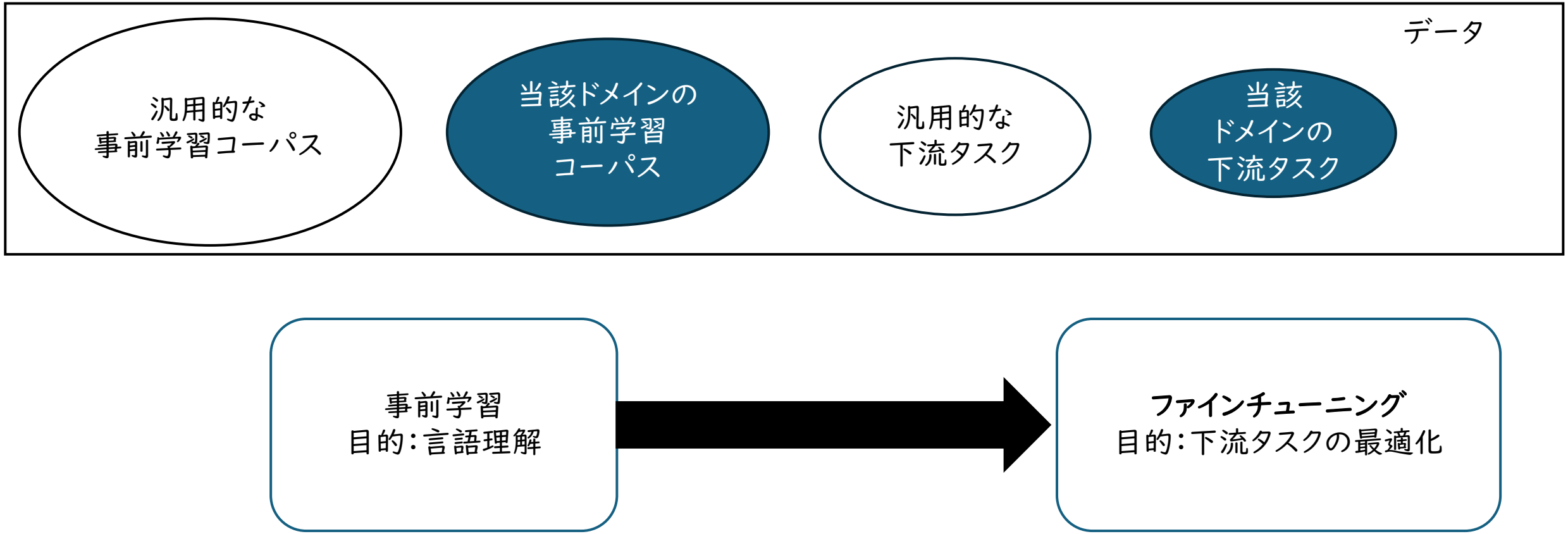
ドメイン適応のためのトークン単位の擬似尤度に基づくマスク戦略

木村 優介, 同志社大学大学院文化情報学研究科
✉ usk@acm.org
駒水 孝裕, 名古屋大学数理・データ科学教育研究センター
波多野 賢治, 同志社大学文化情報学部

