自己教師あり学習を用いた文書分類のためのマルチタスク学習フレームワーク

〇木村 優介[1]

[1]チームMIL

自己紹介

名前： 木村優介

所属： 同志社大学大学院文化情報学研究科

研究対象： 文書分類

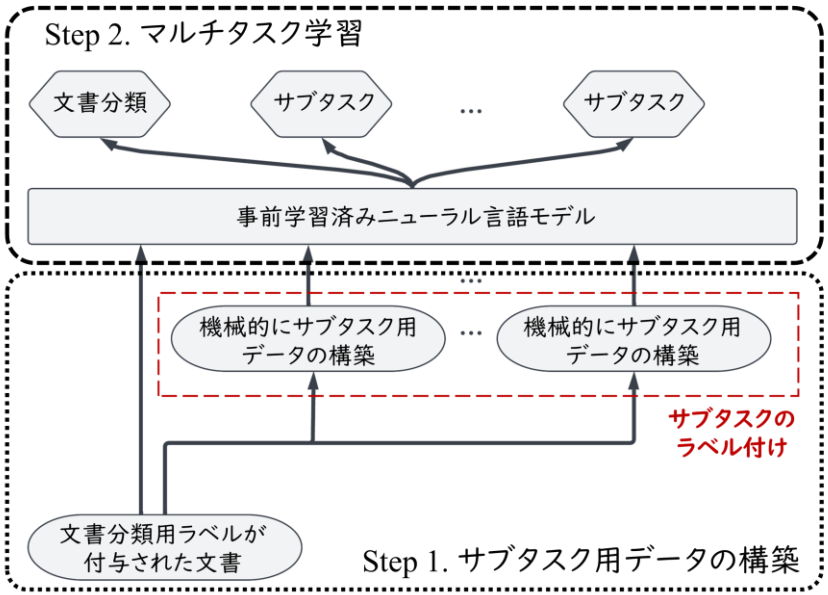
リーダーボードに掲載されている結果：
個人的なベースラインとして複数回実験した結果

最大系列長：256, バッチサイズ：128
（Bi-LSTM 層を複数追加・dropout 層を追加.
MSDは Multi Sample Dropout の略 ）

Rank	Team Name	Submitted on	Description	Micro-F1 (Public) ↓
1	MIL ←	2022/09/01	Roberta-StackedBiLSTM(layer=2, dropout = 0.25)	96.1285
2	Yusuke Kimura ←	2022/09/02	Roberta-StackedBiLSTM（layer=3, MSD=5, dropo=0.25）	95.9571

