マルチタスク学習における低コストな補助タスク設計に関する研究

メディア情報学研究室
木村優介

MEDIA
INFORMATICS
LABORATORY

発表資料のダウンロード

次のQRコードから本日の発表資料を
ダウンロードできます

博論の概要

文書分類タスクにおいて，
言語モデルを効果的に
ファインチューニングする技術を提案

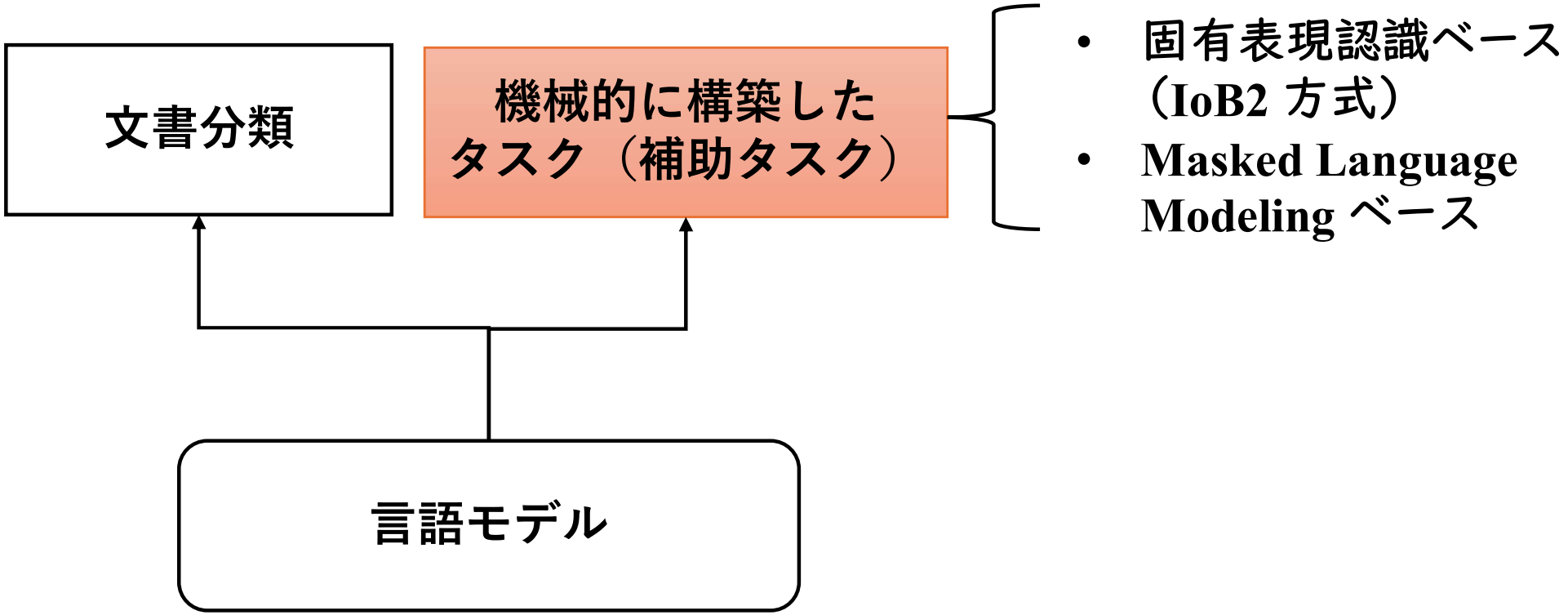
要点：

従来，人手のコストがかかるタスクを機械的に
構築して利用するマルチタスク学習フレームワークを構築

機械的に構築するタスクは，ラベリング方法の違いから
二つのタスク設計の方法を提案

補助タスク設計法の提案

マルチタスク学習の補助タスクをラベリング方法の違いから二つ提案



言語モデル

文における単語や文字の予測確率を算出する確率モデル

古典的な言語モデルである N-gram
（文をn 文字単位で分割）の確率モデルは次の式で表される

$$p(w_i|w_{i-n} \dots w_n) = \frac{C(w_{i-n} \dots w_{i-1} w_i)}{C(w_{i-n} \dots w_{i-1})}$$

ある文中のi 番目のn-gramである w_i の予測確率は，直前の文字列 の出現頻度C でw_i を含む文字列の出現頻度を割ることで求められる

例：$p(dog|The\ quick\ brown\ fox\ jumps\ over\ the\ lazy)$

現代の言語モデルの構造

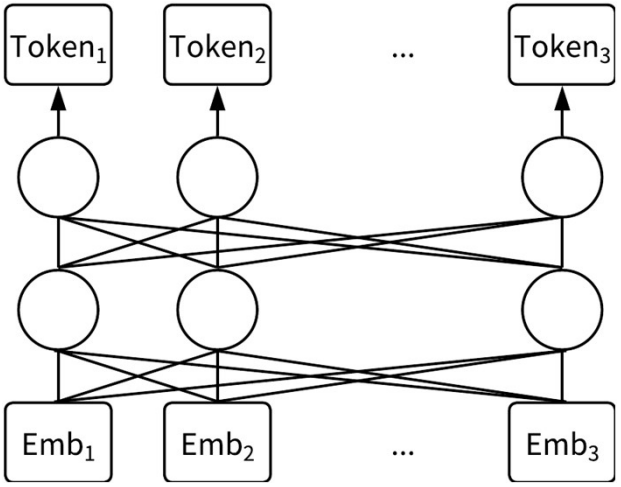
深層学習モデルを用いて大量の文書を学習した
言語モデルは，人間が文書に記載する際の
言語パタンを高精度に予測