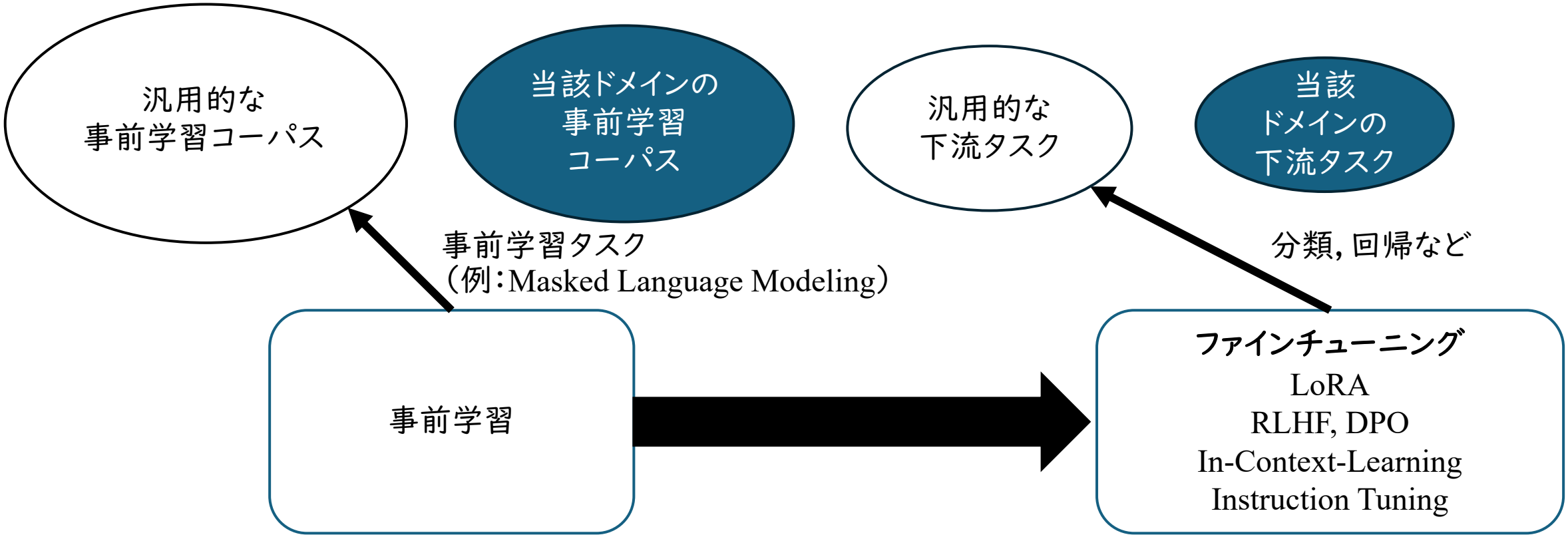
言語モデル

事前学習-ファインチューニングの学習パラダイム



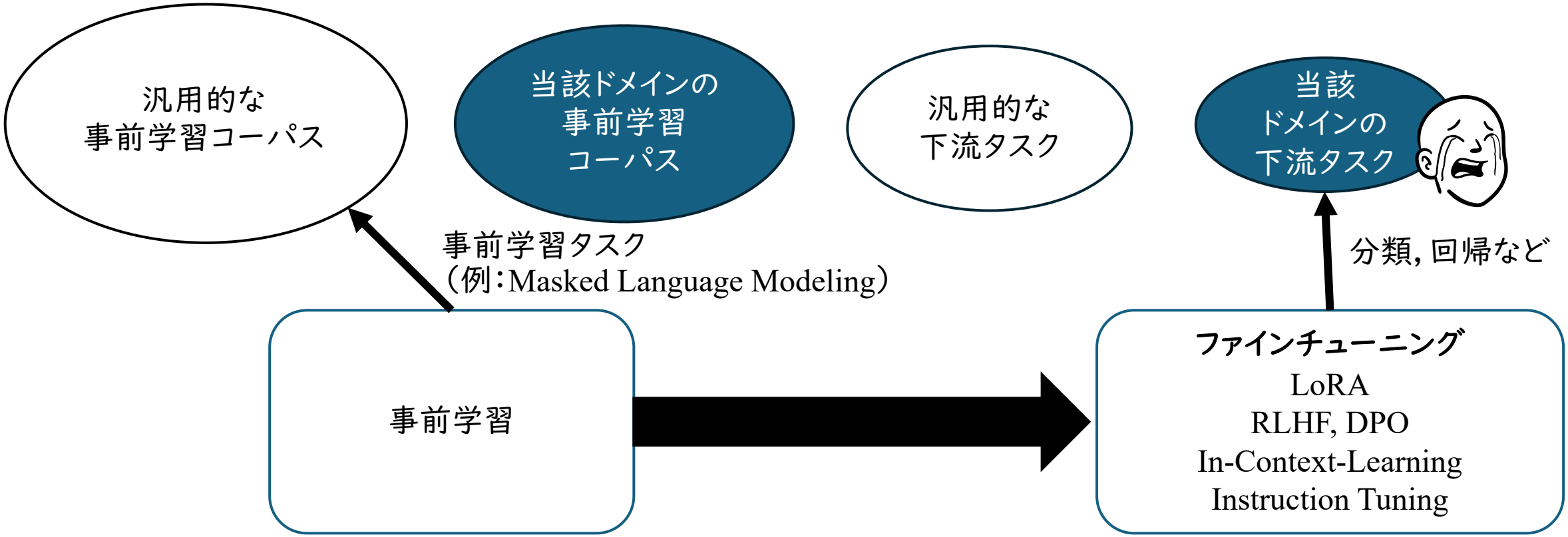
汎用的な言語モデル

一般的な言語モデル（例：bert-base）



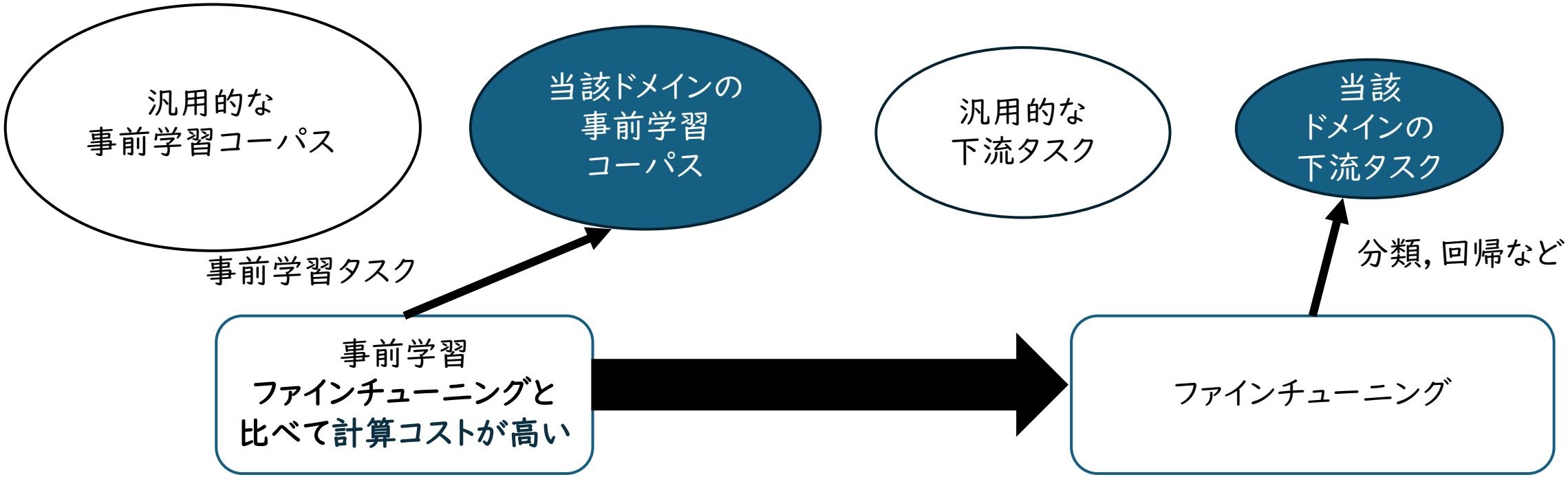
一般的な学習パラダイムが持つ問題

