Vol.2019-NL-241 No.25
2019/8/30

文符号化器のマルチタスク学習による テキスト分類モデルの頑健化

大橋 空[1,a)]　高山 隼矢[1,b)]　梶原 智之[2,c)]　Chenhui Chu[2,d)]　荒瀬 由紀[1,e)]

概要：一般的なニューラルテキスト分類モデルは、文をベクトル化する文符号化器と、文ベクトルを基に分類ラベルが付与される確率を計算する分類器からなる。このようなモデルは、特定の単語が出現する文に対し、文意に関わらずその単語との共起頻度が高いラベルに分類しやすくなるという過学習を起こしやすい。これは、文符号化器が分類に強く寄与する単語を過度に反映した文ベクトルを生成するためであると考えられる。この課題に対し本研究では、同じ（異なる）ラベルを持つ文同士のベクトルはベクトル空間で近傍（遠方）に位置すべき，という直感に基づくマルチタスク学習手法を提案する。具体的には、共通のラベルを持つ文同士の文ベクトルが類似するように、文符号化器を通常のテキスト分類タスクおよび同一ラベル判別タスクのマルチタスク学習によって訓練する。同一ラベル判別タスクでは、 コーパスからサンプリングした複数の文のうち、どれが入力文と同一のラベルを持つかを判別できるように文符号化器を訓練する。これにより、文符号化に特定の単語が過度に影響するのを抑制し、テキスト分類の性能を改善するような文ベクトルが得られると期待できる。提案手法の有効性を検証するため、単一ラベル分類の 6 つのデータセットおよび複数ラベル分類の 3 つのデータセットにおいて、2 種類の文符号化器を用いて実験を行った。また、入力が文書であるデータセットについても 1 種類の文符号化器を用いて実験を行った。これらの実験結果から、6 つのデータセットについて全ての文符号化器で提案手法がベースラインを上回る精度を達成し、提案手法の有効性が示された。

1. はじめに

Twitter などのソーシャルネットワーキングサービスやオンライン上のニュースサイトなどの様々な Web サービスの発展に伴い、日々大量のテキストが Web 上に投稿されている。これに伴い、テキストの自動分類システムへの需要が高まっている。テキスト分類は自然言語処理の中でも、感情分析、トピック分類など様々なタスクにおいて有用であることから基礎的かつ重要な問題である [1]。

テキスト分類とは、入力されたテキストに対して一つまたは複数のラベルを付与する問題である。本稿では、付与するラベルが一つのみの場合は単一ラベル分類、付与するラベルが複数の場合は複数ラベル分類と呼ぶ。テキスト分類モデルは一般的に、入力文にある単語を分散表現へと変換する Embedding 層、 Embedding 層より得られた行列を一つの文ベクトルへと変換する文符号化器、文ベクトルを基に分類を行う分類器の三つから成り立っている。近年では文ベクトルの生成にニューラルネットワークを用いたテキスト分類モデルが提案されており、大きな成功を収めている [2]。しかし、テキスト中にラベルとよく共起する単語が存在する場合、出力がその単語に強く依存してしまう。

表 1　MedWeb データセットの例 : 2 つの文には共に「風邪」という単語が含まれているが、付与されているラベルは全く異なる。

入力文	正解ラベル
風邪もりっぱな病気の 1 つ。	（なし）
最悪だー風邪ひいたー。頭がぼーっとする。	＜ 風邪 ＞

ラベルとよく共起する単語への依存が生じる過程および依存により生じる問題を表 1 を例にとって説明する。表 1 の 2 文には共に「風邪」という単語が含まれており、「風邪」という単語は ＜ 風邪 ＞ というラベルと共起する確率が高い。加えて、「風邪」以外の単語については、ラベル付与に寄与するような情報をさほど含んでいない。従って、モデルは「風邪」という単語は ＜ 風邪 ＞ というラベルに対して重要な意味を持つと学習する。そして文に「風邪」という単語が含まれていた時、文符号化器は「風邪」以外の単語を軽視し「風邪」を強く反映させたような文ベクト

[1] 大阪大学大学院情報科学研究科
[2] 大阪大学データビリティフロンティア機構
[a)] ohashi.sora@ist.osaka-u.ac.jp
[b)] takayama.junya@ist.osaka-u.ac.jp
[c)] kajiwara@ids.osaka-u.ac.jp
[d)] chu@ids.osaka-u.ac.jp
[e)] arase@ist.osaka-u.ac.jp