今日の話の結論

文書分類を目的とした提案手法は，英語で書かれた
有名なデータセット（不均衡データ）において高精度を達成

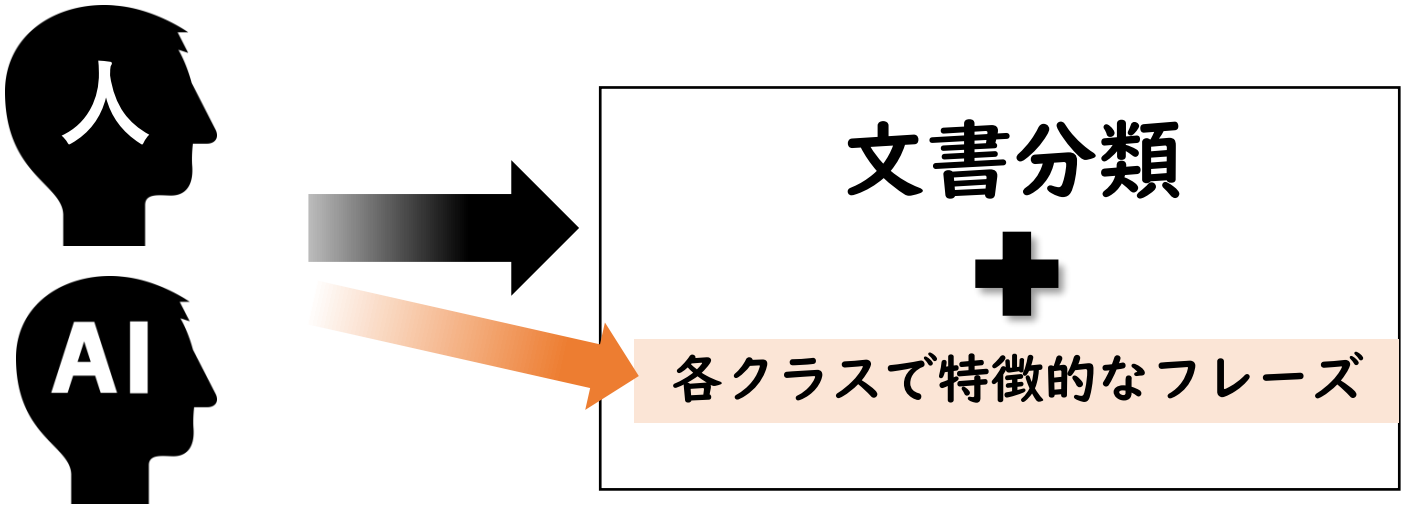


しかし，森羅プロジェクトにおいてはサブタスクがメインタスクに悪影響を及ぼし，提案手法をそのまま適用できなかった

⇒その原因と対策について説明

文書分類におけるフレーズの有用性

昔からフレーズを用いて文書分類を行うことは有用 [1,2]

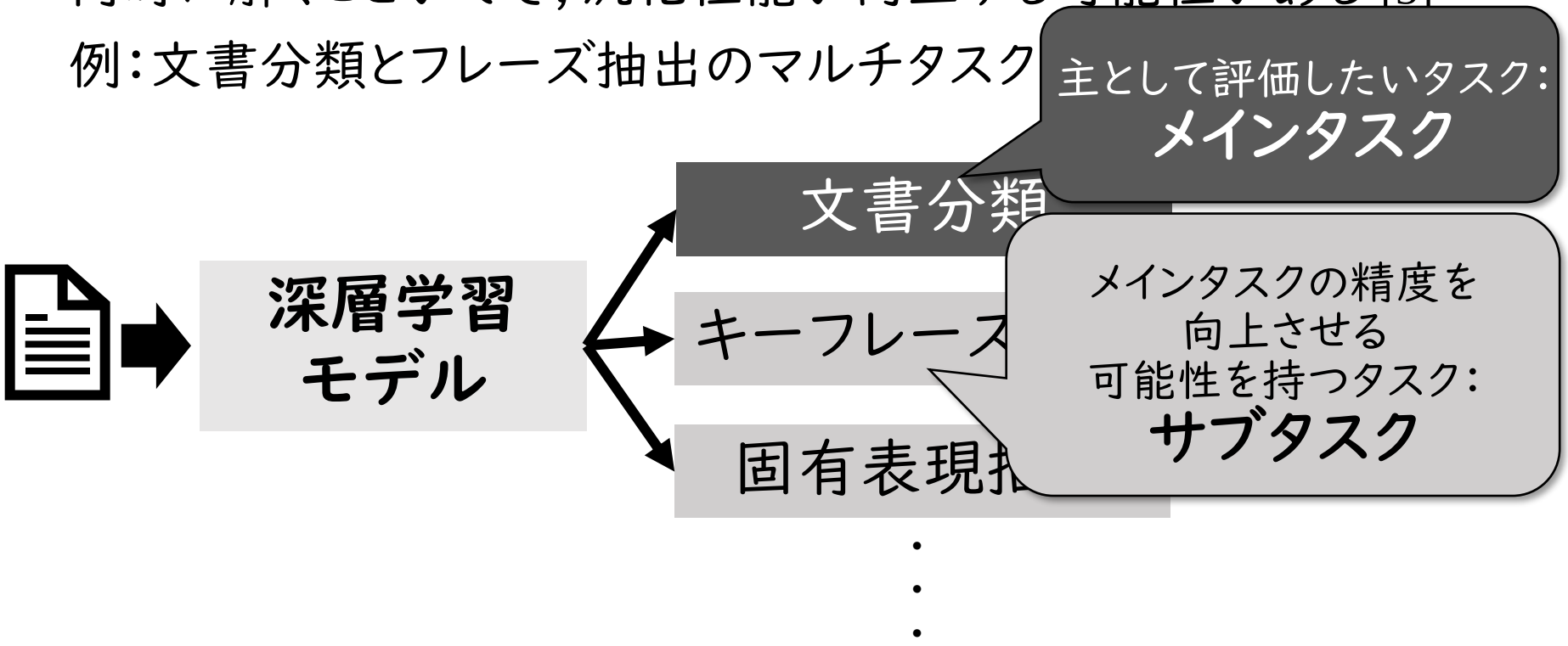


[1] Yongzheng Zhang, Nur Zincir-Heywood, and Evangelos Milios. 2005. Narrative text classification for automatic key phrase extraction in web document corpora. In Proceedings of the 7th annual ACM international workshop on Web information and data management (WIDM '05). Association for Computing Machinery, New York, NY, USA, 51–58. https://doi.org/10.1145/1097047.1097059
[2] Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34, 1 (March 2002), 1–47. https://doi.org/10.1145/505282.505283