事前学習コーパスに少量しか含まれないドメインの下流タスクに
汎用的な言語モデルを適用すると下流タスクの性能は低下（**ドメインシフト**）



当該ドメインで事前学習

事前学習を当該ドメインで行う（SciBERT[1], BioGPT[2] など）

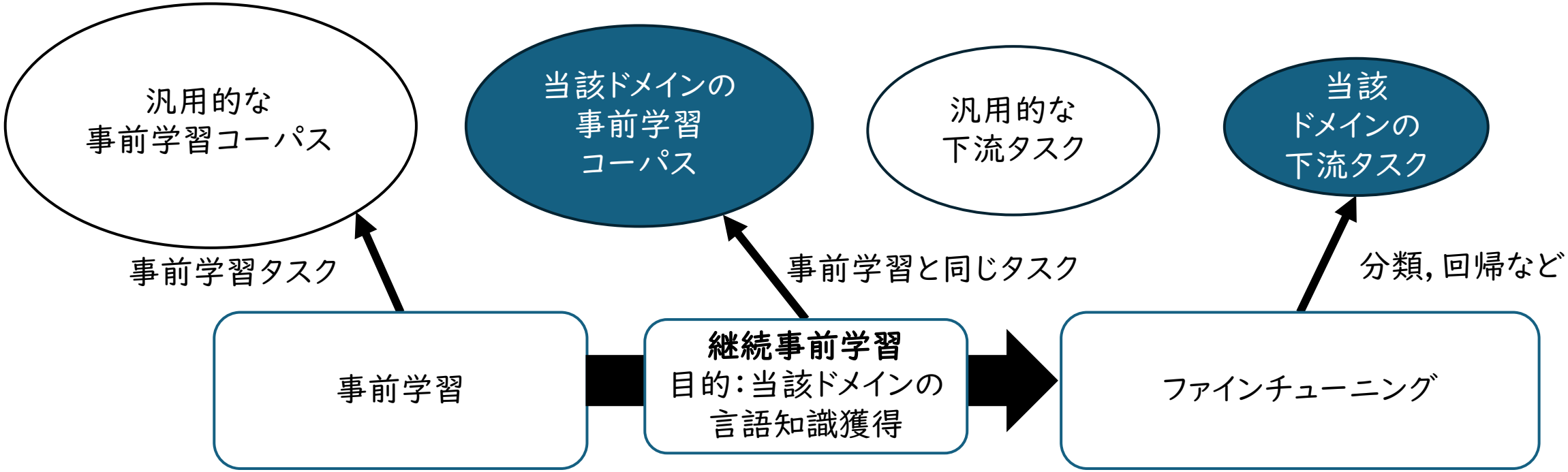


[1] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. EMNLP 2019.