言語モデルの構造として，3種類存在

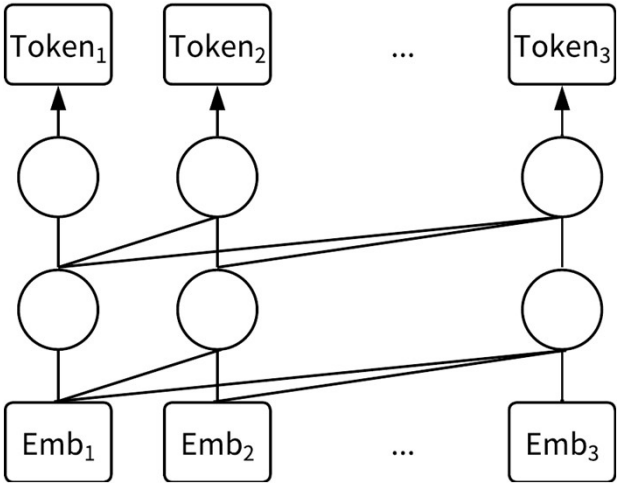
- エンコーダモデル
  - テキストからベクトルを生成
- エンコーダ・デコーダモデル
  - テキストからベクトルへ，ベクトルからテキストを生成
- デコーダモデル（大規模言語モデルの主流）
  - テキストからテキストを生成

言語モデルの構造の違い

現代の言語モデルの多くが Transformer [1] とよばれる深層学習モデルを基盤としている
（トークンは言語モデルが扱う最小の言語単位のこと）



Transformer における
エンコーダの構造



Transformer における
デコーダの構造

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

言語モデルの応用先

言語モデルの構造によって効果的な応用先は異なる

- エンコーダモデルの場合，
  - 文中の人物名や組織名などの固有表現が文中にどこにあるかを予測すること（固有表現認識）
  - 文や文書がどんなカテゴリやラベルに分類されるかを予測すること（文書分類）
- デコーダモデルの場合，
  - 自由記述のQA

分類タスクはエンコーダモデルが効率的な [2, 3]ため，

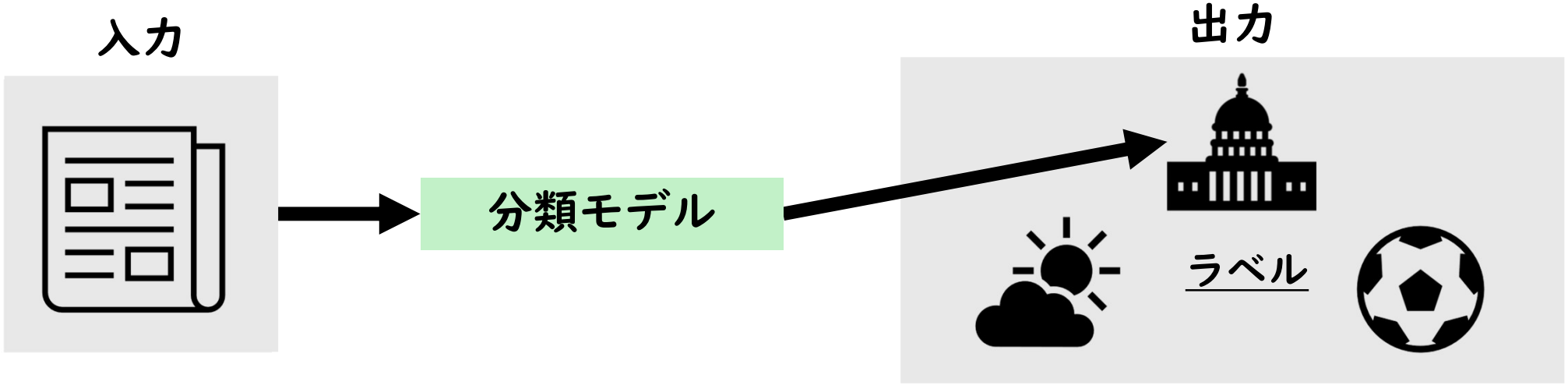
依然としてエンコーダモデルの研究は必要

[2] Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. arXiv, https://arxiv.org/abs/2406.08660

[3] Zhen Zhang, Yuhua Zhao, Hang Gao, and Mengting Hu. 2024. LinkNER: Linking Local Named Entity Recognition Models to Large Language Models using Uncertainty. In Proceedings of the ACM Web Conference 2024 (WWW '24). Association for Computing Machinery, pp.4047–4058. https://doi.org/10.1145/3589334.3645414

文書分類

事前に作成されたカテゴリに文や文書を分類



例：ニュース文書のカテゴリ分類

他にも，文の引用意図分類や病気カテゴリ分類など存在

言語モデルの学習パラダイム

エンコーダモデルの学習は**事前学習**と
ファインチューニングで構成される

1. 事前学習

人間の言語パタンを
学習（大量の文書を学習）

日本語（にほんご、にっぽんご）は、
日本国内や、かつての日本領だった国、
そして国外移民や移住者を含む日本人
同士の間で使用されている言語。

2. ファインチューニング

特定のタスクを学習

例：文書分類タスク

入力：石破総理大臣と
トランプ大統領が～
答え：政治カテゴリ

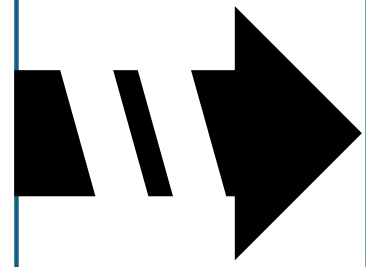
例：固有表現認識タスク

入力：石破総理大臣と
トランプ大統領が～
答え：[石破総理大臣]

事前学習とタスクの文脈のズレ

タスクにおける文脈（例：文法や用語の使われ方）が事前学習と異なる場合，タスクの精度は低下 [4]