マルチタスク学習

深層学習モデルは一つのモデルで複数のタスクを
同時に解くことができ，汎化性能が向上する可能性がある [3]

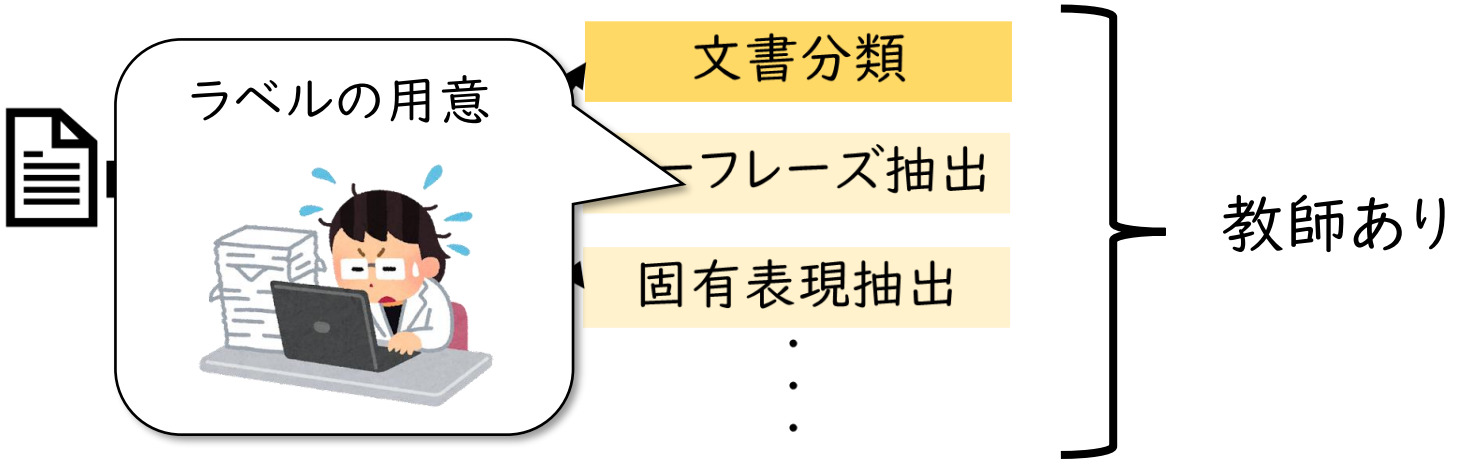
例：文書分類とフレーズ抽出のマルチタスク



[3] Rich Caruana. Multitask learning. Machine Learning, Vol. 28, No. 1, pp. 41–75, 1997