[2] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, Vol. 23, No. 6, 2022.

継続事前学習

事前学習済み言語モデルを用いて当該ドメインで引き続き事前学習を行う

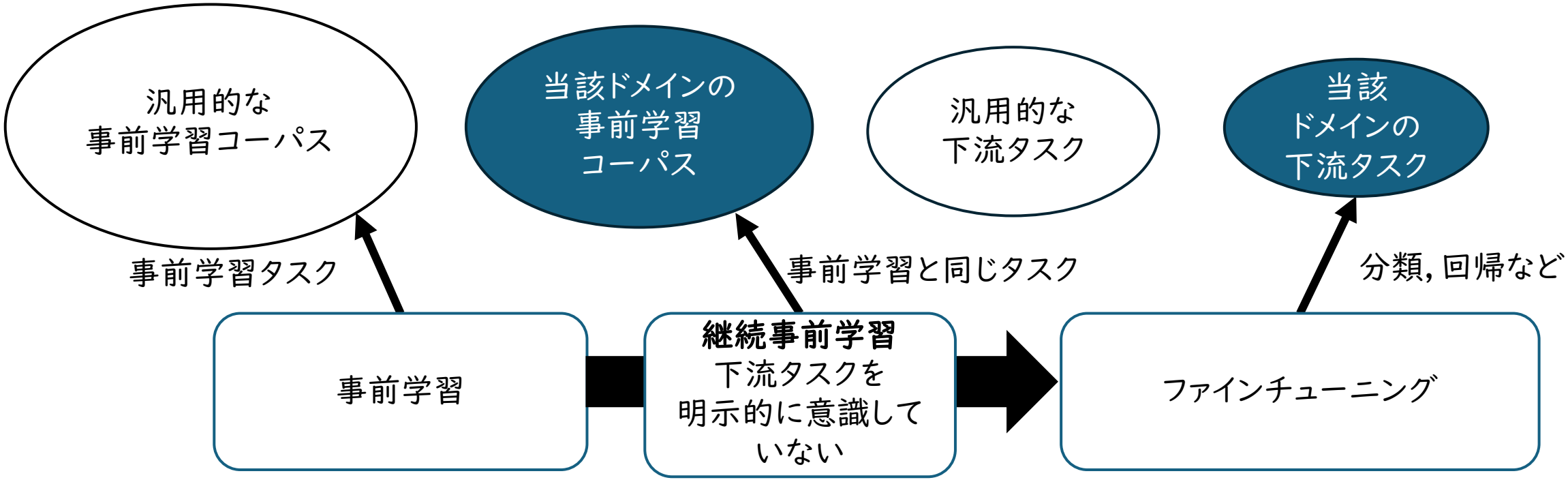
- TAPT・DAPT [3], K2 [4]



