事例研究［論文］

# 育児Q&Aサイトにおける質問の時系列を考慮した複数の子供の月齢予測

東　将己，山根　大輝，原　朋史，梅津　大雅，馬嶋　海斗，松井　諒生，中田　和秀

## 1. はじめに

インターネットが普及し，さまざまなサービスがWebサイト上で提供される現代では，サイト運営会社がユーザーのプロフィールやサイト上での行動履歴を収集することが可能である．これらのデータから個々のユーザーの興味や行動傾向を予測することができれば，ターゲティング広告や販売促進サービスに用いることが可能になる．

コネヒト株式会社が運営するママの疑問や悩みを解決するQ&Aサイト『ママリ』においても，ユーザーの投稿した質問データなどが収集されている．ママリにおけるユーザーの興味は育児に関することであるため，子供の年齢に基づいたターゲティング広告を行うのが有益であると思われる．しかし，全体の約48%のユーザーが子供の誕生年月を登録していないため，年齢の予測が必要になる．

そこで本研究では，ユーザーが過去に投稿した複数の質問データからユーザーのすべての子供の年齢（月齢）を正確に予測することを目指す．予測に用いるママリの質問データは「次男が幼稚園に行きたくないと泣いています．どうしたらいいですか？」のようなテキストである．テキストは可変長で疎性の強い非構造化データであるため，年齢予測を行うには自然言語処理技術を用いて特徴ベクトルへの変換を行う必要がある．そこでわれわれは，長文をパラグラフごとに切り出して各パラグラフをBERT [1] へ入力し，Transformer [2] で複数の質問を紐付けた特徴を得るTransformer over BERT (ToBERT) [3] というモデルを採用し，各質問を複数のパラグラフと捉えてBERTへの入力とすることで，各ユーザーに何人の子供がいて，それぞれ何歳であるかを予測する手法を提案する．ただし，既存のToBERTで設計されているTransformerでは時系列の順序が考慮できるが，時間間隔を考慮できない．そこでPositional Encodingへの入力をユーザーの最初の質問からの経過時間とすることで時間間隔を考慮することを可能にした．また，ユーザーが二人以上の子供をもつケースを考えると，このタスクは複数の正解ラベルを予測するマルチラベル分類となる．しかし，一般的な機械学習で用いられるマルチラベル分類の手法は回帰の手法と異なり予測の正確性を評価することはできない．そのため分類問題で予測の正確性を考慮できるLabel Distribution Learning (LDL) [4] という手法をマルチラベル分類に拡張したMulti-Label Distribution Learning (M-LDL) を提案する．後述するママリのデータを用いた年齢予測の数値実験では，これらの工夫により，ユーザーのもつすべての子供の月齢を高精度で予測可能であることが確認できた．

本稿の構成について述べる．2節でママリのデータセットの特徴と解くべきタスクの説明を行い，3節でテキスト情報からデモグラフィック情報の予測を行った関連研究について述べ，4節で提案手法についての説明を行う．そして，5節でママリのデータセットを用いた数値実験と提案手法についての考察を行い，6節で本稿のまとめと今後について述べる．

## 2. データの特徴とタスクの説明

本研究では，経営科学系研究部会連合協議会主催，令和3年度データ解析コンペティションでコネヒト株式会社より提供された『ママリ』のユーザーデータと質問データを使用し，このデータから質問者の子供の月齢を予測する．ここで年齢ではなく月齢を用いる理由としては，子供の成長の早さが挙げられる．たとえば，生後0ヶ月ではベビーカーに乗っている赤ちゃんも，生後6ヶ月にもなると離乳食を開始する．このよ

ひがし　まさき，やまね　だいき，はら　ともふみ，うめつ　たいが，まじま　かいと，まつい　りょう，なかた　かずひで
東京工業大学工学院経営工学系
〒152–8552 東京都目黒区大岡山 2–12–1
受付 22.7.15　採択 22.11.18

うに子供の状態は短い期間で変化するため，ママリに投稿される質問の内容やユーザーの興味・関心はそれに合わせて変化する．そのため，正確なターゲティングを行うには年齢という範囲の広い単位ではなく，月齢という範囲の狭い単位での予測が好ましい．

ユーザーデータには子供の誕生年月が紐付けられており，質問データには各ユーザーが投稿したすべての質問の内容，カテゴリ，日時が含まれている．以下で使用データの特徴と解くべきタスクについて述べる．

- **子供の人数が未知**
  ユーザーには複数の子供がいる可能性を考慮した予測を行う必要がある．したがって，回帰ではこの問題を扱うことはできず，月齢ごとのマルチラベル分類を行う．しかし，回帰とは異なり，一般的に機械学習分野で用いられるマルチラベル分類の手法では出力が各ラベルに対して「合っているか」もしくは「間違っているか」でしか評価することができない．そのため，月齢（数値）の予測という回帰的な「正確性」を考慮したマルチラベル分類を行うことが求められる．
- **各質問が対象とする子供が不明**
  ママリは投稿の自由度が高く，兄弟喧嘩のように複数の子供を対象とした質問，複数の子供の中の一人を対象とした質問，個人的な悩みの相談や夫の愚痴など子供を対象としない質問などが混在している．そのため，質問単位で子供全員の月齢を予測するタスクを設計することは適切ではなく，ユーザーが過去に投稿した質問をまとめてユーザー単位で子供全員の月齢を予測するというタスクを解く必要がある．

以上より，本論文では複数の質問データをまとめて入力として，ユーザーのすべての子供の月齢をマルチラベルで予測するモデルを設計する．

## 3. 関連研究

本節では，テキストからの年齢予測に関する既存研究を紹介し，本研究との関係について述べる．

テキストには作成者のデモグラフィック情報に応じた内容的特徴と文体的特徴が現れるということがさまざまな研究から判明している [5, 6]．そのため，テキストから年齢や性別を予測する際には，その内容的特徴のみでなく文体的特徴も利用することが多い [7, 8]．