既存のサブタスクのコスト

従来のサブタスク[4]には**人手によるラベル付け**が必要なため，
金銭的・人的コストがかかる

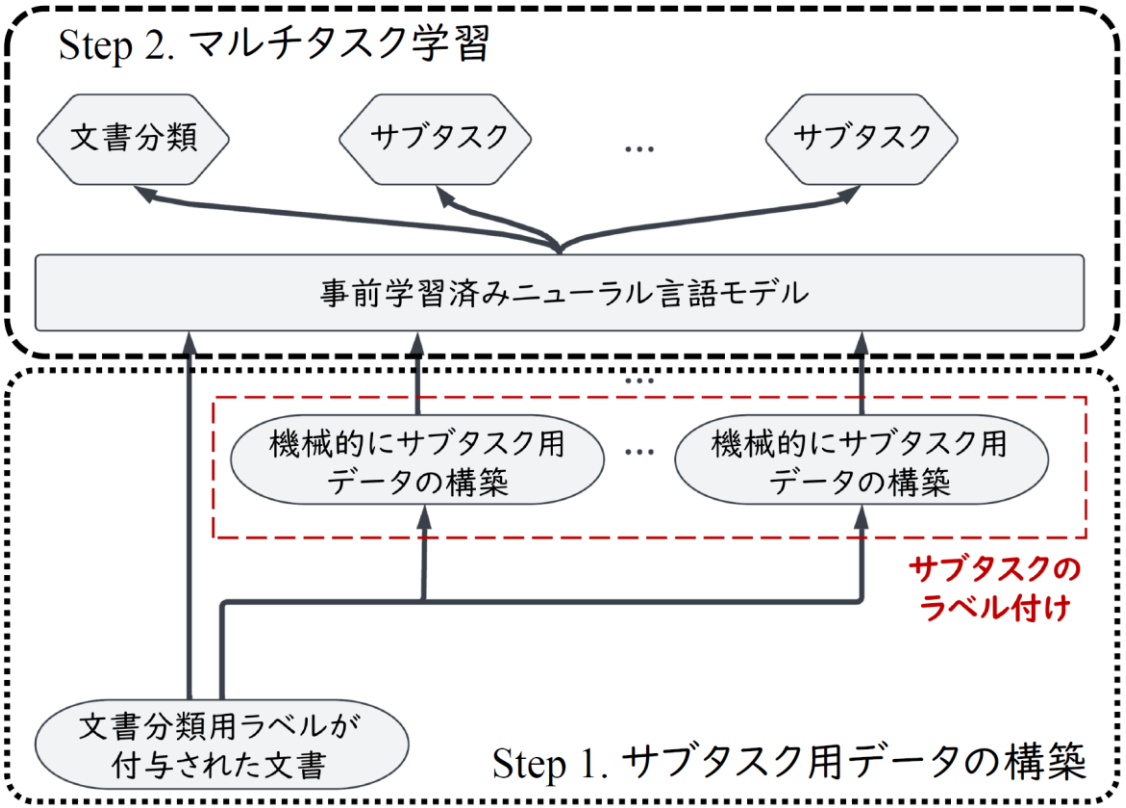


➡ 一般的に，サブタスクのラベルが付与されていない**文書分類用データセットにマルチタスク学習を適用することは難しい**

[4] 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.

文書分類フレームワーク

本研究では，人手によるラベル付けが必要ないサブタスク（自己教師あり学習）を用いた**メインタスク（文書分類）のための**
マルチタスク学習フレームワークを提案



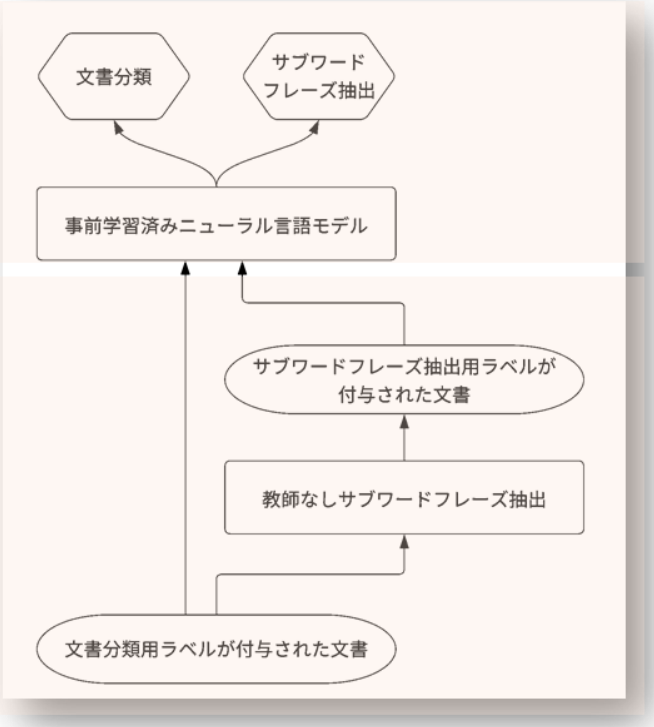
サブタスクの設計

マルチタスク学習において，難易度が高いタスクは**難易度が低いタスク**から情報を得て，容易に学習が進められる（**盗み聞き**）[5]