事前学習
Wikipedia や 新聞などの
多くの人が読みやすい
文書が多い



ファインチューニング
例：医療文書を
病気カテゴリに
分類するタスク

事前学習コーパスの中で
比較的少量のドメイン

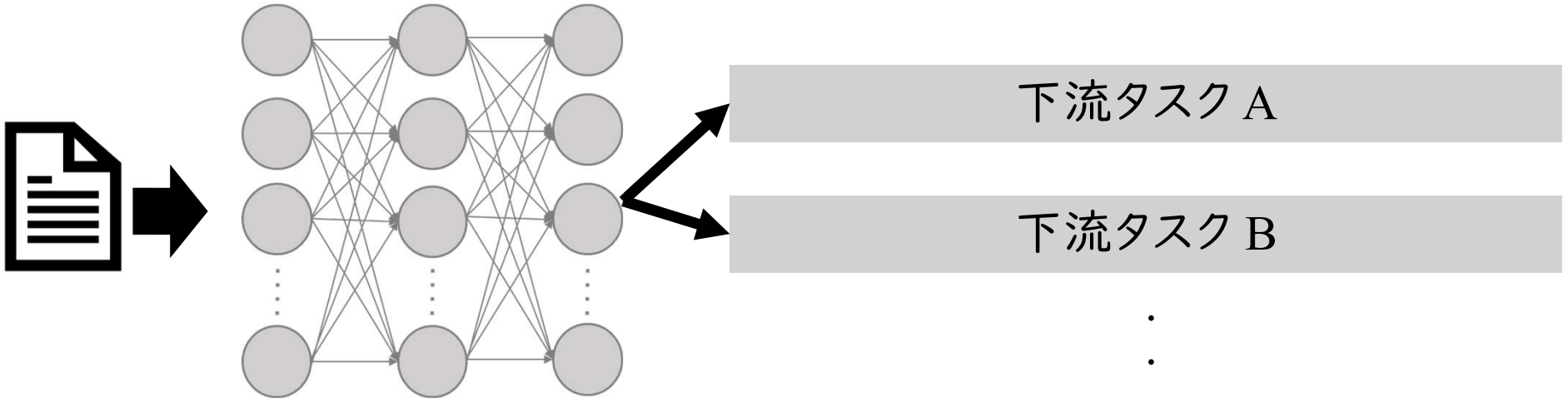
[4] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

効果的なファインチューニング

少量データで効果的に最適化する技術が必要

マルチタスク学習 [5]

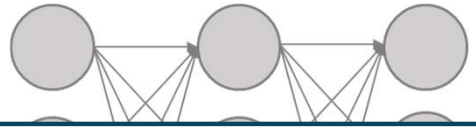
- 共通のパラメタを持つモデルで複数のタスクを学習
- 各タスクの共通性や相互補完を用いて各下流タスクを学習



[5] Rich Caruana. 1997. “Multitask Learning”, Machine Learning,Vol 28, No.1, pp.41–75.

補助タスク構築に関わる人的・金銭的コスト

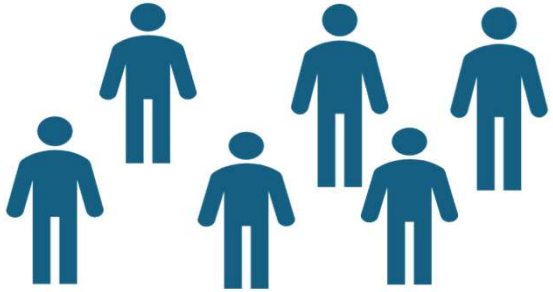
文書分類タスクのデータには，文書分類タスク以外のラベルが付与されていないことが多い



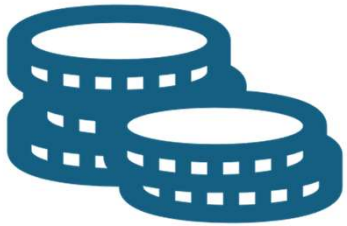
主タスク（文書分類タスク）

…

ラベル付きデータを用意する場合：



専門家・クラウドソーシングで集めた人など



データ作成の依頼料

補助タスク構築コストの削減

本論文では，補助タスクを機械的に構築することで，補助タスクのラベル付けに必要なコストを削減することを目指す

ラベリング種類

主タスクの言語パタンを効果的に学習する補助タスクを機械的に構築するにあたり,
これまで使われてきたラベリング方法を利用

- 固有表現認識におけるラベリング
  各トークンが特定の単語やフレーズに該当するかでラベリングされる
- Masked Language Modeling（MLM）タスクにおけるラベリング
  マスクされたトークンを正しいトークンに分類するタスクでトークンIDがラベリングされる

固有表現認識ベースのラベリング

文書分類を目的とした サブワードベースの トークン N-gram 認識タスク

マルチタスク学習における固有表現認識

文書分類を目的としたマルチタスク学習において，人手によってラベリングされた固有表現認識タスクが有効とされてきた [6]

固有表現：人物名や組織名などのエンティティ

Token	Barack	Obama	visited	the	United	Nations	Headquarters
Entity Label	PER	PER	O	O	ORG	ORG	LOC

固有表現はそれ自体に意味を持つ単語単位のラベリング

[6] Honglun Zhang, Liqiang Xiao, Yongkun Wang, and Yaohui Jin. A generalized recurrent neural architecture for text classification with multi-task learning.
In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pp. 3385–3391, 2017.

サブワード

固有表現と異なり，言語モデルの最小単位は
トークンであり，BERT やModernBERT などの多くの
言語モデルは単語よりも小さな文字列（サブワード）を
トークンとして用いる

単語単位	Barack	Obama	visited		the	United	Nations	Headquarters		
サブワード単位	Barack	Obama	visit	##ed	the	United	Nations	Head	##quarter	##s

サブワードはそれ自体に意味を持たない

サブワードベースの認識タスク

意味を持たないサブワードの連続が
補助タスクとして有効かは未知数

下流タスクの言語パタンを獲得する点で，
各文書分類クラス（カテゴリ）によく出現するトークンの N-gram が有効な可能性有

Token	Barack	Obama	visit	##ed	the	United	Nations	Head	##quarter	##s
Entity	PER	PER	O	O	O	ORG	ORG	LOC	LOC	LOC
Token N-gram	**0**	**0**	**B**	**I**	O	O	O	O	**B**	I

