日本語生成コーパスにおける時間関係の分布推定

小國 怜美　(指導教員：小林 一郎)

1　はじめに

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．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う．その際に，よりラベル分布が均等で，ラベル付けの根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] の定義を与えて，新たに日本語時間関係識別データセットを作成し，そのデータセットを用いて時間関係の識別を行う．

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 4o を用いた．

パターン別生成　データ作成時のプロンプトでは，時間関係の定義を与え，そのラベルは Allen の区間代数 [2] を簡易にした表 1に示す「時間関係ラベル」の 5 つではなく，定義である 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．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（与えたプロンプトと GPT-4o の回答の一部を Appendix に示す）．

文法的パターンには，重複，曖昧なもの，ラベルが間違ったものが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（一覧を Appendix に示す）．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．

$\mathbf{A \geqq B}$, $\mathbf{A > B}$ である場合の生成　$A \geqq B$, $A > B$ 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 [3] 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ぼ全体を占め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を分析し，それらを具体例として GPT-4o に与えた．

$A \geqq B$ 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意した．

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	$A \leqq B$
A equals B A during B A contains B	$A = B$
A overlapped by B A finishes B A started by B	$A \geqq B$
A after B A met by B	$A > B$

データの妥当性の確認　各文法的パターンはいずれも，GPT-4o 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 [4] 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．続いて文を自然言語処理モデル BERT 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．

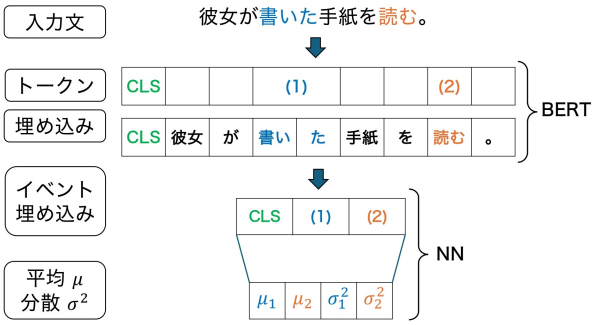


図 1: 時間関係識別モデルの概要

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] に従って以下のように算出する．文章中の二つのイベント A，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 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ に従うと仮定する．数式の対称性から，ここでは 3 つの関係にのみ言及する．

$$\begin{aligned} P(A > B) &= P(A - B > 0) \\ &= \int_0^\infty N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) dx \end{aligned} \tag{1}$$

$$\begin{aligned} P(A \geqq B) &= \exp\left((x_1 - x_2)^2; \beta\right) \\ &= \exp(-\beta(\mu_1 - \mu_2 + 1.64(\sigma_1 - \sigma_2))^2) \end{aligned} \tag{2}$$

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．

$$P(A = B) = \exp\left((\mu_1 - \mu_2)^2; \beta\right) \tag{3}$$

先行研究 [1] にならい，$\beta = 1.5$ を採用した．

3 実験

3.1 データセット

先行研究で用いられたデータ [3]（教師データとする），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．教師データには様々なパターンの文が含まれるため，教師データを可能な限り多く取り入れつつ，生成データとのデータ数比が 1:1 となるように調整する．

3.2 実験設定

言語モデルは日本語 BERT モデル tohoku-nlp/bert-large-japanese[1] を採用する．BERT の更新は 4 エポック目以降停止し，ドロップアウト率をデフォルトの 0.1 から 0.3 に変更して過学習を防止する．最適化には Adam [5] を使用し，バッチサイズは 32，学習率は 5×10^{-5}，エポック数は 20 とした．評価指標には Accuracy を採用する．また，各ラベルのデータ数の違いによる学習の偏りを防ぐため，Focal Loss [6] を参考にし，算出された損失に重みづけを行う．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する．また，教師データを学習に用いる場合は，データ数分布の偏りから，データを 5 分割しクロスバリデーションを実施する．