Rozen et al. [9] のニュース記事の閲覧データとそれに対するユーザーコメントに関する研究では，本研究と同様に多数の文書を同時に処理する手法を提案している．各ユーザーコメントを元の記事の文脈に埋め込むという手法を用いることで，ユーザーのデモグラフィック情報の予測を行っている．また，近年では Twitter などのオープンソースのテキストからの年齢予測が Klein et al. [10] を始めとして盛んに行われている．

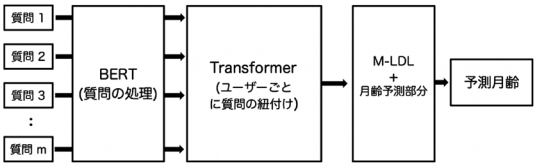


**図 1**　提案モデルの全体像

これらの既存研究の多くでは，テキストの内容と文体の両方を用いて年齢予測を行っている．一方で，本研究で用いるママリのデータでは，テキストの作成者（ママ）の文体から子供の月齢を予測することは難しいと考えられるため，テキストの内容のみから子供の月齢予測をしていく必要がある．また，予測対象である子供の人数が不明であるため，その推定を並行して行う必要がある．

## 4. 提案手法

本研究で提案するユーザーが過去に投稿した質問データからユーザーのすべての子供の月齢を予測するモデルについて説明する．

### 4.1 全体像

提案モデルは質問データの紐付け部分と月齢予測部分に分けられる．4.2 節で質問データの紐付け方法について，4.3 節で 4.2 節で抽出した特徴量を用いた月齢予測部分と月齢の順序関係を考慮した学習方法 (M-LDL) について説明し，4.4 節で 4.2 節，4.3 節で説明した提案モデルを用いてどのように質問文から子供の月齢を出力するか説明する．本研究の提案モデルの全体像を図 1 に示す．

2 節で説明したとおり，本研究の課題として複数の質問から月齢予測を行わなくてはいけない点，順序関係を伴ったマルチラベル分類となる点という 2 点があった．提案モデルでは質問データの紐付け方法を工夫することで 1 点目の課題に，月齢予測モデルにおいて正解ラベルの与え方を工夫することで 2 点目の課題に対処している．

### 4.2 質問データの紐付け

本研究では，複数の質問を入力としてユーザーの子供の月齢を予測するモデルとして，Transformer over BERT (ToBERT) [3] をベースとしたモデルを採用す

  

る．提案モデルによるユーザー $i$ の質問データの紐付けの流れは式 (1)〜(3) で表される

$$\boldsymbol{x}_j^i = \text{BERT}(question_j^i) \quad \forall j \in \{1, 2, \ldots, q_i\} \tag{1}$$

$$\{\hat{\boldsymbol{x}}_1^i, \ldots, \hat{\boldsymbol{x}}_{q_i}^i\} = \text{Transformer}(\{\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_{q_i}^i\}) \tag{2}$$

$$\hat{\boldsymbol{Y}}^i = \text{Mean}(\hat{\boldsymbol{x}}_1^i, \ldots, \hat{\boldsymbol{x}}_{q_i}^i) \tag{3}$$

なお, $question_j^i$ はユーザー $i$ の $j$ 番目の質問, BERT$(\cdot)$ は BERT [1] による特徴量の抽出，$\boldsymbol{x}_j^i$ は $question_j^i$ から抽出した特徴量，$q_i$ はユーザー $i$ の質問の総数，Transformer$(\cdot)$ は Transformer [2] の Encoder による複数の質問情報の紐付け，Mean はベクトルの要素単位で算術加算し，ベクトル数で割ることによる紐付けを表している．

提案モデルでは，まずユーザーの質問文をそれぞれ BERT に入力し，特徴量を抽出する（式 (1)）．BERT は，後述の Transformer の Encoder を基本構造とした高い精度で文章の意味を捉えることのできる自然言語処理モデルであり，近年の自然言語処理の研究における中心技術である．BERT の学習は事前学習とファインチューニングの二段階に分かれている．事前学習では大規模なテキストデータを用いて教師なし学習を行い，ファインチューニングでは解きたいタスクで事前学習済みモデルを再学習させ，モデルの微調整を行う．本研究では，オープンソースとして公開されている事前学習済みモデルを用いることで計算コストの高い事前学習は行わず，ママリのデータを用いた月齢のマルチラベル分類に対してファインチューニングを行う．ファインチューニングに関しては，5.1 節で後述する．なお，BERT には [CLS]，[SEP]，[UNK]，[PAD] と呼ばれるスペシャルトークンがある．[CLS] は文の先頭についており文全体の特徴を表している．[SEP] は文の末尾についており二つの文の比較を行う際，二つの文の特徴を表している．[UNK] は事前に準備されたトークンの辞書に存在しない単語に用いられる．そして [PAD] は長い文と短い文を入力する際，文の長さを長い文に合わせなければいけないため [PAD] で穴埋めすることで短い文を長い文と同じ長さにしている．本研究では，文章を BERT に入力した際の最終層の開始トークン [CLS] のベクトルを文章の特徴量としている．

次に，抽出された特徴量をすべて Transformer の Encoder に入力し複数の質問情報の紐付けを行い，要素単位で平均を取る（式 (2), (3)）．Transformer は，順序付きの入力を処理する深層学習モデルであり，自然言語処理，時系列，画像処理などの分野で広く活用されている．Transformer の Encoder は図 2 のように主に Positional Encoding，Multi-Head-Attention，Feed Forward の三つで構成され，Multi-Head-Attention と Feed Forward の後に変換前の特徴量と変換後の特徴量の和をとる skip-connection と特徴量の正規化を行っている．まず入力は Positional Encoding を通る．その後 Multi-Head-Attention を通り，skip-connection，正規化を行い，続いて Feed Forward を通り，skip-connection，正規化を行い出力されている．以降本稿では Transformer の Encoder のことを Transformer と記載する．なお本稿では，ページ数の関係で BERT や Transformer の厳密な説明を省かせていただいた．詳細については文献 [11, 12] などを参考にしていただきたい．