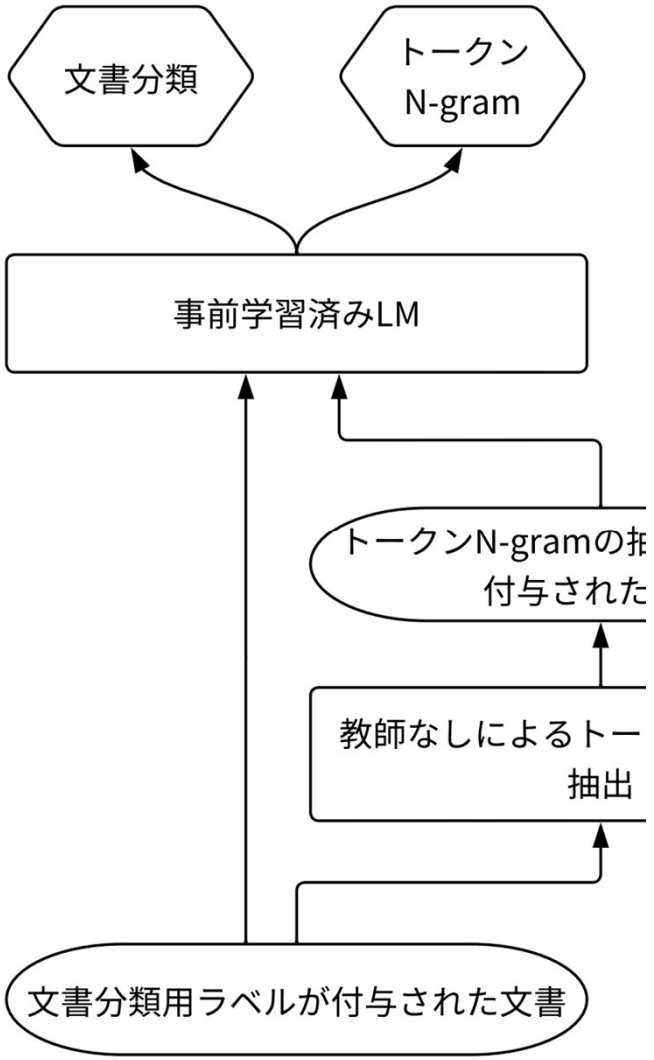
補助タスクを用いた文書分類

教師なし手法で補助タスクを構築する MTL ベースの文書分類を目的としたフレームワークを提案

このフレームワークは,

1. トークンN-gramを抽出,
   ラベリングするステップ
   （機械的に補助タスクを構築）
2. マルチタスク学習を行うステップ

で構成



トークン N-gram の抽出

Byte Pair Encoding (BPE) [7] を用いて，
よく共起するトークンN-gramを抽出

ステップ	マージ	共起頻度	更新対象のトークン
0	-	-	[laenderbank, expects, modest, profit, rise, in, oesterreichische, laenderbank, ag, olbv, vi, expects, ...]
1	(expects, modest)	2	[laenderbank, expects_modest, profit, rise, in, oesterreichische, laenderbank, ag, olbv, vi, expects_modest, ...]
2	(laenderbank, expects_modest)	2	[laenderbank_expects_modest, profit, rise, in, oesterreichische, laenderbank_expects_modest, ag, olbv, vi, ...]
3	(profit, rise)	1	[laenderbank_expects_modest, profit_rise, in, oesterreichische, laenderbank_expects_modest, ag, olbv, vi, ...]

[7] Rico Sennrich, Barry Haddow, Alexandra Birch (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725

文書分類クラス固有のラベリング

どのクラス固有のトークン N-gram かを明らかにするラベリング（Disregard, Common-Label, Bit-Label）

- All-Phrase: クラスを考慮しない, Disregard: クラス固有のN-gram, Common-Label: クラスにまたがるN-gram はØを付ける
  Bit-Label: どこのクラスに出現したかを記述

Index	Token	All-Phrase	Disregard	Common-Label	Bit-Label
1	laenderbank	O	O	O	O
2	expects	O	O	O	O
3	a	O	O	O	O
4	modest	B-Phrase	B-Class1	B-Class1	B-10
5	profit	I-Phrase	I-Class1	I-Class1	I-10
6	rise	I-Phrase	I-Class1	I-Class1	I-10
7	in	O	O	O	O
8	oesterreichische	B-Phrase	B-Class1	B-Ø	B-11
9	laenderbank	I-Phrase	I-Class1	I-Ø	I-11
10	ag	I-Phrase	I-Class1	I-Ø	I-11

データセット

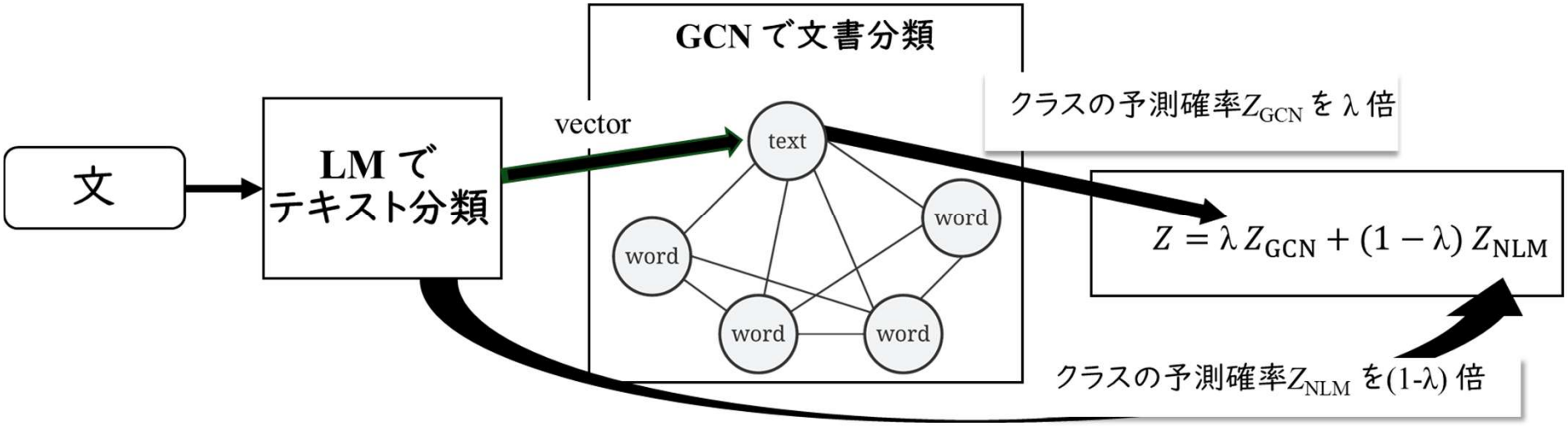
文書分類のみを学習したモデル（Baseline）と高精度な既存手法と比較するために，文書分類タスクで一般的な 5 種類のデータセットを利用

- 20 News Group (20NG), R8, R52: ニュース文書（トピック分類）
- Ohsumed (OHS): 医療文書（病気への分類）
- Movie Review (MR): 映画のレビュー（映画のスコア分類）

	MR	20NG	R8	R52	OHS
#Train	6,398	10,183	4,937	5,879	3,022
#Valid	710	1,131	548	653	335
#Test	3,554	7,532	2,189	2,568	4,043
#Class	2	20	8	52	23
Avg.#Instances/Class	5,331	942	959	175	321
Std.#Instances/Class	0	94	1,309	613	305

高精度な既存手法

BertGCN（BERTとGCNを組み合わせた手法）[8] を本研究の比較相手として利用（BERT の部分は他エンコーダモデルでも可）



[8] Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, Fei Wu (2021). “BertGCN: Transductive text classification by combining GNN and BERT”, In *Findings of the Association for Computational Linguistics”*, pp. 1456–1462.