[3] Suchin Gururangan, Ana Marasovi´c, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don’t stop pretraining: Adapt language models to domains and tasks. ACL 2020, pp. 8342–8360.
[4] Cheng Deng, Tianhang Zhang, Zhongmou He, Yi Xu, Qiyuan Chen, Yuanyuan Shi, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, Junxian He. K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization, 2023, arXiv: 2306.05064

継続事前学習の問題点

継続事前学習で**下流タスクの性能を向上させる**ためには，
事前学習タスクや学習のEarly Stop などの**実験設定に注意を向ける必要がある** [3]

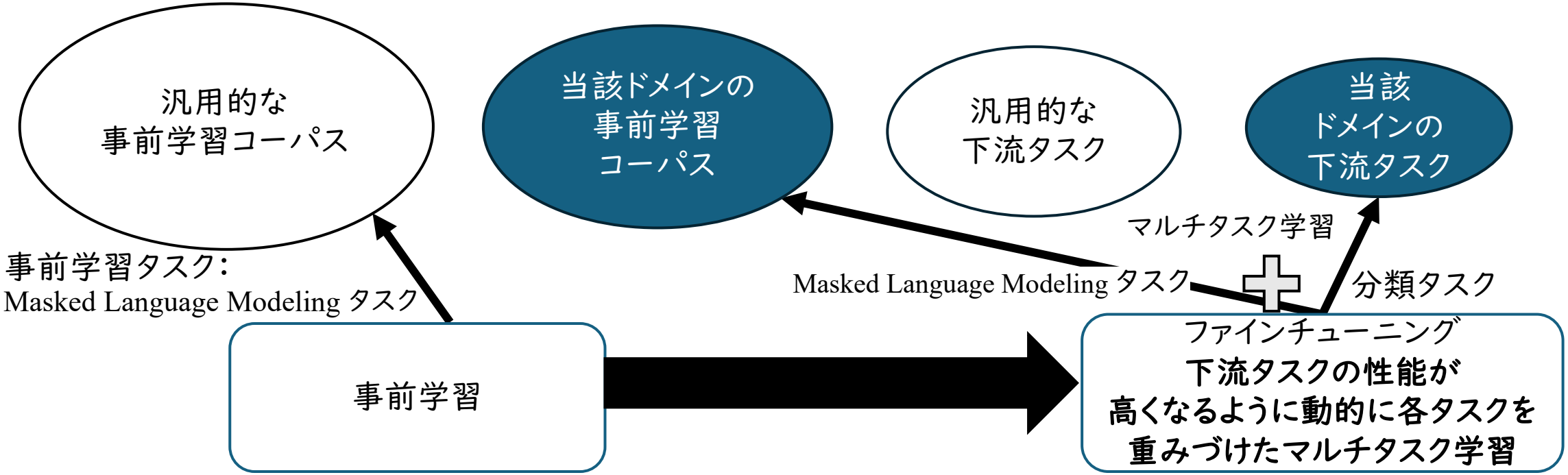


下流タスクを意識したドメインシフトの対策法が必要

[3] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don’t stop pretraining: Adapt language models to domains and tasks. ACL 2020, pp. 8342–8360.