ルを生成するように学習してしまう。この時、二つの文ベクトルは類似しているため、分類器は 2 文の違いが判別できず、2 つの文両方に風邪についての情報が含まれていると判断する。結果、両方の文に < 風邪 > というラベルを付与してしまう。正解ラベルと照らし合わせると、1 番目の文については誤分類となってしまう。これを防ぐため、学習の過程で正解ラベルの情報とは別に、文ベクトルそのものの良し悪しについて、文符号化器にフィードバックを与える必要があると考えられる。

そこで本研究では、文ベクトル生成が適切に行われている場合は、同一のラベルが付与された文の文ベクトルは類似するという直感に基づき、文符号化器をマルチタスク学習させることで上記の問題を解決することを目指す。具体的には、モデルが解くテキスト分類問題（以下、主タスクと述べる）に加えて、主タスクの入力文に付与されているラベルと同じラベルを持つ文を判別する問題（以下、補助タスクと述べる）を学習させる。補助タスクでは主タスクの入力文とは別に、コーパスからサンプリングした入力文が複数与えられる。それぞれの入力文の文ベクトルを元に、どの文が主タスクの入力文と同一のラベル文かを判別する。具体的には、主タスクの入力文と補助タスクの各入力文の文ベクトルの内積を取り、これを判別に用いる。この補助タスクの学習を通じて、文符号化器は同一のラベルを持つ文について、それらの文ベクトルが類似するように学習される。言い換えると、適切な文ベクトルの生成が可能になり、性能の改善へと繋がることが期待される。

単一ラベル分類のデータセット 6 つ、複数ラベル分類のデータセット 4 つそれぞれについて、双方向 GRU および様々な分類問題に有効である BERT [3] を文符号化器として用いた場合について実験を行い、双方の分類問題における提案手法の有効性を確認した。さらに文書分類における有効性を検証するため、文書分類では Hierarchical Attention Network (HAN) [2] を用いた。実験結果より、双方向 GRU については 7 つのデータセットについて性能の改善が見られ、 BERT について 6 つのデータセットで性能の改善が見られた。また、 HAN についても文書分類タスクにおいて性能の改善が見られた。

2. 関連研究

2.1 テキスト分類に関する研究

文書分類を目的とする HAN では、文書中の各文はそれぞれ分類への貢献度に違いがあることに着目し、文単位での情報の取捨選択を可能としたモデルとなっている [2]。このモデルでは、まず各文を注意機構を持った RNN に入力し、得られた各文の分散表現を注意機構を持った RNN へと入力することで、単語単位のみではなく文単位での情報の取捨選択を可能としている。この研究では、文符号化器の構造を変化させることで分類精度の向上を図っているのに対し、本研究では文符号化器から出力される文ベクトルに制約を加えることで精度の向上を図る。

一方、入力が一文のみである文分類タスクでは、複数文からなるテキストとは違い入力される単語数が少なく、その影響で抽出できる特徴量が限られる [4]。これに対処するため、外部知識を利用したモデルが提案されている [5, 6]。また、近年では、外部知識などの追加リソースを必要とせず、文単位での単語の共起に着目したモデルが提案されており、短文分類の精度向上へ貢献している [1]。この手法は大量の短文データがある場合は有効であるが、短文データが少量しか入手できない場合には単語の共起パターンが十分に得られず、分類精度に悪影響を及ぼす可能性がある。本手法では単語同士の共起パターンは考慮していないため、上記のような問題は起こりづらい。

複数ラベル分類では、複数のラベルが付与される可能性があるため、ラベル間の相関やラベルの依存関係を利用して分類性能を向上することが期待される。ラベル間の相関を利用したモデルとして、分類器に再帰的ニューラルネットワークを用い、ラベルの系列を生成するモデルが提案されている [7]。ラベル間の依存関係を利用したモデルとして、依存関係を木構造とみなし、この木構造を探索してラベルを付与するモデルが提案されている [8]。しかしながら、ラベル間の依存関係を木構造で表現できるドメインは限られている。

2.2 マルチタスク学習に関する研究

マルチタスク学習では、複数のタスクを同時に解く事によって、モデルが各タスク間に共通かつ有益な情報を抽出するようになるため、あらゆる分野で大きな成功を成し遂げている。感情分析では、 sentiment と emotion の分類という、二つのタスク間の依存関係をマルチタスク学習により利用する手法 [9] が提案されている。また観点付き感情分析において、観点の分類および抽出をマルチタスク学習することで得られる共通の特徴量を用いて、分類精度を向上する手法も提案されている [10]。これらの手法では、タスク間に共通の特徴量を抽出するためにマルチタスク学習を適用している。本研究では、文符号化器がラベルとよく共起する単語へ依存するのを防ぐという、補助的な役割を追加するためにマルチタスク学習を適用している。