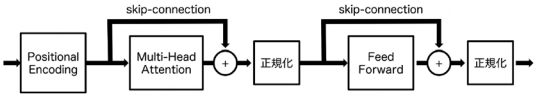


**図 2**　Transformer Encoder の構造

BERT のファインチューニングに対応している特徴量の紐付け方法として，Transformer のほかに，平均，多層パーセプトロン，LSTM [13] や GRU [14] といった RNN [15] の構造をベースとした RNN 型モデルが考えられる．平均や多層パーセプトロンでは，入力である質問の前後関係を考えることができず，RNN 型モデルは学習効率が悪く大規模データへの適用が難しい．このため，本研究では BERT による複数文書分類の先行研究である Pappagari et al. [3] に倣い，Transformer による複数の質問情報の紐付けを行った．なお，文献 [3] では BERT の最大入力長を超える長い文書の分類のために文書を分割してそれぞれ BERT に入力していたのに対し，本研究では同一ユーザーの複数の質問をそれぞれ BERT に入力していることに注意されたい．

また，提案モデルでは Transformer の Positional Encoding に改良を加えたモデルを使用している．Positional Encoding は Transformer の入力の位置情報を保存する構造であり，入力の順番を pos，入力の特徴ベクトルの要素番号（インデックス）を $2i$，$2i+1$，入出力する特徴ベクトルの次元数を $d$ としたとき，式 (4)，(5) のように表される．

$$\text{PE}(\text{pos}, 2i) = \sin\left(\text{pos}/10000^{2i/d}\right) \tag{4}$$

$$\text{PE}(\text{pos}, 2i+1) = \cos\left(\text{pos}/10000^{2i/d}\right) \tag{5}$$

先行研究 [3] における ToBERT の入力は長い単一の文書を分割したものであるため，pos を入力の順番とすることでそれぞれの入力の位置情報を考慮することができた．しかし，本研究ではそれぞれの入力が同一ユーザーの質問に対応しており，各質問は時間的に等間隔に配置されていないため，従来の Positional Encoding [2] では質問の時間間隔を考慮できていない．そこで，提案モデルでは Relative Temporal Encoding (RTE) [16] を参考に，Positional Encoding への入力を設計している．RTE では二つの時系列データ target と source が与えられ，それらの時刻をそれぞれ $T_t$，$T_s$ とおくと以下の式のように表される．

$$\mathrm{Base}(T_t, T_s, 2i) = \sin\left((T_t - T_s)/a^{2i/d}\right) \quad (6)$$

$$\mathrm{Base}(T_t, T_s, 2i+1) = \cos\left((T_t - T_s)/a^{2i/d}\right) \quad (7)$$

$$\mathbf{Base}(T_t, T_s) = \begin{pmatrix} \mathrm{Base}(T_t, T_s, 0) \\ \mathrm{Base}(T_t, T_s, 1) \\ \vdots \\ \mathrm{Base}(T_t, T_s, d-1) \end{pmatrix} \quad (8)$$

$$\mathrm{RTE}(T_t, T_s) = W_t \mathbf{Base}(T_t, T_s) \quad (9)$$

ここで $a = 10{,}000$ である．RTE では式 (6)〜(8) を用いて時間差を考慮した Positional Encoding を考え，式 (9) のように学習パラメータ $W_t$ で線形変換をし，タスクに応じた損失関数を用いて，損失関数の値を最小化するように $W_t$ の値を変えることで網羅していない時間差についても汎化的に性能を向上させている．ここで文献 [16] では 1900 年〜2019 年の 120 年といった長い期間の時間間隔を考えていたのに対し，本論文では高々 2 年 7 ヶ月という短い期間の時間間隔を考慮すればよい．そのためわれわれは RTE の Base の部分を参考にして Positional Encoding を強化することで，各質問が時間的に等間隔に配置されていない問題を解決した．ユーザーが $k$ 個の質問を行った際，質問文の投稿時期を $T_1, T_2, \ldots, T_k$ とおく．ここで $j$ 個目 $(1 \leq j \leq k)$ の質問文について以下のように Positional Encoding を設定する．なお，式 (4)，(5) と同様に入力の特徴ベクトルの要素番号（インデックス）を $2i$，$2i+1$，入出力する特徴ベクトルの次元数を $d$ とする．

$$\mathrm{PE}(T_j, 2i) = \sin\left((T_j - T_1)/10000^{2i/d}\right) \quad (10)$$

$$\mathrm{PE}(T_j, 2i+1) = \cos\left((T_j - T_1)/10000^{2i/d}\right) \quad (11)$$

つまり，Positional Encoding の pos をユーザーの最初の質問からの経過時間（月単位）とすることで，質問の時間間隔を考慮した予測を可能にした．

### 4.3 月齢の順序関係を考慮した学習方法

本研究が行う月齢予測では，ユーザーに複数の子供がいることを想定する必要がある．そのため任意の人数の出力に対応しなければならず，一般的な回帰のような単一の出力しか行えない手法では不十分である．よって本研究では，複数の出力を行えるマルチラベル分類の手法を採用する．

Transformer によって変換された特徴量 $\hat{\boldsymbol{Y}}^i$ は，式 (12) のように行列 $W \in \mathbb{R}^{C \times 768}$ で線形変換することで次元数を $d = 768$ から予測値の個数 $C$ に変換する．その後シグモイド関数に通すことで $z \in \{1, 2, \ldots, C\}$ について出力を $\hat{y}^i_z \in (0, 1)$ にする．

$$\hat{\boldsymbol{y}}^i = (\hat{y}^i_1, \ldots, \hat{y}^i_C)^T = \mathrm{sigmoid}(W\hat{Y}^i) \quad (12)$$

提案モデルでは，マルチラベル分類によって月齢予測を行うため，予測値の次元数は出力する月齢の候補数であり，それぞれのラベルには区間 $[0, 1]$ に含まれる値が出力される．