ファインチューニング時のドメイン適応

- META-TARTAN [5], AANG [6]（文書分類に強い**双方向言語モデルを用いた実験で，継続事前学習よりも効果的に下流タスクの性能向上に繋がったと報告**）



[5] Lucio M. Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig. Should We Be Pre-training? An Argument for End-task Aware Training as an Alternative. ICLR 2022,.
[6] Lucio M. Dery, Paul Michel, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. AANG: Automating auxiliary learning, ICLR 2023

Masked Language Modeling (MLM)

双方向言語モデルのための事前学習タスクで，文の一部をマスクするタスク

従来のマスク戦略：

- **Random Token Masking (META-TARTAN, AANG で使用) [7]**
  同じ確率でランダムにトークンをマスク（15% の確率でトークンを[MASK] トークンに置換，
  その内10% をマスクではなく，ランダムなトークンに置換し，
  他10% は元のトークンに置換する）
- Knowledge Masking [8]
  人手で見つけた言語知識に該当する部分を多くマスクするマスク戦略
- PMI Masking [9]
  機械的に見つけた言語知識に該当する部分を多くマスクするマスク戦略

事前学習で言語知識に該当する部分をマスクすると，下流タスクにおける知識問題の性能を向上させる

[7] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019, pp. 4171–4186.
[8] Lucio M. Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig. Should We Be Pre-training? An Argument for End-task Aware Training as an Alternative. ICLR 2022.
[9] Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. PMI-MASKING: PRINCIPLED MASKING OF CORRELATED SPANS. ICLR, 2020.

ドメイン適応における既存のマスク戦略の問題点

これまでのドメイン適応では，**事前学習と事前学習より後段でのMLM の設定は全く同じ**

ファインチューニングにおけるドメイン適応では，**事前学習ではあまり学習されていない言語知識の獲得に注意を向けるべき（事前学習と同じマスク戦略だと効果的ではない）**

理由：

- ファインチューニング時のドメイン適応は検証データにおける下流タスクの性能を最適化するため，継続事前学習のようにMLM が収束するまで学習を回す必要はない
（従来のマスク戦略は事前学習ですでに獲得した言語知識を再度学習する可能性があるため，冗長な可能性がある）

本研究では，ドメイン適応時に既に存在する事前学習済み言語モデルの存在を考慮して，**ドメインごとにトークン単位でマスク確率を変更する**ドメイン適応のためのマスク戦略を提案

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獲得していない言語知識を発見するために，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