評価実験の結果

マイクロF値（各クラスのインスタンス数が考慮される）

Model	MR	20NG	R8	R52	OHS
RoBERTaGCN	0.880	**0.894**	0.979	0.944	**0.736**
Baseline (RoBERTa)	0.881	0.831	0.977	0.962	0.690
Proposed - All-Phrase	**0.888**	0.838	0.979	0.967	0.705
Proposed - Common-Label	0.860	0.850	0.978	0.967	0.704
Proposed - Bit-Label	0.882	0.846	0.979	0.968	0.711
Proposed - Disregard	0.866	0.851	0.979	**0.969**	0.711

提案手法は，Baseline よりも大きく精度が向上せず，
インスタンス数が多いクラスの精度向上には寄与していない

評価実験の結果

マクロF値（各クラスのインスタンス数を考慮しない）

Model	MR	20NG	R8	R52	OHS
RoBERTaGCN	0.880	**0.861**	0.925	0.756	0.605
Baseline (RoBERTa)	0.881	0.825	0.943	0.836	0.594
Proposed - All-Phrase	**0.888**	0.832	0.948	0.842	0.622
Proposed - Common-Label	0.860	0.845	0.947	0.841	0.610
Proposed - Bit-Label	0.882	0.840	0.953	**0.866**	0.636
Proposed - Disregard	0.866	0.845	**0.955**	0.851	**0.637**

提案手法は各クラスのインスタンス数が不均衡なデータ（R8, R52, OHS）でBaseline よりも 1~3 ポイント精度向上した
（少量データに効果的）

分析

以下の各クラス固有の高頻度トークンN-gram の上位5語を見ると, クラス固有の単語やフレーズで構成されるトークンN-gramや意味解釈不可能なサブワードの連結が見られる

クラス名			
bop	money-fx	orange	クラス共通の Token N-gram
money supply	o ##pec	consumer prices	re ##uter
mln dlrs ibillion dlrs	b ##pd	consumer price index	fe ##b
week ended	ce ##uador	statistics institute	l ##me
borrow ##ings	crude oil	rose pct february	jan ##uary
business loans	sa ##udi	pct march	feb ##ruary

クラス固有のサブワードのN-gram が文書分類に効果的に寄与した可能性有

MLMベースのラベリング

L3Masking を用いた MLM タスク

MLM タスク

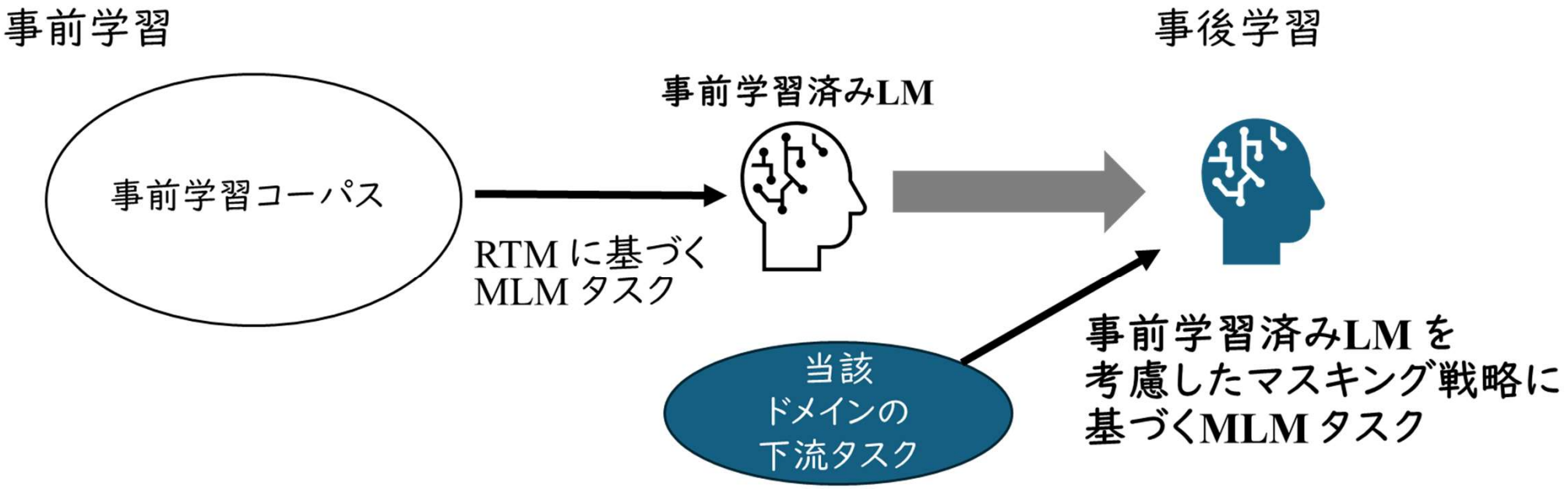
基本的には, Random Token Masking（RTM）とよばれるトークンのマスキング戦略が用いられる

- 文中の 15 % のトークンを選択し, それらのトークンの内 80 % を[MASK] トークンに置換, 残りの 10 % をランダムなトークンに置換, 残りの10 % をそのままにする（BERTの設定）

RTM の例	The	quick	brown	fox	jumps	over	the	lazy	dog	.
①	The	**cat**	brown	**[MASK]**	jumps	over	the	lazy	dog	.
②	The	quick	brown	fox	jumps	over	the	**[MASK]**	dog	.
③	The	quick	brown	fox	**[MASK]**	**[MASK]**	the	lazy	dog	.