3. 文符号化器のマルチタスク学習

提案手法の全体図を図 1 に示す。提案手法は主タスクを解く要素と、補助タスクを解く要素より構成されている。補助タスクの入力には、訓練用データよりランダムにサンプリングした文を用いる。ただし、サンプリングした複数文のうち、1 つは主タスクと同一のラベルの文をサンプリングする。

訓練時には、主タスクと補助タスクを同時に行い、各タ

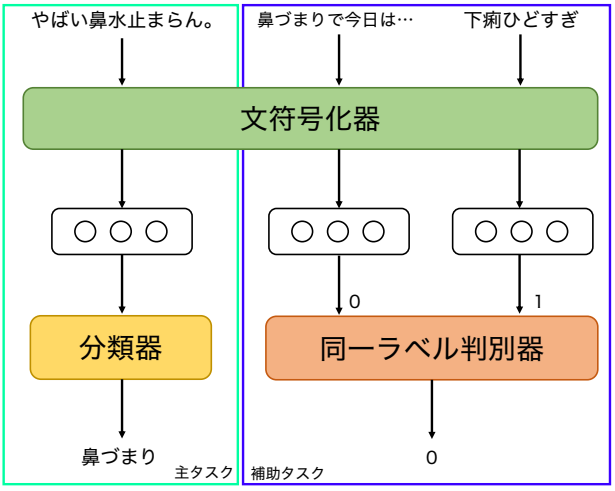

図 1　提案手法の全体図。文符号化器は主タスク、補助タスク間で重みを共有している。

スクの損失を足し合わせた結果を最終的な損失とする。つまり、損失関数 $\mathcal{L}(\boldsymbol{y}_m, \boldsymbol{y}_a, \boldsymbol{p}_m, \boldsymbol{p}_a)$ は次式となる。

$$\mathcal{L}(\boldsymbol{y}_m, \boldsymbol{y}_a, \boldsymbol{p}_m, \boldsymbol{p}_a) = \mathcal{L}_m(\boldsymbol{y}_m, \boldsymbol{p}_m) + \mathcal{L}_a(\boldsymbol{y}_a, \boldsymbol{p}_a) \quad (1)$$

ここで、各記号の意味は以下の通りである。

- $\mathcal{L}_m(\boldsymbol{y}_m, \boldsymbol{p}_m)$: 主タスクの損失関数
- $\mathcal{L}_a(\boldsymbol{y}_a, \boldsymbol{p}_a)$: 補助タスクの損失関数
- $\boldsymbol{y}_m$: 主タスクの正解ラベル
- $\boldsymbol{p}_m$: 主タスクにおいて、モデルから出力されたラベルの付与確率
- $\boldsymbol{y}_a$: 補助タスクの正解ラベル
- $\boldsymbol{p}_a$: 補助タスクにおいて、判別器から出力された確率

ただし、推論時には補助タスクは行わず、主タスクの部分のみを利用する。

以下では、節 3.1 で主タスク、節 3.2 で補助タスクについての詳細を述べる。

3.1　主タスク：テキスト分類

主タスクでは目標としている分類問題を解く。以下では文符号化器より出力される文ベクトルの次元数を d_s、単語分散表現の次元数を d_w と表記する。またラベルの種類の数を C と表記する。まず L 個の単語からなる入力文が与えられ、その入力文中に含まれる各単語は、Embedding 層を通じて分散表現へと変換される。その後、それらを連結して行列 $\boldsymbol{X} \in \mathbb{R}^{L \times d_w}$ へと変換する。この行列 $\boldsymbol{X}$ は文符号化器を用いて 1 つのベクトル $\boldsymbol{v}_m \in \mathbb{R}^{d_s}$ へと変換され、このベクトルが次に続く分類器へと入力される。提案手法では文符号化器として任意のモデルを用いることができる。

分類器では、入力された文ベクトルを基に各ラベルが付与される確率を計算する。具体的には、以下の式に従って計算される。

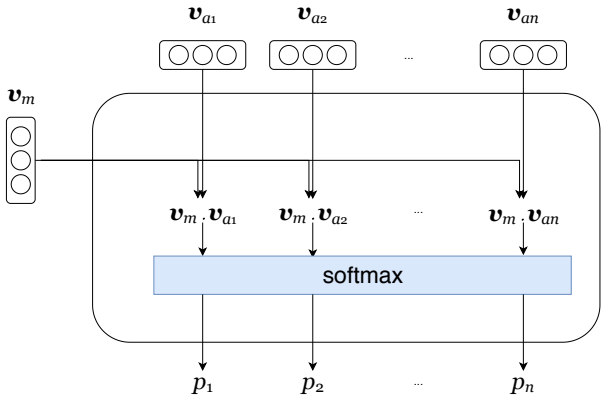

図 2　補助タスクの判別器の詳細。

$$\boldsymbol{l} = \boldsymbol{W}\boldsymbol{v}_m + \boldsymbol{b}$$
$$\boldsymbol{p} = g(\boldsymbol{l}) \quad (2)$$