⇒ シンプルな**サブタスクとして，カテゴリ別で特徴的なサブワードフレーズを認識するタスクを用いる（頻度ベース）**

②マルチタスク学習：
サブワードフレーズ抽出と
文書分類

①入力
（教師なしサブワード
フレーズ抽出でラベリング）

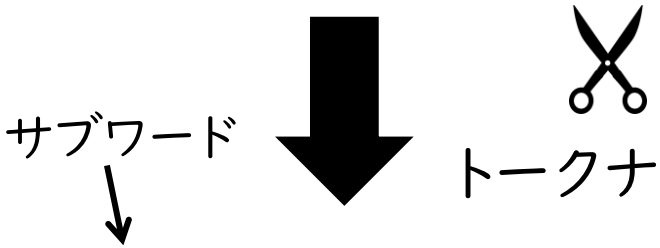


[4] Rich Caruana. Multitask learning. Machine Learning, Vol. 28, No. 1, pp. 41–75, 1997

サブワードフレーズ

トークナイザの分割結果に基づいて，機械的にフレーズのラベリングを行うために，サブワードのフレーズ（以後，**サブワードフレーズ**と呼称）を利用

例：Prices are lower in the country than in the city.

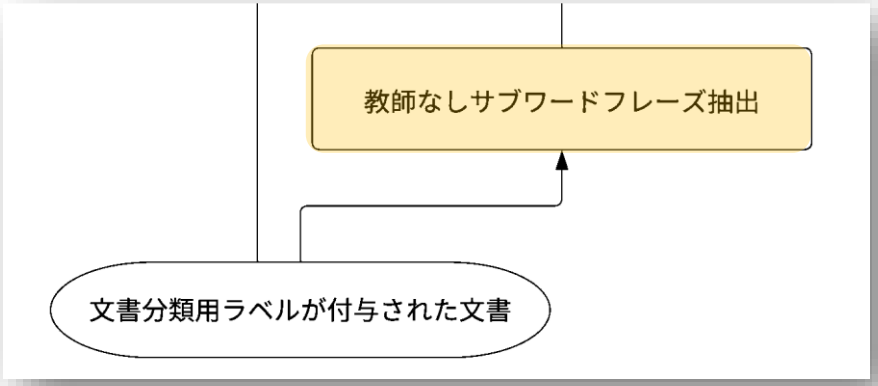


['_Price', 's', '_are', '_lower', '_in', '_the', '_country', '_than', '_in', '_the', '_city', '.']

サブワードフレーズ

サブワードフレーズ
※構成要素が単語でもサブワードフレーズと呼称することとする

高頻度なサブワードフレーズの作成（1 / 2）



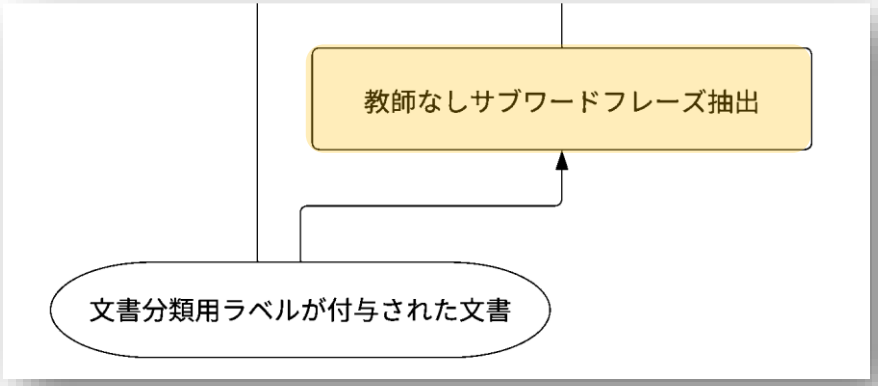
1. 事前学習済み言語モデルのトークナイザで文書を分割

例：Prices are lower in the country than in the city.

トークナイザ

['_Price', 's', '_are', '_lower', '_in', '_the', '_country', '_than', '_in', '_the', '_city', '.']

高頻度なサブワードフレーズの作成（2 / 2）



2. 高頻度なサブワードフレーズをクラス別にN 語抽出するために，**特定の語彙サイズになるまで頻度が高い文字同士を結合する Bite Pair Encoding(BPE) [6]をサブワード単位に変更する**

例：サブワードフレーズを 2 個得る場合（共起頻度が高いサブワードから結合）

['_Price', 's', '_are', '_lower', '_in', '_the', '_country', '_than', '_in', '_the', '_city', '.']