このとき，一般的なマルチラベル分類の手法では，ラベル間の順序関係を考慮できないという問題が存在する．たとえば，正解ラベルが 12 ヶ月のユーザーに対して予測を行う場合，12 ヶ月と正確に予測を行えなくとも 12 ヶ月に近い予測を行うことが望まれる．しかし，一般的な分類手法では正解ラベルと予測の順序関係を考慮しない．すなわち予測を 2 ヶ月としても 10 ヶ月としても誤差を同じものとして扱ってしまう．このような手法は，月齢のような順序関係が重要な値に対して不適切である．また，学習を安定させるために各ラベルに対して十分なデータ数があることが望まれるが，今回使用するデータは育児向けのサイトのものであるため，0 から 12 ヶ月歳付近のデータが極端に多く，それ以外のデータは非常に少ない．このような少量のデータでは正しく学習を行うことが困難である．

この二つの課題に対して，Geng et al. は年齢予測 [4]，頭部姿勢推定 [17] において，ラベルの連続性を利用した LDL を提案している．LDL は正解ラベルを平均とした確率分布を目的変数にすることでラベルの順序関係の評価を可能にする．また，これは正解ラベルに近いラベルを部分的に正解ラベルとすることになるため，各ラベルにおけるデータ数を擬似的に増やすことにもなる．具体的には，順序付きラベル集合を $L = \{l_1, l_2, \ldots, l_C\}$，$\sigma$ を標準偏差とすると，正解ラベルが $l_i$ のときのラベル分布 $\boldsymbol{y}^i = (y^i_1, y^i_2, \ldots, y^i_C)^T$ は式 (14) のようになる．

  

$$p\left(l_j \mid l_i, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(l_j - l_i\right)^2}{2\sigma^2}\right) \quad (13)$$

$$y_j^i = \frac{p\left(l_j \mid l_i, \sigma\right)}{\sum_{l_k \in L} p\left(l_k \mid l_i, \sigma\right)} \quad (14)$$

このように LDL は正解ラベルとの近さに応じて，学習に用いるラベルの生起確率を変更することで，ラベル間の順序を考慮することができる．

本研究の月齢予測タスクの場合，子供が複数存在する可能性があるため，単一のラベルのみにしか適応できない従来の LDL では不十分である．さらに単に複数のラベルを元にした確率分布を使用するだけでは複数の子供をもつユーザーの確率が低く評価されてしまう問題がある．たとえば，1 歳の子供 1 人をもつユーザーと 1 歳と 5 歳の子供 2 人をもつユーザーの確率分布を考えた場合，それぞれの確率の和が 1 にならなければならないため，後者のユーザーの 1 歳の確率は，前者のユーザーの半分程度になる．これでは子供の人数によって，分布の大きさが変化してしまうため学習が安定しない．よって，われわれはこの問題に対処するために M-LDL を提案する．

M-LDL では，複数の正解ラベルに対して，それらを平均とした分布の最大値を各ラベルの確信度の分布とする．具体的には，月齢の集合を $L = \{l_1, l_2, \ldots, l_C\}$ とすると，ユーザーがもち得る子供の月齢の組合せ集合は $L^{'} = 2^L \backslash \emptyset$ となる．ここで任意のユーザー $i$ の子供らの月齢の組合せを $l_i^{'} \in L^{'}$ としたとき，ユーザー $i$ についての月齢 $l_j \in L$ の確信度 $y_j^i$ は式 (16) で表される．

$$p\left(l_j \mid \hat{l}, \sigma\right) = \exp\left(-\frac{\left(l_j - \hat{l}\right)^2}{2\sigma^2}\right) \quad (15)$$

$$y_j^i = \max_{\hat{l} \in l_i^{'}} \{p(l_j \mid \hat{l}, \sigma)\} \quad (16)$$

たとえば，ユーザー $i$ の子供の組み合わせが $l_i^{'} = \{5, 16\}$ のとき，$l_j = 8$ の確信度は $y_j^i = \max\{p(l_j = 8 \mid \hat{l} = 5, \sigma),\ p(l_j = 8 \mid \hat{l} = 16, \sigma)\}$ となる．図 3 は横軸として $l_j$，縦軸として $y_j^i$ を取ったものである．この図の場合は，$y_j^i$ として $p(l_j = 8 \mid \hat{l} = 5, \sigma)$ の方が採用される．また，このユーザーの M-LDL による各ラベルの確信度の分布は図 3 の実線部になる．

この各ラベルの確信度の分布を用いて式 (1) の BERT と式 (2) の Transformer 内にある学習パラメータと，式 (12) 内の学習パラメータ $W$ について学習を行う．その際の損失関数として，式 (17) で表される Binary Cross Entropy を用いる．

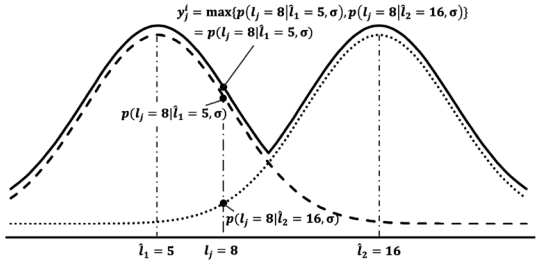


**図 3** 子供の組み合わせが $l_i^{'} = \{5, 16\}$ の場合の M-LDL を適応させた各ラベルの確信度の分布

$$H(\boldsymbol{y}^i, \hat{\boldsymbol{y}}^i) = -\sum_{j \in \{1,\ldots,C\}} \Big(y_j^i \log \hat{y}_j^i + (1 - y_j^i) \log\left(1 - \hat{y}_j^i\right)\Big) \quad (17)$$