ここで、p_i はラベル i が付与される確率であり、$\boldsymbol{W} \in \mathbb{R}^{d_s \times C}, \boldsymbol{b} \in \mathbb{R}^C$ はそれぞれ分類器のパラメータである。また、関数 $g(\boldsymbol{x})$ はタスクの種類によって変化し、単一ラベル分類の場合は softmax 関数であり、複数ラベル分類の場合は sigmoid 関数である。各関数は次式で定義される。

$$\mathrm{softmax}(\boldsymbol{x}_i) = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}} \quad (3)$$
$$\sigma(\boldsymbol{x}_i) = \frac{1}{1 + e^{-x_i}} \quad (4)$$

損失関数は Negative Log Likelihood を用いて計算される。単一ラベル分類の時、

$$\mathcal{L}_m(\boldsymbol{y}_m, \boldsymbol{p}_m) = -\sum_i [y_{m_i} \log(p_{m_i})] \quad (5)$$

となり、複数ラベル分類の場合は次のように計算される。

$$\mathcal{L}_m(\boldsymbol{y}_m, \boldsymbol{p}_m) = -\sum_i [y_{m_i} \log(p_{m_i}) + (1 - y_{m_i}) \log(1 - p_{m_i})] \quad (6)$$

3.2　補助タスク：同一ラベル判別

補助タスクでは、入力される複数の文から、どの文が主タスクの入力と同一のラベルを持つかを予測する。以下では、主タスクに入力された文のベクトルを $\boldsymbol{v}_m$、補助タスクに入力された文のベクトルを $\boldsymbol{v}_{a_1}, \boldsymbol{v}_{a_2}, ..., \boldsymbol{v}_{a_n}$ と表記する。補助タスクの判別器の詳細を図 2 に示す。

補助タスクでは、まず主タスクとは別の入力文を訓練データよりランダムにサンプリングする。この時、サンプリングした入力文のうち 1 つのみ主タスクの入力と同一のラベルが付与されている入力文を選び出すように行う。これらの文は主タスクと共通の文符号化器によってベクトルへと変換される。その後、各 $\boldsymbol{v}_{a_i} (1 \leq i \leq n)$ について、$\boldsymbol{v}_m, \boldsymbol{v}_{a_i}$ の類似度を計算する。類似度の計算には内積を用いる。最後に、 softmax 関数を通じて、入力文 i が主タス

表 2　データセットの統計情報：MR, CR, SUBJ については交差検定時におけるデータ数を示している。

	入力の単位	ラベル数	訓練データ数	開発データ数	テストデータ数
MR	文	2	8,529	1,066	1,066
CR	文	2	3,020	377	377
SST-2	文	2	67,349	872	1,821
SST-5	文	5	8,544	1,101	2,210
TREC	文	6	4,361	1,090	500
SUBJ	文	2	8,000	1,000	1,000
MedWeb	文	8	1,536	384	640
arXiv	文書	40	38,188	9,548	11,935

クと同じラベルを持つ確率が計算される。つまり、図 2 における p_i は、$\boldsymbol{c} = (\boldsymbol{v}_m \cdot \boldsymbol{v}_{a_1}, \boldsymbol{v}_m \cdot \boldsymbol{v}_{a_2}, ..., \boldsymbol{v}_m \cdot \boldsymbol{v}_{a_n})$ としたとき、式 (7) で表現される。

$$p_{a_i} = \mathrm{softmax}(\boldsymbol{c}_i) \tag{7}$$

次に損失関数について述べる。補助タスクの i 番目の入力 $\boldsymbol{v}_{a_i}$ が主タスクの入力と同じラベルを持つとき、$\boldsymbol{y}_a$ の各要素 $y_{a_j}(1 \leq j \leq n)$ は次のように表される

$$y_{a_j} = \begin{cases} 1 & (j = i) \\ 0 & (j \neq i) \end{cases} \tag{8}$$

補助タスクの損失関数には式 (9) で定義されるクロスエントロピーを用いる。

$$\mathcal{L}_a(\boldsymbol{y}_a, \boldsymbol{p}_a) = -\sum_{k=1}^{n} [y_{a_k} \log(p_{a_k}) + (1 - y_{a_k}) \log(1 - p_{a_k})] \tag{9}$$