↓

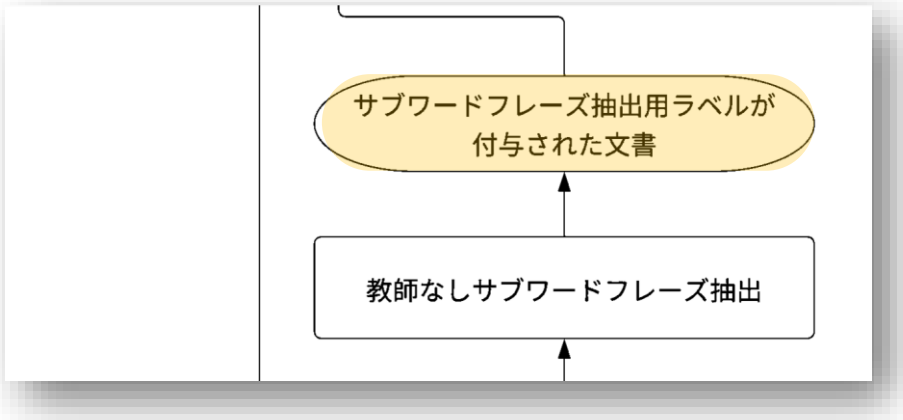
['**_Prices**', '_are', '_lower', '_in', '_the', '_country', '_than', '_in', '_the', '_city', '.']

↓

['**_Prices**', '_are', '**_lower_in**', '_the', '_country', '_than', '_in', '_the', '_city', '.']

[6] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

サブワードフレーズのラベリング



3. 教師なしサブワードフレーズ抽出手法で得られたサブワードフレーズを固有表現抽出で使われるIOB2タグ[7] でラベリング

例：_Prices, _lower_in がサブワードフレーズの場合

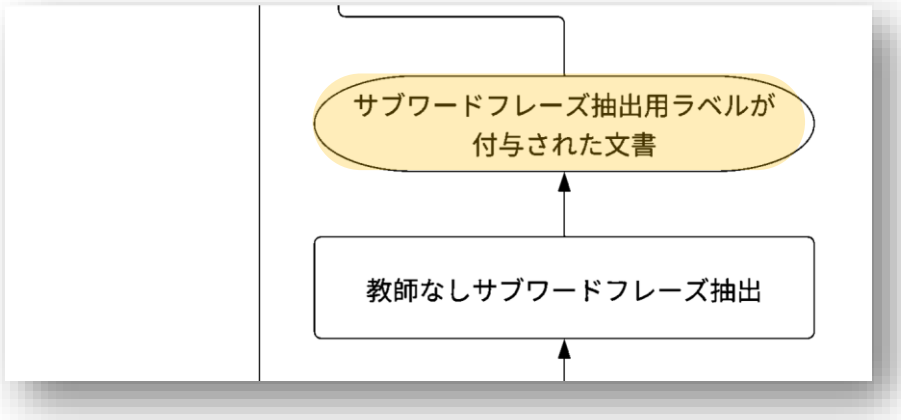
['_Price', 's', '_are', '_lower', '_in', '_the', '_country', '_than', '_in', '_the', '_city', '.']

['B', 'I', 'O', 'B', 'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O']

‘B’: サブワードフレーズの頭に位置するトークン,
‘I’: サブワードフレーズの2番目以降に位置するトークン,
‘O’: サブワードフレーズに含まれないトークン

[7] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In Third Workshop on Very Large Corpora, 1995.