この関数で計算された値は正解ラベルとの近さを表現できるため，ラベル間の順序を考慮した学習が可能となる．さらに和が 1 であることを必要としないため，ユーザーの子供の人数にかかわらず同じ月齢のラベルから発生する分布の大きさは等しくなる．これにより安定した学習を行うことができる．なお，$L^{'}$ はすべてのラベルの組で構成されているが，実験では訓練用データに含まれる正解ラベルの組のみ計算を行う．組み合わせの数は 12,523 種類であった．

ここまでの提案モデルの学習の流れを疑似コードを用いて説明する．まず，訓練データに含まれるユーザー $i$ に対して，Algorithm 1 の 1 行目を用いて予測値 $\hat{\boldsymbol{y}}^i$ を得る．次に，Algorithm 1 の 2 行目から 9 行目を用いて $i$ に対する損失 $H(\boldsymbol{y}^i, \hat{\boldsymbol{y}}^i)$ を計算する．そして，ここで計算した損失を基に提案モデル内で使用されるパラメータを更新する．

**Algorithm 1** ユーザ $i$ への予測に対する損失の計算

**Input:** $\boldsymbol{question}^i$：ユーザ $i$ が過去にした複数の質問
**Output:** $H(\boldsymbol{y}^i, \hat{\boldsymbol{y}}^i)$：ロス値
1: $\hat{\boldsymbol{y}}^i \leftarrow \boldsymbol{question}^i$を式 (1), (2), (3), (12) に順に入力
2: $loss_i \leftarrow 0$
3: **for** $j$ in $\{1, 2, ..., C\}$ **do**
4: $\quad y_j^i \leftarrow l_i^{'}$と式 (15), (16) を用いて計算
5: **end for**
6: **for** $j$ in $\{1, 2, ..., C\}$ **do**
7: $\quad loss_i \leftarrow loss_i - y_j^i \log \hat{y}_j^i - (1 - y_j^i) \log\left(1 - \hat{y}_j^i\right)$
8: **end for**
9: $H(\boldsymbol{y}^i, \hat{\boldsymbol{y}}^i) \leftarrow loss_i$
10: **return** $H(\boldsymbol{y}^i, \hat{\boldsymbol{y}}^i)$

### 4.4 現実を考慮した月齢出力方法

訓練データを用いたモデル内のパラメータの学習が終了した後は，モデル内のパラメータの値を固定することで未知のデータに対して推論を行うことが可能となる．一般的なマルチラベル分類では，モデルの予測値 $\hat{\boldsymbol{y}}^i$ に対して閾値以上のラベルをすべて解とするが，本研究ではこの方法の適用は精度悪化につながる可能性が高い．なぜなら，提案モデルでは近い月齢の予測値は似た数値になりやすいのに対し，正解ラベルの間隔は空いている場合が多いためである．たとえば，本研究において 10 ヶ月の子供がいるユーザーに対して 9 ヶ月，10 ヶ月，11 ヶ月を示すラベルの予測値を高く出すモデルは性能が良いといえる．その一方で，9 ヶ月，10 ヶ月，11 ヶ月の予測値が閾値以上であった場合，一般的なマルチラベル分類の方法では 9 ヶ月，10 ヶ月，11 ヶ月の 3 人の子供がいると出力してしまうが，この解は非現実的である．そこで，推論時には Algorithm 2 を用いて，提案モデルの予測値 $\hat{\boldsymbol{y}}^i = (\hat{y}^i_1, \ldots, \hat{y}^i_C)$ に対して現実的な制約を考慮して予測月齢を決定する．具体的には，1 行目において提案モデルを用いてユーザー $i$ に対する予測値 $\hat{\boldsymbol{y}}^i$ を得た後，3 行目で予測値「$1, \ldots, C$」に含まれる要素のうち，確信度が隣接する月齢の確信度以上かつ，閾値 $\theta$ 以上のものを暫定解とし，4–8 行目で暫定解の間隔が 9 ヶ月以内である場合は予測値が低い暫定解を除いて月齢の最終的な予測 $A$ を出力する．なお，閾値 $\theta$ の決定方法については 5.3.1 節で後述する．

**Algorithm 2** 推論時における現実を考慮した月齢予測

**Input:** $\boldsymbol{question}^i$：ユーザ $i$ が過去にした複数の質問
**Output:** $A := \{\hat{a}^i_1, \ldots, \hat{a}^i_{\hat{n}_i}\}$：予測月齢

1: $\hat{\boldsymbol{y}}^i \leftarrow \boldsymbol{question}^i$ を式 (1), (2), (3), (12) に順に入力
2: $A \leftarrow \emptyset$：空集合
3: $R \leftarrow \{r \in \{1, \ldots, C\} | \hat{y}^i_r \geq \theta, \hat{y}^i_r \geq \hat{y}^i_{\max\{r-1,1\}}, \hat{y}^i_r \geq \hat{y}^i_{\min\{r+1,C\}}\}$
4: **for** $r$ in $R$ **do**
5: $\quad R_2 \leftarrow \{r_2 \in R | -9 \leq r - r_2 \leq 9\}$
6: $\quad r_{\max} \leftarrow \underset{r_2 \in R_2}{\operatorname{argmax}} \hat{y}^i_{r_2}$
7: $\quad A \leftarrow A \cup r_{\max}$
8: **end for**
9: **return** $A$

また，Alogrithm 2 からもわかるように提案手法では間隔が 9 ヶ月より大きくなるように予測月齢を出力している．9 ヶ月という間隔は妊娠期間を考慮した数値であり，9 ヶ月以内に二度以上子供が生まれることは稀であるという知見に基づいている．なお，子供が双子の場合は月齢に差がないが，ターゲティング広告を目的とした月齢予測タスクにおいて双子を 2 人の子供と区別する必要性は低いため，本研究では双子は一人の子供とみなして処理している．

## 5. 数値実験

令和 3 年度データ解析コンペティションにおいてコネヒト株式会社から提供された，乳幼児をもつ親向けポータルサイト『ママリ』の質問データを用いて，提案手法の有効性を確かめる．