4. 実験

4.1 データセット

本手法の有効性を評価するため、単一ラベル分類データセットとして SentEval [11] の MR, CR, SST-2, SST-5, TREC, SUBJ データセットを用いて実験を行った。MR, CR, SUBJ データセットについては開発データが含まれていないため、5 分割交差検定を用いて評価を行った。また、複数ラベルデータセットとして NTCIR-13 MedWeb タスク [12] のデータセットと arXiv データセットを使用した。表 2 に各データセットのサイズを示す。

4.1.1 単一ラベル分類

単一ラベル分類データセットの詳細は以下の通りである。

MR　映画のレビューについて、 positive, negative 2 値分類を行う単一ラベル分類問題のデータセットである。

CR　MR と同様に、商品レビューについて 2 値の感情分析を行う。

SST-2　MR と同様に、映画のレビューについて 2 値の感情分析を行う。

SST-5　映画のレビューについて negative, somewhat negative, neutral, somewhat positive, positive の 5 値分類を行う。

TREC　質問文が入力として与えられ、この質問文が何を尋ねているかを分類する。ここで使用されているラベルとその意味を表 A·2 に示す。

SUBJ　入力文が主観的であるか客観的であるかの 2 値分類を行う。

MR, CR, SST-2, SST-5, TREC, SUBJ データセットについては、SentEval の github ページ[*1]より得られるダウンロード用のスクリプトを用いて入手した。

4.1.2 複数ラベル分類

複数ラベル分類データセットの詳細は以下の通りである。

MedWeb　入力文が含んでいる疾病情報をラベルとして付与する複数ラベル分類問題を行う。このデータセットで使用するラベルとその意味を表 A·2 に示す。英語、日本語、中国語の 3 言語について実験を行った。

arXiv　arXiv[*2]に公開されている論文のカテゴリを付与する複数ラベル分類を行う。入力は論文の概要である。

MedWeb データセットは NTCIR-13 MedWeb の Web ページ[*3]より入手した。 arXiv データセットは、arXiv API [*4]を用いて 2019 年 1 月 1 日から 2019 年 6 月 4 日までに投稿された論文の概要とカテゴリを収集し、コンピュータサイエンスのカテゴリに属する論文のみをデータセットとして利用した。[*5]

4.2 評価指標

本実験の評価の指標として、単一ラベル分類では Accuracy を、複数ラベル分類では式 (10) で定義される Exact

*1 https://github.com/facebookresearch/SentEval
*2 https://arxiv.org/
*3 http://research.nii.ac.jp/ntcir/permission/ntcir-13/perm-ja-MedWeb.html
*4 https://arxiv.org/help/api
*5 Yang ら [7] によって構築された arXiv データセットでは論文中の改行情報が失われており、復元不可能であった。そのため、本実験では独自に収集した arXiv データを用いる。

表 3　単一ラベル分類の実験結果。

	MR	CR	SST-2	SST-5	TREC	SUBJ
	Accuracy					
BERT	86.3	89.3	**92.8**	51.7	96.6	**96.6**
BERT + prop	**86.7**	**89.7**	92.5	**52.0**	**97.2**	96.5
双方向 GRU	77.1	78.8	85.8	42.6	86.8	92.2
双方向 GRU + prop	77.9	78.0	85.8	42.7	87.4	92.4

表 4　複数ラベル分類の実験結果。

	MedWeb(en)	MedWeb(ja)	MedWeb(zh)	arXiv
	Exact Match			
SOTA（文献 [13] より引用）	79.5	82.5	80.9	-
BERT	**83.5**	85.5	85.8	-
BERT + prop	**83.5**	**86.0**	**86.0**	-
双方向 GRU	76.3	83.0	81.3	-
双方向 GRU + prop	78.3	83.6	81.5	-
HAN	-	-	-	41.7
HAN + prop	-	-	-	**42.8**

Match を用いる。

$$\text{ExactMatch} = \frac{1}{N}\sum_{i=1}^{N} E(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) \tag{10}$$

ここで、$\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i$ は正解ラベルおよび予測したラベルをベクトルで表現したものであり、ラベル j が付与される時 $y_{ij} = 1$ となり、それ以外の場合 0 となる。また、$E(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)$ は以下の式で定義される。

$$E(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) = \begin{cases} 1 & (\boldsymbol{y}_i = \hat{\boldsymbol{y}}_i) \\ 0 & (\text{otherwise}) \end{cases} \tag{11}$$

4.3　実験設定

文符号化器には BERT [3] を用いた場合と、双方向 GRU [14, 15] を用いた場合のそれぞれについて実験を行った。 arXiv データセットについては入力が文書であるため、 HAN [2] を使用した。

BERT については、英語および中国語はそれぞれの言語の訓練済みモデルを、日本語は Multilingual モデルを用い、各タスクにおいて fine-tuning を行う。[*6] また、双方向 GRU について、Embedding 層のパラメータは fastText [16] の訓練済み単語ベクトルを用いて初期化した。