高頻度なサブワードフレーズのラベリング



4. クラスごとに文書の意味を表さないサブワードフレーズを表現するために B, I タグにクラスの番号（クラス名を一意の番号に変換）を付与

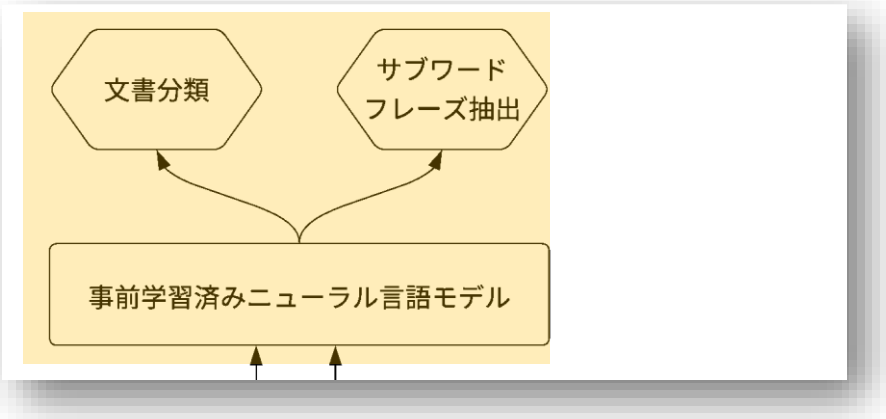
例：_Prices, _lower_in がクラス1で特徴的なサブワードフレーズの場合

['_Price', 's', '_are', '_lower', '_in', '_the', '_country', '_than', '_in', '_the', '_city', '.']

['B-1', 'I-1', 'O', 'B-1', 'I-1', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

※2クラス以上に共通するサブワードフレーズには，B-0, I-0 タグを付与

マルチタスク学習



4. 文書分類とサブワードフレーズ抽出のマルチタスク学習

文書分類:
$$P_r(c \mid X_{cls}) = \mathrm{softmax}(W_{dc}^{\top} \cdot \mathbf{x}_{cls})$$

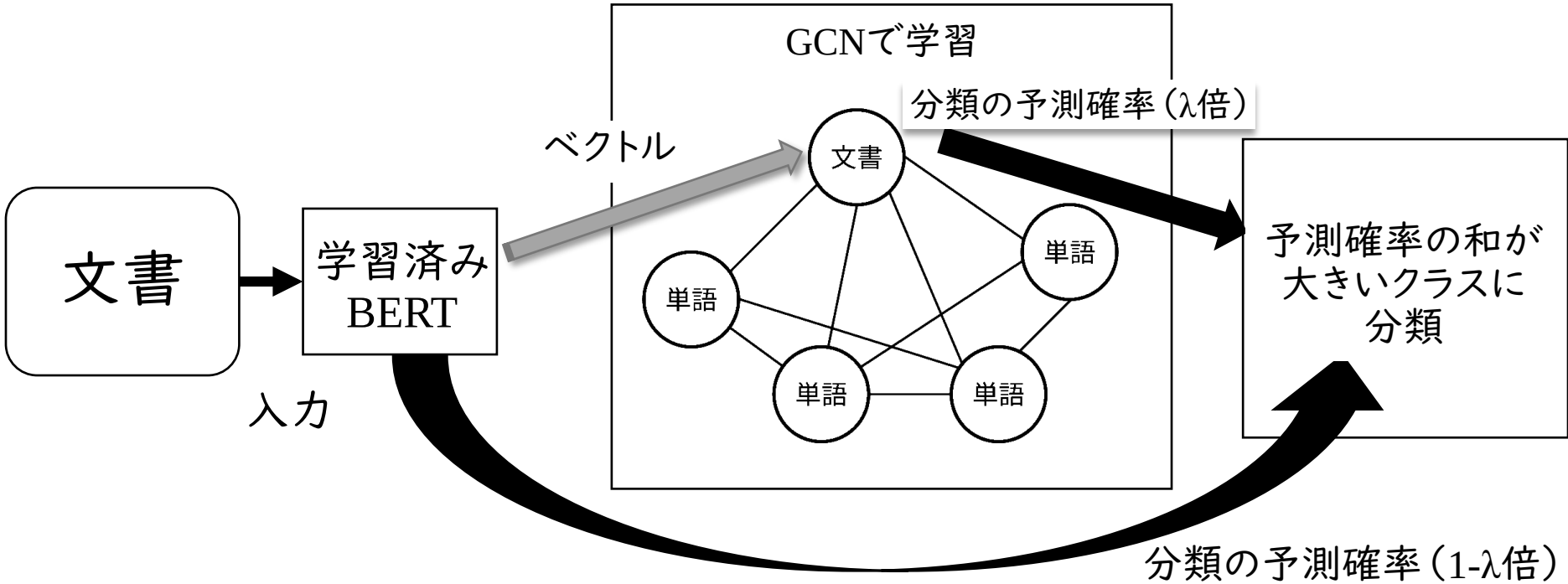
サブワードフレーズ抽出:
$$P_r(c \mid X_{token}) = \mathrm{softmax}(W_{sb}^{\top} \cdot \mathbf{x}_{token})$$

各ラベルへの分類と考え, Binary Cross Entropy Loss で得られた双方のタスクの Loss を足した値で学習を行う(今後改良)

$$L_{joint} = L_{dc} + L_{sb}$$

BertGCN

本研究の比較相手として，文書分類の精度が
高いBertGCN（BERTとGCNを組み合わせた手法）[7]と比較



[7] Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. BertGCN: Transductive text classification by combining GNN and BERT. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1456–1462, Online, August 2021. Association for Computational Linguistics.