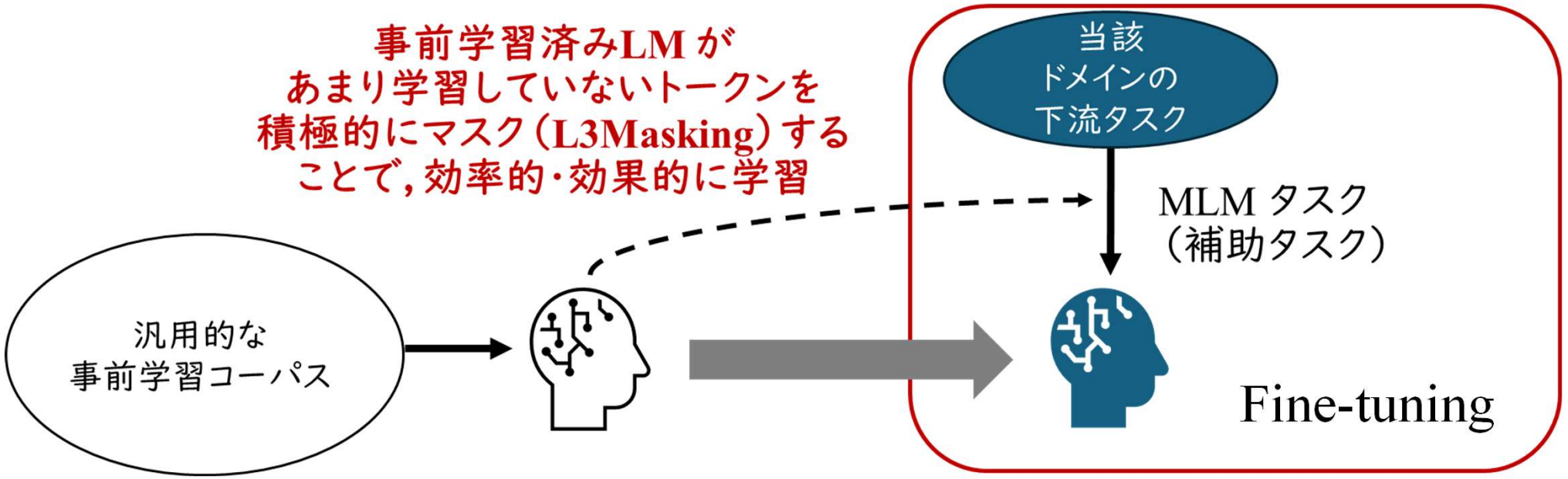
タスクのファインチューニングにおけるマスキング戦略

事後学習においては，事前学習済み言語モデルを考慮したマスキング戦略を考案しないと，非効率・非効果的な可能性有



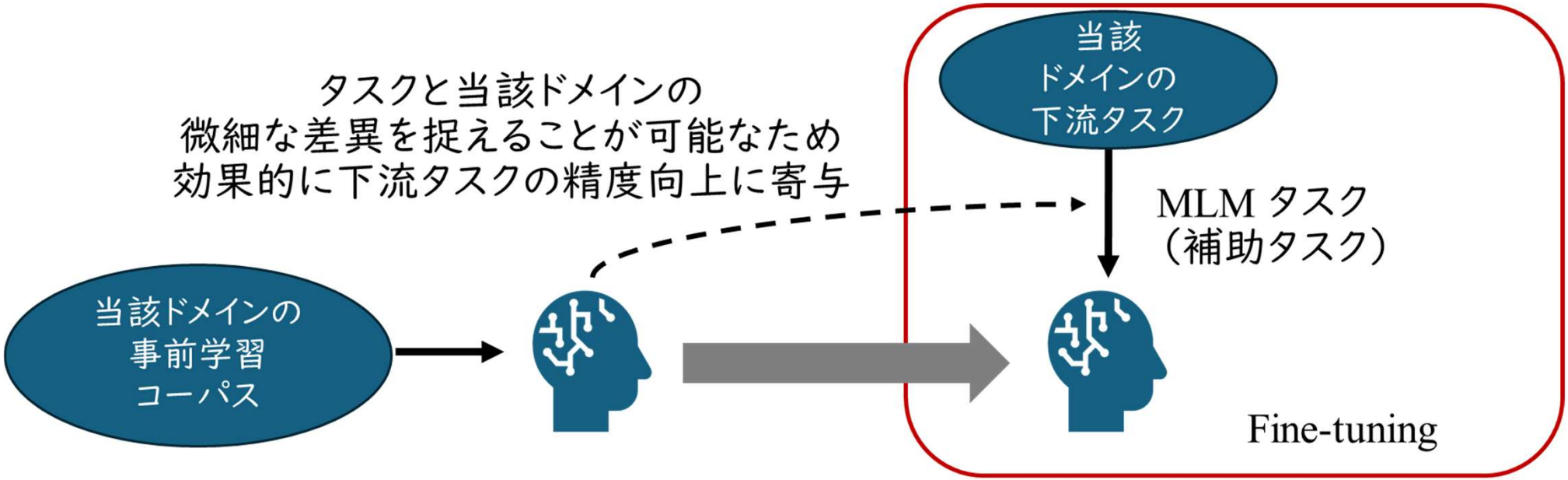
L3Masking による効率的な学習

Multi-task Fine-tuning for Language Models
by **L**everaging **L**essons **L**earned from Vanilla Models



L3Maskingによる効果的な学習

Multi-task Fine-tuning for Language Models
by **L**everaging **L**essons **L**earned from Vanilla Models



文の擬似対数尤度

ある文において，言語モデルが事前学習であまり学習していないトークンを発見するために，Encoder（双方向言語モデル）の文の擬似対数尤度（Pseudo-Log Likelihood; PLL）の考えを導入