- 単方向言語モデルと同じ計算法で文の対数尤度を定義することはできない

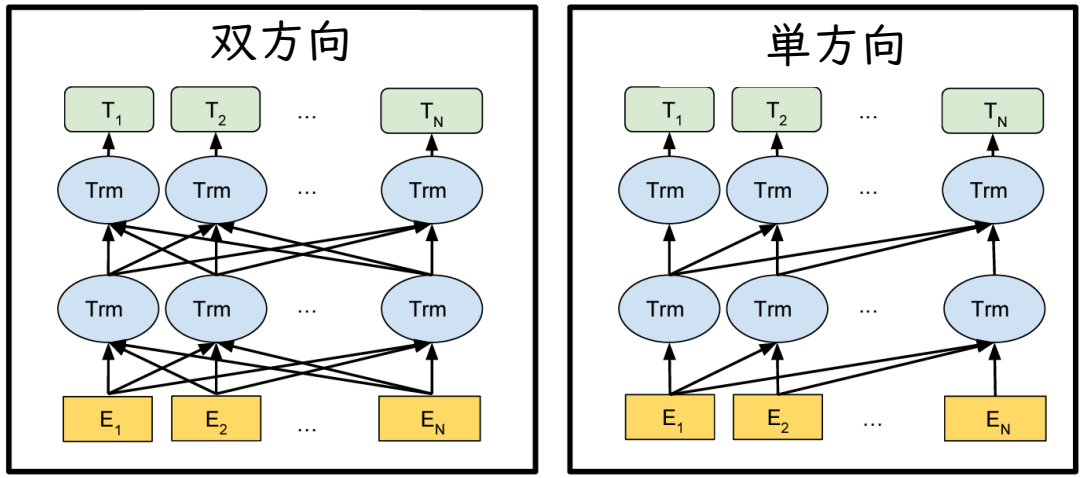


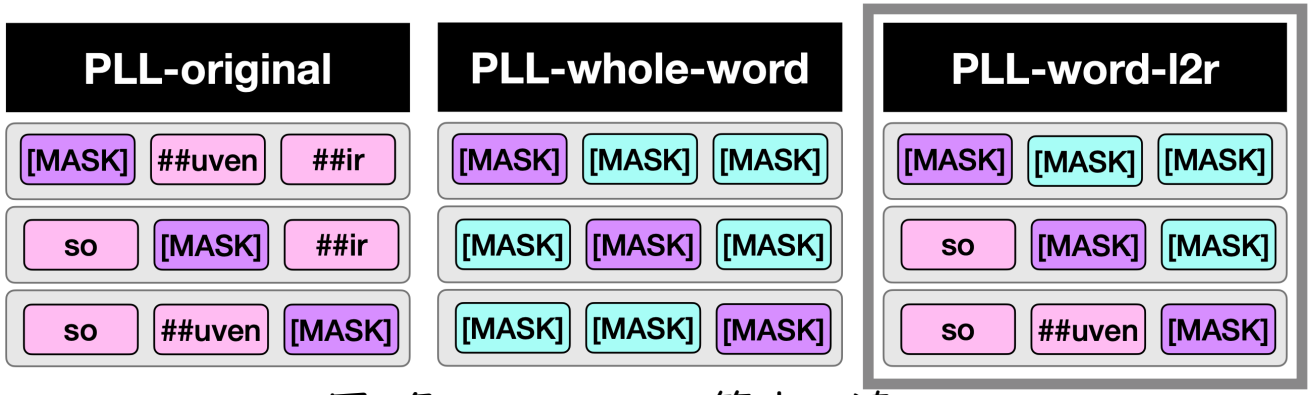
図: 双方向と単方向言語モデルの違い [7]

理由：
双方向言語モデルはあるトークン（言語モデルが扱う言語の最小単位）の予測の際に，単方向と異なり前後のトークンを考慮できるため

[7] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019, pp. 4171–4186.

PLL Score の定義

PLL Score には，複数の計算法が存在



図：各PLL Score の算出の違い [11]
（紫色：予測する対象，水色：予測対象以外のマスク，ピンク：予測対象以外のトークン）

- PLL-original [10]：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 [11]：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** [11]：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**）

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

PLL-word-l2r

PLL-word-l2r [11] の算出式

$$\mathrm{PLL}_{\mathrm{l2r}}(S) := \sum_{w=1}^{|S|} \sum_{t=1}^{|w|} \log P_{\mathrm{MLM}}\left(S_{\backslash t} \mid S_{\backslash w_{t' \geq t}}\right)$$

- n 個のトークンで構成される 文S
- 疑似対数尤度を算出する t 番目のトークンをマスクした際の表記：

$$S_{\backslash \mathrm{t}} := (s_1, \cdots, s_{t-1}, [\mathrm{MASK}], s_{t+1}, \cdots, s_n)$$