主タスクのバッチサイズは 32、補助タスクのサンプリング数は 4 とした。補助タスクのサンプリング数については、双方向 GRU を用いた時の MR データセットを用いて設定した。具体的には、サンプリング数が $4, 8, 16$ の場合における開発データの Accuracy が最も高いものを採用した。モデルの訓練には early stopping を適用し、10 エポック連続で開発データの評価指標が上昇しなかった時訓練を打ち止め、最大のスコアを記録したエポックのモデルで評価を行った。モデルの最適化アルゴリズムには Adam [17] を用い、 Adam の各パラメータについて、学習率 $\alpha = 0.001, \beta_1 = 0.999, \beta_2 = 0.9$ とした。

[*6] https://github.com/google-research/bert

4.4　比較手法

比較手法として、主タスクのみで訓練を行った場合の BERT および双方向 GRU をベースラインとする。MedWeb においては最高性能を記録している既存手法 [13] のうち、提案手法同様シングルモデルと比較する。

4.5　結果

結果の表を表 4 に示す。双方向 GRU について、CR、SST-2 を除く全てのデータセットにおいて精度の改善が見られ、 BERT については SST-2 を除く全てのデータセットで精度の改善が確認できた。また、入力が文書である arXiv データセットについても精度が改善された。加えて、MedWeb においては、既存の最高性能を更新する性能を達成している。提案手法である補助タスクにより、ラベルとよく共起する単語が文ベクトルに過度に影響するのを防げたためと考えられる。

5.　考察

5.1　ラベルとよく共起する単語への効果

ラベルとよく共起する単語への効果について評価するため、まず各データセットにおいて単語とラベルとの共起を

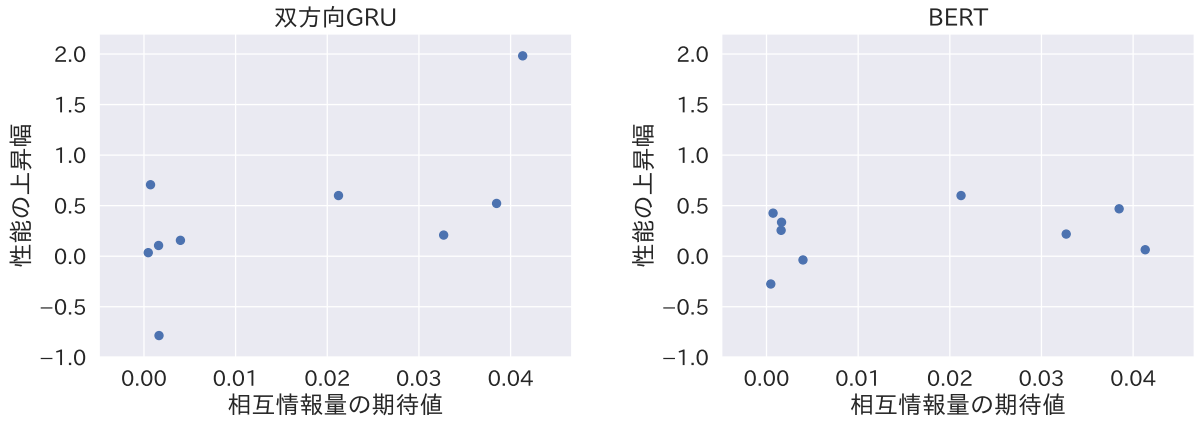

図 3 相互情報量と性能変化の散布図： GRU の方が BERT より相関が強い

表 5 各データセットにおける相互情報量の期待値：SST-2 の期待値は最も低い

データセット	相互情報量の期待値
MR	7.16×10^{-4}
CR	1.64×10^{-3}
SST-2	4.67×10^{-4}
SST-5	1.59×10^{-3}
TREC	2.12×10^{-2}
SUBJ	3.98×10^{-3}
NTCIR (en)	4.13×10^{-2}
NTCIR (ja)	3.84×10^{-2}
NTCIR (zh)	3.27×10^{-2}
arXiv	7.19×10^{-3}

調査した。具体的には、1 文中の各単語の出現の有無とラベルとの相互情報量を求め、それぞれの相互情報量について単語の出現頻度での重み付き平均をとった。この値は、文中の 1 単語が持つ相互情報量の期待値となる。ラベルと単語がよく共起する場合相互情報量は大きくなるため、あるデータセットおいて相互情報量の期待値が大きいことはラベルとよく共起する単語が多いことを示す。そのため、提案手法がラベルとよく共起する単語への依存を解消している時、この期待値と性能の上昇幅に正の相関があることが期待される。