### 5.1 データとモデルの設定

本研究の数値実験では，子供の誕生日を登録しているユーザー（全体の約 48%）について，カテゴリが「妊娠・出産」「子育て・グッズ」である質問から子供の月齢を予測する．また GPU メモリの上限が 64GB であるため，ユーザーの質問をすべて学習に用いることができない．そのため 16 個以上の質問履歴があるユーザーは最新の 15 個の質問を使用する．そして質問履歴が 15 個未満のユーザーについては BERT 内の特殊文字である [PAD] で穴埋めしたダミー質問で 15 個になるように補正を加えた．質問数についての統計データを表 1，新しいものから 15 個抽出した各ユーザーの質問のうち，一番新しいものと一番古いものの期間に関する統計データを表 2 に記載する．

2019 年 1 月 1 日から 2021 年 7 月 31 日までの質問のテキストデータ 2,351,000 件，質問を行ったユーザー 159,265 人の子供の誕生日データを利用し，訓練用データ，検証用データ，テストデータが 6:2:2 になるように分割した．訓練データはファインチューニングに使用し，検証データは学習時のエポック数の決定と月齢抽出のための閾値の決定に用い，テストデータは評価に使用している．

実験タスクとして，ユーザーが最後に質問した時点で

**表 1** 質問数の統計的データ

| 平均 | 最大値 | 最小値 | 15 個以下の割合 |
|---|---|---|---|
| 14.8 個 | 3301 個 | 1 個 | 81.5% |

**表 2** 質問期間の統計的データ

| 平均 | 最大値 | 最小値 | 中央値 |
|---|---|---|---|
| 5.78 ヶ月 | 30 ヶ月 | 0 ヶ月 | 3 ヶ月 |

  

のユーザーのすべての子供の月齢について月単位でのマルチラベル分類を行った．なお，ラベルは出産前の子供をマイナスの月齢として扱ったうえで，−9 ヶ月から 72 ヶ月までを用い，ラベルの個数は全部で $C = 82$ となっている．（72 ヶ月以上のデータは 72 ヶ月とした）．

また，提案手法の一部である BERT の事前学習済みモデルには，東北大学乾研究室の公開している cl-tohoku/bert-base-japanese-v2 [18] を使用する．ここで BERT は $L = 12$, $H = 768$, $A = 12$ の BERT base モデルを用い，BERT の後ろにある Transformer については隠れ層の次元数を 768，Multi-head-attention の head 数を 8 個に設定している．

またファインチューニングは 4.2 節で説明したモデルに損失関数として式 (17) で表される Binary Cross Entropy を用い，M-LDL のラベルを用いて予測値 $\{\hat{y}_1^i, \ldots, \hat{y}_C^i\}$ を推定するタスクに対して行っている．予測値から子供の予測月齢と予測人数の導出は，4.4 節に記載した Alogrithm 2 で求めているため，直接的に子供の月齢と人数の推定タスクをファインチューニングしているわけではない点に注意されたい．

### 5.2 評価指標

複数の子供の月齢予測タスクでは，子供の人数についての予測精度，各子供の月齢についての予測精度の両方が求められる．本研究では，以下の二つの指標によりモデルの性能を評価する．

一つ目の指標は，モデルが予測する子供の人数と実際の子供の人数の平均絶対誤差（人数 MAE）である．テストデータのユーザー数を $N$，$i$ 番目のユーザーの子供の数の正解を $n_i$ 人，予測として出力した子供の数を $\hat{n}_i$ 人とすると，人数 MAE は式 (18) によって算出できる．

$$\text{人数 MAE} = \frac{1}{N}\sum_{i=1}^{N} |n_i - \hat{n}_i| \tag{18}$$

二つ目の指標は，モデルが予測する子供を基準とした最も近い正解の月齢との平均絶対誤差である．$i$ 番目のユーザーの子供の月齢の正解を $\{a_1^i, \ldots, a_{n_i}^i\}$，予測として出力した子供の月齢を $\{\hat{a}_1^i, \ldots, \hat{a}_{\hat{n}_i}^i\}$ とすると，月齢 MAE は式 (19) によって算出できる．

$$\text{月齢 MAE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{n}_i}\sum_{j=1}^{\hat{n}_i}\min_{k\in\{1,\ldots,n_i\}}\left|a_k^i - \hat{a}_j^i\right| \tag{19}$$

評価指標として，正解データの子供を基準とする指標も考えられるが，本研究の目的であるターゲティング広告への活用を考えた際，予測情報に基づいて出稿した広告が実際の子供の月齢にどの程度適したものであるかが重要であるため，モデルの予測を基準とする指標としている．

### 5.3 実験結果

本研究における提案は大別して

1. BERT を用いた特徴量抽出
2. Transformer による特徴量の紐付け
3. 時間間隔を考慮した Positional Encoding
4. マルチラベルに拡張した LDL (M-LDL)

の四つである．ここではこれらの効果を検証する．

#### 5.3.1 提案 2 から 4 の効果

Transformer による特徴量の紐付け（提案 2），時間間隔を考慮した Positional Encoding（提案 3），マルチラベルに拡張した LDL（提案 4）の効果の検証のために，それぞれの提案の有無を変更して実験を行った．Transformer による特徴量を紐付けを行わない場合，代わりに質問の特徴量の平均を取ることで特徴量の紐付けを行っている．

本実験の結果を表 3 に示す．なお，時間間隔を考慮した Positional Encoding（提案 3）は Transformer による特徴量の紐付け（提案 2）を前提とした提案であるため，前者のみを適用した実験は行っていないことに注意されたい．また，予測月齢決定の際の閾値 $\theta$ については検証用データの人数 MAE が最も低くなる値を採用し，提案 4 の M-LDL のハイパーパラメータとして $\sigma = 2$ を採用している．提案 2 から 4 それぞれの適用によって人数 MAE と月齢 MAE は改善する傾向があり，すべてを適用した場合が最も性能が良いことが見て取れる．以上の結果から，提案 2 から 4 は月齢予測に有効であるといえる．