- 文S 内のw 番目の単語の末尾のトークンt' までマスクした上で算出する

[10] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. ACL 2020, pp. 2699–2712.
[11] 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

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）を定義
   - 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ため
   - log がない理由： トークン単位のマスク確率とするため

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

トークン単位の擬似尤度

事前学習であまり学習していないトークンを多くマスクするマスク戦略

1. トークン単位の擬似尤度を定義

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

2. 下流タスクのデータで**擬似尤度が低い（≒事前学習であまり学習していない言語知識）**トークンのマスク確率が高くなるように 1からPLT スコアを引く

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

マスク確率の調整

3. 事前学習において，**著しく高いマスク確率は下流タスクの性能を低下させる** [12]
   PLT スコアの値をマスク確率としてそのまま用いると，文内のマスク確率の平均値を調整できない

そのため，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なるようにする

1. 現在の平均値を指定した値にするために，文内の各トークンのマスク確率に対して
   同じ値を掛ける
2. 各トークンのマスク確率が1 を超えた場合，超えた分だけ他トークンのマスク確率に
   等しく分配する処理を1 を超えるトークンがなくなるまで繰り返す

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

学習設定

評価実験では, ファインチューニング時にドメイン適応を行う手法 META-TARTAN [5] の枠組みで **Random Token Masking をベースラインとした際の比較実験を行う**

- ベースラインと本研究の提案手法ともに 次のMETA-TARTAN の実験設定に倣う

項目	値
言語モデル	roberta-base
学習率	0.0001
トークン長※	128
バッチサイズ※	64
ドロップアウト率	0.1
平均マスク確率	**0.15**
エポック数	150
EarlyStopping（エポック）	3
オプティマイザ	AdamW

※著者の環境（V100 (32GB) × 4）に合わせて, トークン長とバッチサイズに関してはMETA-TARTAN の元論文と異なる

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

データセット

基本統計量（META-TARTANに倣う）

ドメイン	タスク	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	評価指標
生物医学	CHEMPROT [13]	relation classification	4,169	2,427	3,469	13	Acc
計算機科学	SCIERC [14]	relation classification	3,219	455	974	7	F_1
計算機科学	ACL-ARC [15]	citation intent	1,688	114	139	6	F_1

relation classification: 記号で囲まれた二つの名詞がどのような関係にあるか
citation intent: その文がどのような引用意図を持つか

[13] Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. ChemProt-3.0: a global chemical biology diseases mapping. Database: The Journal of Biological Databases and Curation, Vol. 2016, , 2016.
[14] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. EMNLP 2018, pp. 3219–3232.
[15] David Jurgens, Srijan Kumar, Raine Hoover, Dan Mc Farland, and Dan Jurafsky. Measuring the Evolution of a Scientific Field through Citation Frames. TACL 2018, Vol. 6, pp. 391–406.

結果

3種類のデータセットにおける評価実験

- 異なるシード値(0~9)で 10回実験した平均値と標準偏差
- 黒太字は平均値の差の検定で有意差があったことを示す

	CHEMPROT	SCIERC	ACL-ARC
	Acc	F_1	F_1
ベースライン	0.840 ± 0.008	0.818 ± 0.009	0.703 ± 0.022
提案手法	0.841 ± 0.006	0.815 ± 0.009	**0.733 ± 0.029**

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, 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

まとめ

- **ドメイン適応のために，事前学習済み言語モデルによる擬似尤度を用いて**
  ドメインごとにトークン単位でマスク確率を変更する新しいマスク戦略を提案
- 本研究の提案手法は**ACL-ARC でRandom Token Masking よりも高性能**
  - ただし，他の実験として行った Relation Classification タスクでは事前学習のデータには
    ないタスク固有の記号に対して高いマスク確率が付与されたため，
    評価実験の対象として行うべきデータセットではなかった
- 今後は，少量のパラメータのみをチューニングした際の実験や
  他の文書分類データセットでの更なる実験を行う

本研究の提案手法の
コードはGithub で公開中