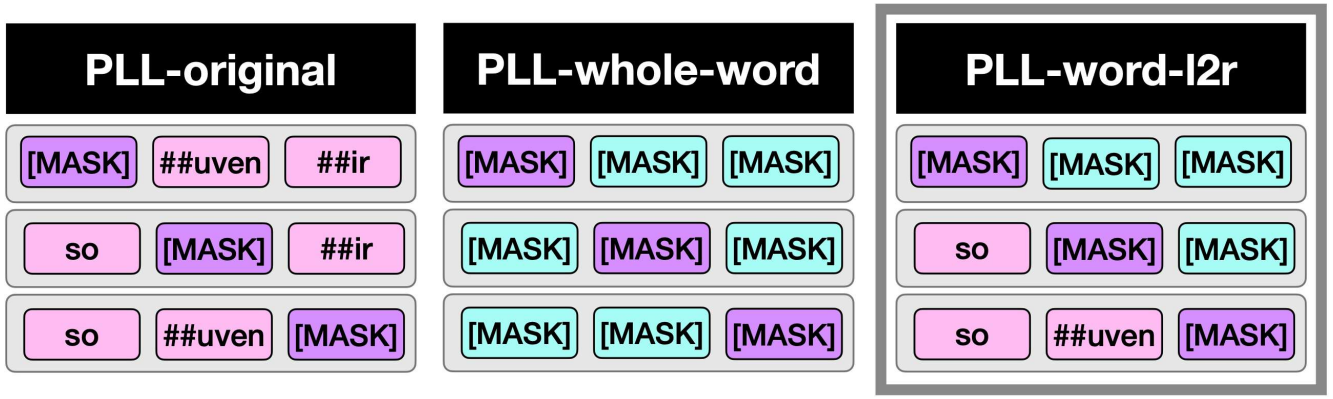
- 尤度が低い≒事前学習コーパスであまり出現していない

理由：
単方向言語モデルの尤度算出は以下の式で定義されるが，双方向言語モデルの場合，予測するトークンの前後が見えてしまうため，これまで定義されてこなかった

$$P(x) = \prod_{t=1}^{N} P(x_t | x_{<t})$$

PLL Score の定義

PLL Score には，複数の計算法が存在



- PLL-original [8]：
  各トークンをマスクした時のトークンの疑似対数尤度の総和
- PLL-whole-word [9]：
  単語単位でマスク．疑似対数尤度を算出したトークンのマスク箇所は外さない
- **PLL-word-l2r** [9]：
  単語単位でマスクして，単語内のトークンの疑似対数尤度を左から右に計算
  疑似対数尤度を算出したトークンのマスク箇所は外す（**最良**）

[8] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. ACL 2020, pp. 2699–2712.
[9] Carina Kauf and Anna Ivanova. A Better Way to Do Masked Language Model Scoring. ACL 2023, pp. 925–935.

トークン単位の擬似尤度に基づくマスク戦略

下流タスクのデータに対して，トークン単位の擬似尤度を計算し，擬似尤度が低いほど（事前学習であまり学習していないドメイン固有のトークン）多くマスクするマスク戦略を提案

1. トークン単位の擬似尤度（PL score of Token; PLT）
   - トークン単位の理由： 当該ドメインの言語知識に該当するトークンの発見の
   - log がない理由： トークン単位のマスク確率とするため

$$\mathrm{PLT}_{\mathrm{l2r}}(s_t \mid S) := P_{\mathrm{MLM}}\left(S_{\setminus t} \mid S_{\setminus s_{t' \geq t}}\right)$$

マスク確率の調整

2. 事前学習において，**著しく高い文内のトークンの平均マスク確率は下流タスクの性能を 低下させる** [10]
（例：15% ⇒ 80%）

事後学習においても平均マスク確率は下流タスクの精度に影響を及ぼす可能性有

文内のマスク確率の平均値を変化させる関数f を導入

$$\mathrm{PLT}_{\mathrm{l2r}}(s_t \mid S) := f\left(1 - P_{\mathrm{MLM}}\left(S_{\setminus t} \mid S_{\setminus s_{t' \geq t}}\right)\right)$$

関数f は，文内におけるマスク確率の平均値を特定の値に変更するために，各トークンのマスク確率が 0 以上 1 以下の制約を守りつつ，平均値が特定の値にす

[10] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023,Should You Mask 15% in Masked Language Modeling? EACL , pp. 2985–3000.

評価実験

主タスクのみを学習したモデル（Single Task Learning; STL），Random Token Masking（RTM）を補助タスクとしたモデル[11] との比較実験を行う

表：使用するデータセットについて

ドメイン	データセット	教師ラベルの種類	学習データの数	検証データの数	テストデータの数	クラス数
計算機科学	ACL-ARC	引用意図	1,688	114	139	6
医療	Ohsumed	病気カテゴリ	3,022	4,043	4,043	23
映画レビュー	IMDb	感情分析	25,000	2,500	22,500	2

[11] Lucio M. Dery, Paul Michel, Ameet Talwalkar, Graham Neubig. 2022. “Should We Be Pre-training? An Argument for End-task Aware Training as an Alternative”, ICLR, pp.1--18

結果（ACL-ARC：計算機科学）

Model	Framework	Masking	マイクロF1	マクロF1
BERT-base（汎用的な文書で事前学習）	STL	-	**71.34** ±0.35	63.07 ±0.69
	MTL	RTM	70.77 ±0.86	62.15 ±0.48
	MTL	L3Masking	71.31 ±0.98	**63.15** ±0.90
RoBERTa-base（汎用的な文書で事前学習）	STL	-	71.73 ±4.06	59.44 ±6.70
	MTL	RTM	78.94 ±1.76	70.30 ±2.20
	MTL	L3Masking	**79.12** ±1.60	**73.30** ±2.90
SciBERT（科学関連の文書で事前学習）	STL	-	80.36 ±2.45	71.84 ±2.73
	MTL	RTM	80.14 ±1.38	70.88 ±3.06
	MTL	L3Masking	**82.50** ±1.90	**74.10** ±2.40