BERT および双方向 GRU について、各データセットにおけるベースラインからの性能変化と相互情報量の期待値をプロットしたものを図 3 に示す。双方向 GRU についてはある程度の正の相関が見られ、提案手法によってラベルとよく共起する単語への依存がある程度解消されているのが確認できる。BERT については、双方向 GRU と比較して相関が弱い。原因の一つとして、 BERT は各層において multi-head のアテンションを適用しており、文中の 1 箇所のみではなく複数の箇所に注目を置くことが可能であるため、1 単語が及ぼす影響が双方向 GRU と比較して小さいことが挙げられる。

5.2 分類問題の種類に関する依存性

表 4 に示すように、単一ラベル分類と複数ラベル分類双方について、提案手法を用いることで精度は概ね上昇している。また、入力が文書である場合も精度が上昇している。単一ラベル分類である SST-2 データセットでは精度の上昇が確認できなかったが、これは表 5 からもわかるように、ラベルとよく共起する単語が他と比べて少ないなどデータセットの特徴に起因していると考えられる。以上を踏まえると、タスクが単一ラベル分類か複数ラベル分類かには依存せず、提案手法を用いることで精度を上昇させることが可能であると考えられる。

5.3 出力例

表 6 に日本語 MedWeb データセットの、BERT における出力例を示す。また、図 4 に 11 層目から 12 層目における、各ヘッドのアテンションスコアを示す。

一番目の例では、風邪が治った事について言及しているため、この文には疾病情報は含まれていない。ベースラインでは、「風邪」と言う単語に依存してしまい、誤って < 風邪 > のラベルを付与しているが、提案手法では、「風邪」という単語に影響を受けず、正確に分類ができている。図 3 より、ベースラインでは「風邪」の箇所に注目しており、これが誤って < 風邪 > のラベルを付与してしまったと思われるが、提案手法では文中にある「風邪」の箇所にはあまり注目せず、「治って完全だせ」の箇所にアテンションが集中しており、この文に疾病情報が含まれていないことを認識できているのが確認できる。これらのことから、提案手法によって、ラベルとよく共起する単語に依存せずより適切な文ベクトル生成が行われるように文符号化器が訓練され、それが分類性能に寄与したと考えられる。

6. おわりに

本研究では、文符号化器をマルチタスク学習させる事で、

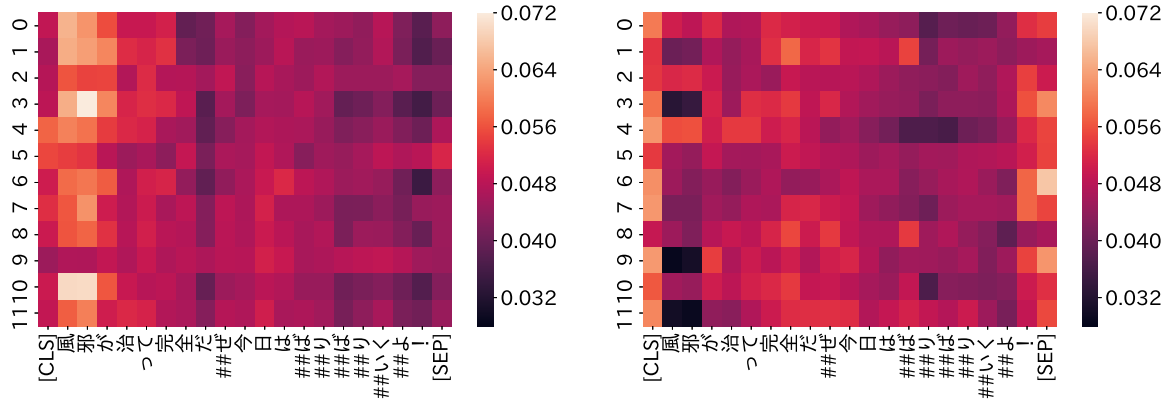

図 4 日本語の MedWeb データセットにおける BERT のアテンションスコア。左はベースラインのアテンションスコアで、右は提案手法のアテンションスコア。ベースラインでは「風邪」に注目しているのに対し、提案手法では「治って完全だぜ」の箇所に注目している。

表 6 日本語の MedWeb データセットにおける BERT 出力の一例：ベースラインではラベルとよく共起する単語への依存が見られるが、提案手法ではそれが見られない。

入力文	出力ラベル		
	ベースライン	提案手法	正解
風邪が治って完全だぜ今日はばりばり行くよ！	風邪	（なし）	（なし）

分類ラベルと頻繁に共起する様な特定の単語が過度に影響を与える過学習を防ぐ手法を提案した。評価実験の結果、6つの文分類タスクにおいて、双方向 GRU および BERT それぞれの性能を改善することを確認した。また、文書分類においても、HAN の性能を改善することを確認した。分析により、文符号化器によって程度の差は見られるが、提案手法を用いることでラベルとよく共起する単語への依存を防げることが示された。今後は、補助タスクの入力について、判断を誤りやすい文ペアを優先的にサンプリングするなど、サンプリング方法についてさらなる検討を行う。

参考文献

[1] Zeng, J., Li, J., Song, Y., Gao, C., Lyu, M. R. and King, I.: Topic Memory Networks for Short Text Classification, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3120–3131 (2018).