[1] https://huggingface.co/tohoku-nlp/bert-large-japanese

表 2: 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．

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 2に示す．生成データのみで学習した場合，教師データで学習した場合とほぼ同じ精度となったが，両者を組み合わせたデータで学習するとモデルの精度向上が確認できた．ラベルごとの精度では，A = B 以外のラベルで向上が見られる．（Appendix に教師データのみで学習したモデルが誤推定したが，生成データを含むデータで学習したモデルが正しく推定できた例を示す．）

3.4 考察

A ≧ B,A > B の精度向上は，教師データ中の具体例を用いた Few-shot 学習によるデータ生成が主な要因と考えている．一方，全体の精度に大きな向上が見られなかった理由としては，教師データにはイベントが 2 文に分かれて存在するものが多いのに対し，全てのラベルについてそのようなデータを生成することが難しく，生成データでは 2 つのイベントが 1 文内に収まっていることが挙げられる．イベントが 2 文に分かれたデータは時間関係が曖昧なものが多いため，新たなコーパスから教師データを作成することも今後の検討課題としたい．

4 まとめ

本研究では，先行研究において提案された，正規分布の位置関係によってイベントの時間関係を表現し推定する手法に対し，大規模言語モデルを用いて，Allen の区間代数 [2] に基づいて生成したデータセットでモデルを学習させることで，その精度向上を目指した．今後はデータセットの検討に加え，時制や動詞の持続時間の学習にも取り組み，イベントの絶対的な発生時点や時間幅を考慮した，人間の直感に合致する分布推定を目指したい．

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] 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] 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.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, Vol. abs/1412.6980, , 2014.

[6] 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.

付録 A　与えたプロンプトと得た回答 (A ＜ B の場合)

プロンプト：日本語の文に，2 つの動詞が登場するとする．先に登場するものを A，後に登場するものを B とする．A が起こってから B が起こるような場合は，具体的にはどのような文法的バリエーションが考えられますか？ “A before B”という関係です．A と B に時間的重なりはありません．B はすぐ起こるのではなく，A と B には時間的隔たりがあります．また，動詞の直前に [ENT] を配置してください．

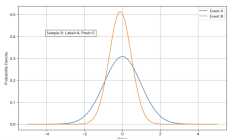
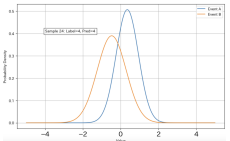
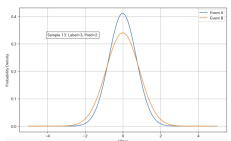
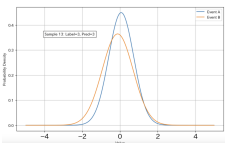
ChatGPT：1.A てから B
A が終わった後に B が始まることを表すシンプルな接続形式です．
例: ご飯を [ENT] 食べてから、宿題を [ENT] 始めた．
雨が [ENT] 降ってから、川が [ENT] 増水した．
（さらに 19 パターン提示）

付録 B　生成データの文法的パターン

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 ≦ 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 ≧ 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

付録 C　モデルによる推定結果の違い

以下に，教師データで学習させたモデルでは正しく推定できなかったが，生成データ + 教師データで学習させたモデルでは正しく推定できた例を示す．

	教師データ	生成 + 教師データ
陸自の答えは聞かなくても お分かりでしょう． (A>B)	 $\mu_1 = 0.00280, \sigma_1^2 = 0.943$ $\mu_2 = -0.106, \sigma_2^2 = 0.340$	 $\mu_1 = 0.351, \sigma_1^2 = 0.349$ $\mu_2 = -0.476, \sigma_2^2 = 0.590$
これで許されると 思ってるわけ？ (A≧B)	 $\mu_1 = -0.0123, \sigma_1^2 = 0.530$ $\mu_2 = -0.0083, \sigma_2^2 = 0.774$	 $\mu_1 = 0.0646, \sigma_1^2 = 0.444$ $\mu_2 = -0.125, \sigma_2^2 = 0.675$

表の上部には，モデルの学習に用いたデータセットを示す．