#### 5.3.2 提案 1 の効果

BERT を用いた特徴量抽出の効果を検証するため，Word2Vec [19] により特徴量を抽出した場合との比較実験を行う．Word2Vec については，質問文の単語のベクトルの平均を特徴量としており，日本語 Wikipedia

**表 3** 提案 2 から 4 の効果の検証

| 提案 2 | 提案 3 | 提案 4 | 人数 MAE | 月齢 MAE |
|---|---|---|---|---|
| | | | 0.417 | 7.19 |
| ○ | | | 0.312 | 5.08 |
| | | ○ | 0.400 | 5.80 |
| ○ | ○ | | 0.312 | 4.76 |
| ○ | | ○ | 0.259 | 4.36 |
| ○ | ○ | ○ | **0.242** | **3.86** |

で学習を行ったモデル [20] と『ママリ』の質問データで学習を行ったモデルの 2 つを比較対象とした．また，前項の結果を踏まえて，特徴量抽出以外の部分では提案 2 から 4 すべてを適用したモデルを使用する．

本実験の結果を表 4 に示す．人数 MAE，月齢 MAE ともに BERT，Word2Vec（ママリ），Word2Vec (Wikipedia) の順に数値が良いことが見て取れる．ママリのデータによって学習することで日本語 Wikipedia で学習するよりも月齢予測の精度が良くなる理由としては，育児サイトの質問文固有の表現や文法を的確に捉えることができたからであると考えられる．BERT が Word2Vec よりも精度が高い理由としては，BERT による質問文からの特徴量の抽出も含めて月齢予測タスクで学習できるため，質問文から月齢予測に必要な部分をよく反映した特徴量を抽出できたからであると考えられる．

#### 5.3.3 M-LDL の標準偏差変化による考察

次に，M-LDL のハイパーパラメーターである標準偏差 $\sigma$ の影響を調べる．提案 1 から 4 をすべて採用したモデルのハイパーパラメーターを $\sigma = 1, 2, 4, 8$ の 4 通りに変化させ，結果を比較した．

本実験の結果を表 5 に示す．人数 MAE については $\sigma = 2$ が最も良いことがわかる．月齢 MAE は $\sigma$ を大きくすることで精度が良くなっているが，月齢予測数も減少しており，予測が難しそうな月齢は予測しなくなったと思われる（$\sigma = 2$ に対し $\sigma = 8$ の予測数は約 1%減）．そのため人数 MAE を確認しながら適切な標準偏差を選ぶ必要がある．

### 5.4 提案モデルの実用性の考察

最後に，提案 1 から 4 すべてを適用し $\sigma = 2$ としたモデル（以後提案モデルと呼ぶ）の実用性について考察する．

提案モデルの予測人数と正解人数の関係を図 4，予測月齢と正解月齢の関係を図 5 で示す．グラフの横軸が提案モデルの予測，縦軸が正解に対応している．なお，5.2 節の指標の設計と同様の理由から，図 4，図 5 は予測人数/月齢を基準としたグラフとなっている．具体的には，図 4 の数字は予測人数基準でのそれぞれの正解人数の割合を示しており，縦の列の和は 1 となる．そして，図 5 は提案モデルによって出力されたすべての月齢に対する，最も近い正解月齢が散布図としてプロットされている．図 4，図 5 ともに原点を通る傾き 1 の直線の近くに集中して分布していることから，子供の人数や月齢を高い精度で予測できているといえる．

次に，提案モデルが予測した子供の月齢の誤差を表 6 で示す．

**表 4**　言語モデルによる精度の差異

| | 人数 MAE | 月齢 MAE |
|---|---|---|
| Word2Vec (Wikipedia) | 0.416 | 7.26 |
| Word2Vec（ママリ） | 0.334 | 5.13 |
| BERT | **0.242** | **3.86** |

**表 5**　標準偏差による精度の差異

| 標準偏差：$\sigma$ | 人数 MAE | 月齢 MAE |
|---|---|---|
| 1 | 0.266 | 4.48 |
| 2 | **0.242** | 3.86 |
| 4 | 0.243 | 3.72 |
| 8 | 0.253 | **3.68** |

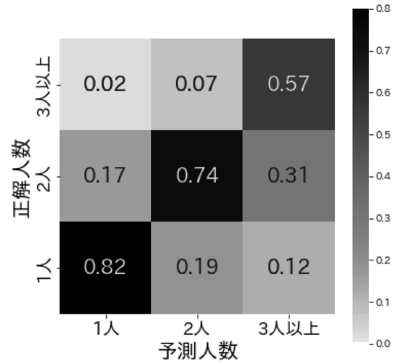


**図 4**　提案モデルの予測人数と正解人数の関係

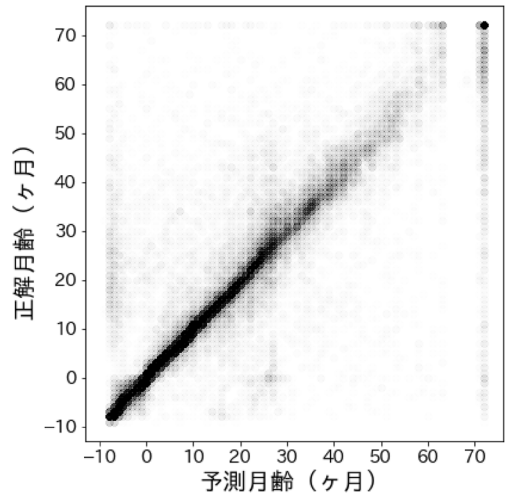


**図 5**　提案モデルの予測月齢と正解月齢の関係

  

**表 6**　提案モデルの子供の月齢の誤差

| | 一致 | 1 ヶ月以内 | 2 ヶ月以内 | 6 ヶ月以内 |
|---|---|---|---|---|
| 全体 | 39.2% | 73.1% | 81.9% | 91.8% |
| 1 歳未満 | 48.7% | 84.8% | 91.9% | 97.6% |