全体的に，本研究の提案手法は高精度

当該ドメインで事前学習を行ったSciBERT に対して効果的

結果（Ohsumed）

Model	Framework	Masking	マイクロF1	マクロF1
BERT-base	STL	-	76.69 ±3.41	**68.76 ±3.47**
	MTL	RTM	**76.98 ±2.03**	67.47 ±2.40
	MTL	L3Masking	76.81 ±1.49	66.1 ±3.50
RoBERTa-base	STL	-	70.07 ±0.54	60.92 ±0.91
	MTL	RTM	69.92 ±0.64	64.83 ±0.37
	MTL	L3Masking	**73.38 ±0.48**	**65.02 ±0.61**
ClinicalBERT（医療関係の文書で事前学習）	STL	-	71.02 ±0.42	62.85 ±0.63
	MTL	RTM	70.75 ±0.36	62.7 ±0.61
	MTL	L3Masking	**71.66 ±0.78**	**63.7 ±0.60**

全体的に，本研究の提案手法は高精度
当該ドメインで事前学習を行ったClinicalBERT に対して効果的

結果（IMDb）

Model	Framework	Masking	Accuracy	F1
BERT-base	STL	-	88.05 ±0.05	87.15 ±0.56
	MTL	RTM	88.05 ±0.05	88.19 ±0.08
	MTL	**L3Masking**	**88.10 ±0.21**	**88.08 ±0.08**
RoBERTa-base	STL	-	88.84 ±0.32	88.89 ±0.30
	MTL	RTM	91.29 ±0.27	**91.30 ±0.22**
	MTL	L3Masking	**91.32 ±0.15**	91.13 ±0.09

二値分類のデータには，あまり効果なし

マスク確率の分析

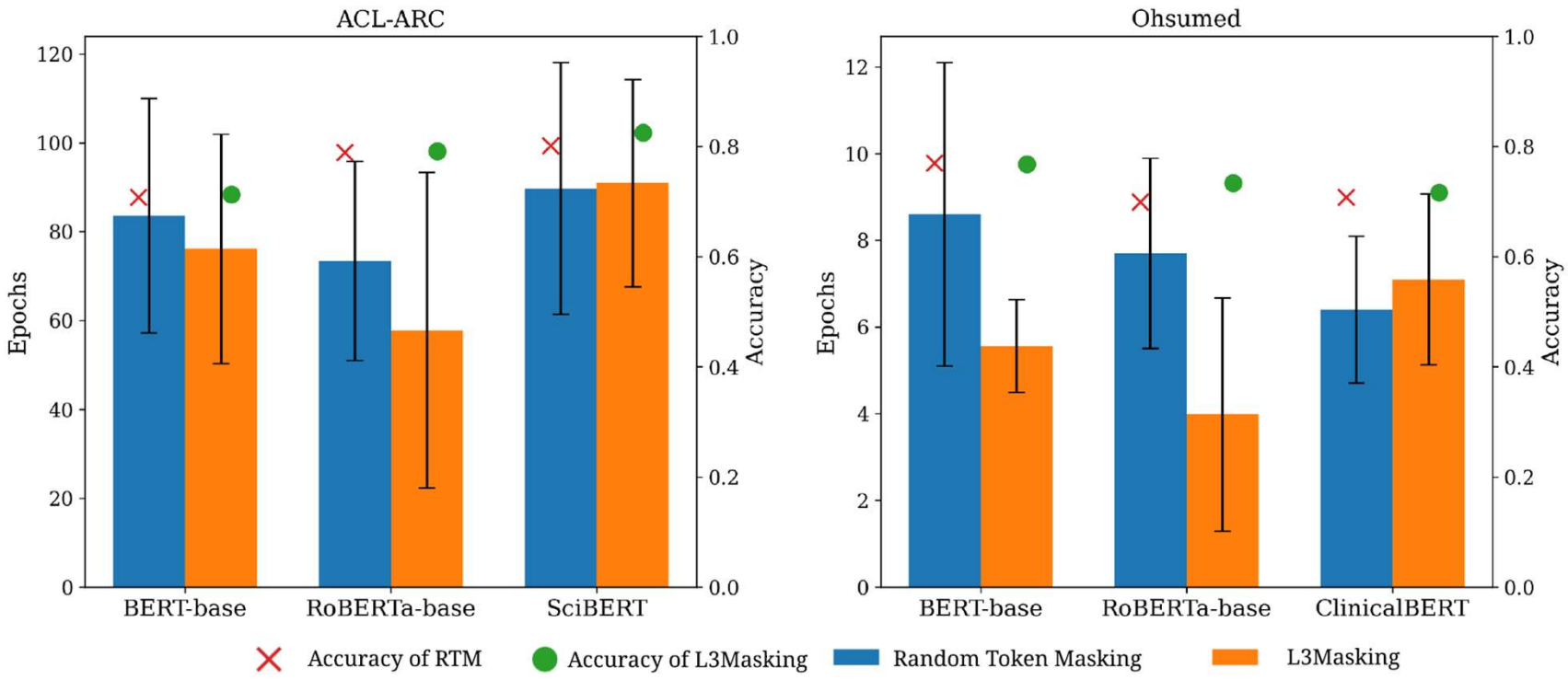
ACL-ARC
（マスク確率が15 % より高いほど赤色, 低いほど青色）

Thus , over the past few years , along with advances in the use of learning and statistical methods for acquisition of full parsers (Collins , 1997 ; Charniak , 1997a ; Charniak , 1997b ; Ratnaparkhi , 1997) , significant progress has been made on the use of statistical learning methods to recognize shallow parsing patterns syntactic phrases or words that participate in a syntactic relationship (Church , 1988 ; Ramshaw and Marcus , 1995 ; Argamon et al . , 1998 ; Cardie and Pierce , 1998 ; Munoz et al . , 1999 ; Punyakanok and Roth , 2001 ; Buchholz et al . , 1999 ; Tjong Kim Sang and Buchholz , 2000) .

本研究の提案手法の目論見通り, 計算機科学ドメイン固有と考えられる用語parser や人物名に高いマスク確率が付与

効率性の分析

本研究の提案手法は，汎用的な事前学習コーパスで学習されたBERTやRoBERTa に対して，効率的な学習を可能にした



本論文のまとめ

本論文ではマルチタスク学習における
補助タスクを低コストに設計するために
ラベリング方法の違いによって二つの手法を提案

- サブワードベースのトークンN-gram 認識タスクの提案
- 擬似尤度に基づくマスク戦略を用いたMLM タスクのためのマスク戦略を提案

文書分類タスクに対して，効果的な学習を可能にした