[2] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E.: Hierarchical Attention Networks for Document Classification, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489 (2016).

[3] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805* (2018).

[4] Phan, X.-H., Nguyen, L.-M. and Horiguchi, S.: Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections, *Proceedings of the 17th International Conference on World Wide Web*, pp. 91–100 (2008).

[5] Lucia, W. and Ferrari, E.: EgoCentric: Ego Networks for Knowledge-based Short Text Classification, *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pp. 1079–1088 (2014).

[6] Wang, W., Feng, S., Gao, W., Wang, D. and Zhang, Y.: Personalized Microblog Sentiment Classification via Adversarial Cross-lingual Multi-task Learning, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348 (2018).

[7] Yang, P., Sun, X., Li, W., Ma, S., Wu, W. and Wang, H.: SGM: Sequence Generation Model for Multi-label Classification, *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3915–3926 (2018).

[8] Shimura, K., Li, J. and Fukumoto, F.: HFT-CNN: Learning Hierarchical Category Structure for Multi-label Short Text Categorization, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 811–816 (2018).

[9] Akhtar, M. S., Chauhan, D., Ghosal, D., Poria, S., Ekbal, A. and Bhattacharyya, P.: Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 370–379 (2019).

[10] Xue, W. and Li, T.: Aspect Based Sentiment Analysis with Gated Convolutional Networks, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2514–2523 (2018).

[11] Conneau, A. and Kiela, D.: SentEval: An Evaluation Toolkit for Universal Sentence Representations, *arXiv preprint arXiv:1803.05449* (2018).

[12] Wakamiya, S., Morita, M., Kano, Y., Ohkuma, T. and Aramaki, E.: Overview of the NTCIR-13: MedWeb Task, *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 40–49 (2017).
[13] Iso, H., Ruiz, C., Murayama, T., Taguchi, K., Takeuchi, R., Yamamoto, H., Wakamiya, S. and Aramaki, E.: NTCIR13 MedWeb Task: multi-label classification of tweets using an ensemble of neural networks, *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 56–61 (2017).
[14] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734 (2014).
[15] Schuster, M. and K. Paliwal, K.: Bidirectional recurrent neural networks, *Signal Processing, IEEE Transactions on*, pp. 2673 – 2681 (1997).
[16] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C. and Joulin, A.: Advances in Pre-Training Distributed Word Representations, *Proceedings of the International Conference on Language Resources and Evaluation* (2018).
[17] Kingma, D. and Ba, J.: Adam: A Method for Stochastic Optimization, *Proceedings of the International Conference on Learning Representations* (2014).

付　　録

A.1 TREC, MedWeb, arXiv データセットで用いたラベル一覧

表 **A·1**　MedWeb で使用されるラベルの一覧とその意味

ラベル	意味
鼻水	鼻水・鼻づまりがあるか否か
咳	咳・痰が出るか否か
インフルエンザ	インフルエンザであるか否か
下 痢	下痢であるか否か
花粉症	花粉症であるか否か
熱	熱があるか否か
頭痛	頭痛があるか否か
風邪	風邪をひいているか否か

表 **A·2**　TREC で使用されるラベルの一覧とその意味

ラベル	意味
ABBREVIATION	略語についての質問をしている
ENTITY	エンティティについての質問をしている
DESCRIPTION	定義や理由などの説明を求めている
HUMAN	ある人物についての質問をしている
LOCATION	場所を尋ねている
NUMERIC	数値を尋ねている

表 **A·3**　arXiv データセットで使用するラベルの一覧

cs.AI	cs.AR	cs.CC	cs.CE	cs.CG
cs.CL	cs.CR	cs.CV	cs.CY	cs.DB
cs.DC	cs.DL	cs.DM	cs.DS	cs.ET
cs.FL	cs.GL	cs.GR	cs.GT	cs.HC
cs.IR	cs.IT	cs.LG	cs.LO	cs.MA
cs.MM	cs.MS	cs.NA	cs.NE	cs.NI
cs.OH	cs.OS	cs.PF	cs.PL	cs.RO
cs.SC	cs.SD	cs.SE	cs.SI	cs.SY