評価方法

シングルタスクと比べて分類精度が向上したかを確認するために，
5種類のデータセットを利用

- 20 News Group (20NG)：ニュース文書
- R8, R52：ロイター
- Ohsumed (OHS)：医療文書
- Movie Review (MR)：映画レビュー

他データセットと比べると，**不均衡**なデータ
また，**R52はクラス数が多い**

	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

結果

マクロ, マイクロF 値で評価

F_{macro}

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 w/ cmn	0.860	0.845	0.947	0.841	0.610
Proposed w/o cmn	0.866	0.845	0.955	0.851	0.637

F_{micro}

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 w/ cmn	0.860	0.850	0.978	0.967	0.704
Proposed w/o cmn	0.866	0.851	**0.979**	0.969	0.711

表の説明：

Proposed w / cmn は
他クラスにも出現する
フレーズを考慮した手法

Proposed w / cmn は
他クラスにも出現する
フレーズを削除した手法

4種類のデータセットでRoBERTa-baseより高く,
クラス数が多くかつ**不均衡なデータでRoBERTaGCNより高精度**

考察

各クラスの上位10語の高頻度なサブワードフレーズを確認

	2 クラス以上で共通するサブワードフレーズ	クラス名 'bop'	...	'orange'
1	Ġre uter	Ġmoney Ġsupply		Ġconsumer Ġprices
2	Ġfe b	Ġmln Ġdlrs Ġbillion Ġdlrs		Ġconsumer Ġprice Ġindex
3	Ġl me	Ġweek Ġended		Ġstatistics Ġinstitute
4	Ġjan uary	Ġborrow ings		Ġrose Ġpct Ġfebruary
5	Ġfeb ruary	Ġbusiness Ġloans		Ġpct Ġmarch
6	Ġspokesman Ġsaid	Ġmoney Ġsupply Ġrises		Ġcost Ġliving
7	Ġp ct	Ġfed Ġsays		Ġpct Ġcompared
8	Ġm ln	Ġbillion Ġdlrs Ġbillion Ġdlrs		-
9	Ġseason ally	Ġweek Ġfed Ġsays		-
10	Ġseasonally Ġadjusted	Ġpct Ġbillion		-

各クラスに特徴的なフレーズが取得できている

Bop（国際収支）：money supply や business loan
orange（オレンジ）：consumer prices や consumer price index（消費者物価指数）

真のサブワードで構成されたフレーズが抽出されていないのは
高頻度なサブワードフレーズを対象としたため

まとめと今後の課題

- **従来のマルチタスク学習のサブタスクは人的・金銭的コストが高く**, 文書分類にマルチタスク学習を適用することは難しい

- 本研究では, メインタスクの精度向上のために**機械的にサブタスクのラベリングを行う汎用的なマルチタスク学習フレームワークを提案**

- 今後, 文書分類に有効な他の低コストなサブタスクを設計・利用する
  - 名詞句抽出
  - MLM など

森羅プロジェクトへの適用

そのままの適用は難しいことが分かった

提案手法をそのまま適用すると, サブワードフレーズ認識タスクの分類先が多すぎて（**343**種類＞文書分類クラス219種類）, サブタスクの精度が低い

⇒「盗み聞き」の性質（簡単なタスクが難しいタスクに良い影響を与える）がうまく働かなかった可能性あり.

対策：**盗み聞き**の性質を考え, サブタスクをより簡単にする

クラス固有にせず, トレーニングデータ全体で高頻度なサブワードフレーズのみ抽出対象とする [8]（B,I,Oの3種類）

[8]木村優介, 駒水孝裕, 波多野賢治. ストップフレーズ抽出を併用した文書分類. 第 14 回データ工学と情報マネジメントに関するフォーラム予稿集, 2022.