なお，子供の月齢の誤差はモデルの出力した月齢と正解の月齢の組合せのうち，最も近いものを採用している．また，全体はすべての子供を対象とした月齢の誤差を，1 歳未満は正解が 1 歳未満の子供に対象を限定した場合の月齢の誤差を示している．

表 6 の全体の結果から，予測月齢の誤差が 2 ヶ月以内のユーザーが 8 割を超えており，高精度な予測であることが確認できる．また，1 歳未満の乳児に関しては 1 ヶ月ごとの成長が大きいためより正確なターゲティングが求められるが，1 歳未満では月齢の一致が 48.7%，誤差 1 ヶ月以内が 84.8%と，より高精度な予測が可能になっている．これらの結果から，提案モデルの性能はターゲティング広告への適用を考えた場合，実用可能な水準であると考えられる．

また，提案モデルは −9 ヶ月から 72 ヶ月までの各 82 ラベルに対して 0 から 1 の値を出力する．これは出力が 1 に近いほどそのラベルの年齢の子供をもつ可能性が高いことを示すため，出力を予測の信頼度とみなすことも可能である．すると信頼度の高さに応じて柔軟なターゲティング広告を行うことも考えられる．たとえば，予測の信頼度が高いユーザーにはランドセルなどの購入期間が限られている商品の広告を，予測の信頼度の低いユーザーには絵本などの対象年齢の広い商品の広告を出すなどの活用方法がある．

## 6. おわりに

本研究では，育児ママ向け Q&A サイト『ママリ』の質問データを用いて子供の月齢の予測を行った．提案した子供の月齢予測モデルは以下の二つの工夫を行っている．一つ目は BERT と Transformer を用いた予測モデルを採用し，順序ではなく時間を埋め込む Positional Encoding を導入したことである．これにより，複数の質問データに対し質問の時間間隔を考慮した予測が可能になった．二つ目は M-LDL である．これによって順序関係を考慮した高精度なマルチラベル分類を行うことが可能になった．

なお，今回は情報量は多いがデータ数が少ないという性質をもつ質問データのみを扱ったため，質問をしたことがあるユーザしか予測できない．そのため，情報量は限定的だがデータ数が多いという性質をもつ検索データも考慮することでより多くのユーザに適用可能な月齢予測モデルを開発していきたい．また，今回用いた ToBERT は説明可能性が乏しく，入力となる質問文のどの単語が予測結果に寄与するか定量化することができない．提案モデルの ToBERT の部分を説明可能性の高いモデルで置き換えることで，予測結果が解釈可能になり，より実用性の高い月齢予測モデルを構築できる可能性がある．最後に，本提案では育児ママ向けの Q&A サイトを適用対象としたが，これ以外のサイトにも応用可能かどうかを今後検討していきたい．

**謝辞**　乳幼児をもつ親向けポータルサイト『ママリ』のデータを提供いただきました，コネヒト株式会社およびデータ解析コンペティション運営の皆様に御礼申し上げます．また，精密な査読および大変有意義なコメントをくださりました匿名の審査員のお二人に御礼を申し上げます．

### 参考文献

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv*, arXiv:1810.04805, 2018.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, “Attention is all you need,” In *Advances in Neural Information Processing Systems*, **30**, pp. 5998–6008, 2017.
[3] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel and N. Dehak, “Hierarchical transformers for long document classication,” In *Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 838–844, 2019.
[4] X. Geng, C. Yin and Z.-H. Zhou, “Facial age estimation by learning from label distributionsm,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, pp. 2401–2412, 2013.
[5] J. W. Pennebaker and L. D. Stone, “Words of wisdom: language use over the life span,” *Journal of Personality and Social Psychology*, **85**, pp.291–301, 2003.
[6] S. Argamon, M. Koppel, J. W. Pennebaker and J. Schler, “Mining the blogosphere: Age, gender and the varieties of self-expression,” *First Monday,* **12**, Number 9, 2007.
[7] S. Goswami, S. Sarkar and M. Rustagi, “Stylometric analysis of bloggers’ age and gender,” In *Proceedings of the International AAAI Conference on Web and Social Media*, **3**, pp. 214–217, 2009.
[8] D. Nguyen, N. A Smith and C. Rose, “Author age prediction from text using linear regression,” In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 115–123, 2011.
[9] O. Rozen, J. Oren and A. Raviv, “Predicting user

demography and device from news comments," In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1995–1999, 2021.
[10] A. Z. Klein, A. Magge and G. Gonzalez-Hernandez, "Reportage: Automatically extracting the exact age of twitter users based on self-reports in tweets," *PloS ONE*, **17**, e0262087, 2022.
[11] M. V. Koroteev, "BERT: A review of applications in natural language processing and understanding," *CoRR*, abs/2103.11943, 2021.
[12] 近江崇宏，金田健太郎，森長誠，江間見亜利，『Bert による自然言語処理入門—transformers を使った実践プログラミング—』，オーム社，2021.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation,* **9**, pp. 1735–1780, 1997.
[14] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint*, arXiv:1406.1078, 2014.
[15] D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning representations by back-propagating errors," *Nature,* **323**, pp. 533–536, 1986.
[16] Z. Hu, Y. Dong, K. Wang and Y. Sun, "Heterogeneous graph transformer," In *Proceedings of The Web Conference 2020*, pp. 2704–2710, 2020.
[17] X. Geng and Y. Xia, "Head pose estimation based on multivariate label distribution," In *Prodeedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1837–1842, 2014.
[18] 東北大学乾研究室,「Pretrained Japanese BERT models」，https://github.com/cl-tohoku/bert-japanese（2022 年 5 月 21 日閲覧）
[19] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint*, arXiv:1301.3781, 2013.
[20] 東北大学乾研究室,「Wikipedia Entity Vectors」，https://github.com/singletongue/WikiEntVec（2022 年 5 月 26 日閲覧）

  