#### 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 tohoku-nlp/bert-large-japanese<sup>1)</sup>を採用した. BERT の更新は 4 エポック目以降は停止し, また, ドロップアウト率をデフォ

表 3: 使用したデータセットのラベル分布

時間関係	データ数		
	生成データ	教師データ	生成, 教師データ
$A < B$	665	267	534
$A \leq B$	639	114	228
$A = B$	664	703	1328
$A \geq B$	621	44	88
$A > B$	660	238	476

表 4: 全体の精度とラベルごとの精度

	生成データ	教師データ	生成、教師データ
全体	20.00	21.41	23.87
A < B	8.614	0.13	0.3745
A ≤ B	50.88	28.74	29.82
A = B	23.61	32.24	18.78
A ≥ B	6.818	25.89	59.09
A > B	9.664	8.38	55.88

累積密度点とは、ラベル A ≥ B の確率関数に用いる累積密度点を指す。

ルトの 0.1 から 0.3 に変更している。これにより訓練データでの過学習を防止する。最適化には Adam [12] を使用した。また、バッチサイズは 32、学習率は  $5 \times 10^{-5}$ 、エポック数は 20 を用いた。評価指標としては Accuracy を採用した。また、各ラベルのデータ数の違いから学習に偏りが生じることを防ぐため、focal loss [13] を参考にし、算出された損失に重みづけをしたものを学習に用いた。重みは、全データ数を該当ラベルのデータ数で割ったものを正規化した値とした。また、教師データの場合は、データ数の分布の偏りを考慮して、データを 5 分割しクロスバリデーションを行って精度を算出した。

### 3.3 実験結果

モデル全体の精度とラベルごとの精度を、学習に用いたデータごとに表 4 に示す。生成データのみを用いてモデルを学習させた場合は、教師データで学習を行った場合とほぼ同じ精度となったものの、2 つを混ぜたデータを用いた場合は、モデルの精度向上が確認できた。ラベルごとの精度に注目すると、A = B 以外のラベルでは精度が向上している。図に、同一のデータに対して、教師データで学習したモデルでは正しい推定が行えなかったが、混ぜたデータで学習したモデルでは正しく推定できた例を、並べて示す。混ぜたデータを用いて学習したモデルを用いて推定できた正規分布の結果を、Appendix に示す。

### 3.4 考察

ラベルごとの精度に注目する。まず教師データにより学習したモデルの結果と混ぜたデータにより学習したモデルの結果を比較すると、 $A \geq B, A > B$  の精度の増加が大きいことがわかる。これは、この 2 つのラベルは、データを生成する際に、教師データから選択した具体例を用いて Few-shot 学習をさせて

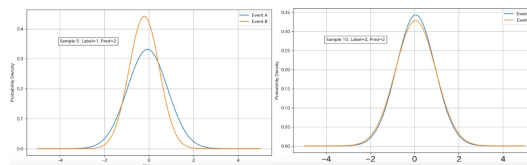


図 3: タネンだけが死ねば、警察はここを徹底的に調べる。

左: 教師データによる学習モデル

右: 混ぜたデータによる学習モデル

生成したことが一番の原因と考えられる。実際、具体例として用いたものや、同様の言い回しを含むものは正解している例が多かった。このことから、今回生成したデータセットは、モデルの汎用性の汎用性を少し上げることはできたが、教師データのバリエーションの多さ、ラベルの曖昧さに対応できていないことがわかる。

また、生成データは“雨が止んでから、野球試合が始まった。”のように 1 つの文中の 2 つのイベントに注目している一方、教師データは、“図書館でこのビデオを見つけたわ。これは誰だか分かる?”のように、複数の文にまたがっている 2 つのイベントに注目したものが多く含まれることが考えられる。これは、データ生成の段階で懸念していたものの、2 文に分かれ、そこに時間関係が現れるようなデータの生成が難しく、1 文中の 2 つの動詞に注目したデータとしたものの、今回の教師データに対応できるようにするにはさらに工夫が必要だったと考える。

## 4 おわりに

本研究では、先行研究において提案された、イベントの時間関係を、正規分布の位置関係によって表現し推定する手法について、大規模言語モデルを用いて、Allen の区間代数 [2] に従って生成したデータセットで学習させることで、その精度を向上させることを目指した。

今後の課題として、現在は、あくまで 2 つのイベントの相対的時間関係を推定しているに過ぎないが、時制を学習させることで、絶対的な発生時点を含めて推定するようになると考えている。さらに、動詞の持続時間を学習させることで、イベントの時間幅を含めた推定が可能になり、より人間の直感に合致する分布を推定できるようにしたい。さらに、学習ならびにテストに用いるデータセットについては、さらに検討を重ねていきたい。

1) <https://huggingface.co/tohoku-nlp/bert-large-japanese>

## 謝辞

本研究は JSPS 科研費 JP23K28143 の助成を受けたものです。

## 参考文献

- [1] 船曳日佳里, 持橋大地, 浅原正幸, 小林一郎. 潜在的正規分布によるイベントの時間関係の推定. 言語処理学会 第 30 回年次大会 発表論文集, pp. 1243–1247, 兵庫, 日本, 2024.
- [2] James F. Allen. Maintaining knowledge about temporal intervals. **Commun. ACM**, Vol. 26, No. 11, p. 832–843, November 1983.
- [3] 吉川克正, リーデルセバスチャン, 浅原正幸, 松本裕治. Markov logic を利用した時間的順序関係の同時推論. 人工知能学会論文誌, Vol. 24, No. 6, pp. 521–530, 2009.
- [4] Nathanael Chambers and Daniel Jurafsky. A probabilistic model of temporal ordering of events in natural language text. In **Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL)**, pp. 173–182, Manchester, UK, 2008. Association for Computational Linguistics.
- [5] 吉川克正, 浅原正幸, 飯田龍. Bccwj-timebank を対象とした時間的順序関係の推定. 言語処理学会 第 20 回年次大会 発表論文集, pp. 1103–1106, 北海道, 日本, 2014.
- [6] 高木宏伸, 嶋田和孝. 事象間の接続関係に基づく時間的順序関係推定. 言語処理学会 第 21 回年次大会 発表論文集, pp. 55–58, 京都, 日本, 2015.
- [7] 稲田和明, 松林優一郎, 乾健太郎. 同一文内の表現対を対象とした日本語における時間関係認識. 言語処理学会 第 20 回年次大会 発表論文集, pp. 348–351, 京都, 日本, 2014. 言語処理学会.
- [8] 吉川克正, 浅原正幸. 言語横断手法による日本語時間的順序関係推定. 言語処理学会 第 21 回年次大会 発表論文集, pp. 353–356, 京都, 日本, 2015.
- [9] 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.
- [10] 松田寛. Ginza - universal dependencies による実用的日本語解析. 自然言語処理, Vol. 27, No. 3, pp. 695–701, 2020.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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.
- [12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **CoRR**, Vol. abs/1412.6980, , 2014.
- [13] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming

He, and Piotr Dollár. Focal loss for dense object detection. **CoRR**, Vol. abs/1708.02002, , 2017.

