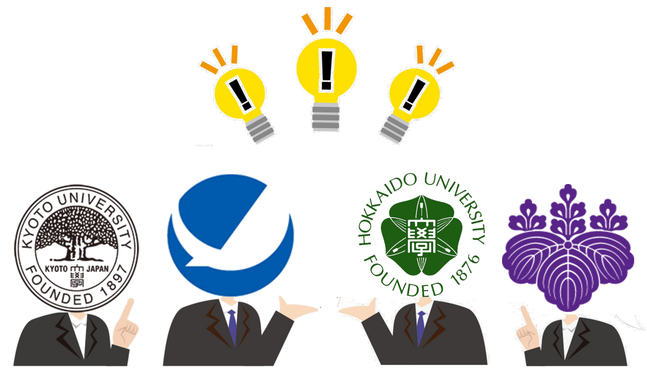
三言語モデル寄れば文殊の知恵

稲葉 達郎[1], **藤井 巧朗**[2], **小原 涼馬**[3], **柴田 幸輝**[4]
[1]京都大学., [2]横浜国立大学, [3]北海道大学, [4]筑波大学

研究目的・概要

- 「**三人寄れば文殊の知恵**」を言語モデルで再現する
  - 性能向上・推論効率化
- 性質・特徴の違う言語モデルを作成し，アンサンブル学習により互いの能力を最大限補完し合うフレームワークを探す



背景

そもそも，文殊の知恵とは…？

- 「凡人でも三人集まって相談すれば、すばらしい知恵が出るものだ」ということ [1]

これを言語モデルで再現するには…?

- 異なる知識・パラメータを持った言語モデルを複数作成し，それぞれの能力を最大限引き出すアンサンブル/マージを行えば良い
- 言語モデルの大規模化が進む今だからこそ，複数モデルに知識を分散させる方法を考えたい！！

どのように異なるモデルを作成する…？

- データセットが一つでメタデータがついていないものを今回は考える

関連研究

Mixture-of-Experts (MoE) [1]

- 複数のエキスパートモデルを作成し，親ゲートネットワークによりどのエキスパートモデルを使用するかを判別する

Branch-Train-Merge (BTM) [2]

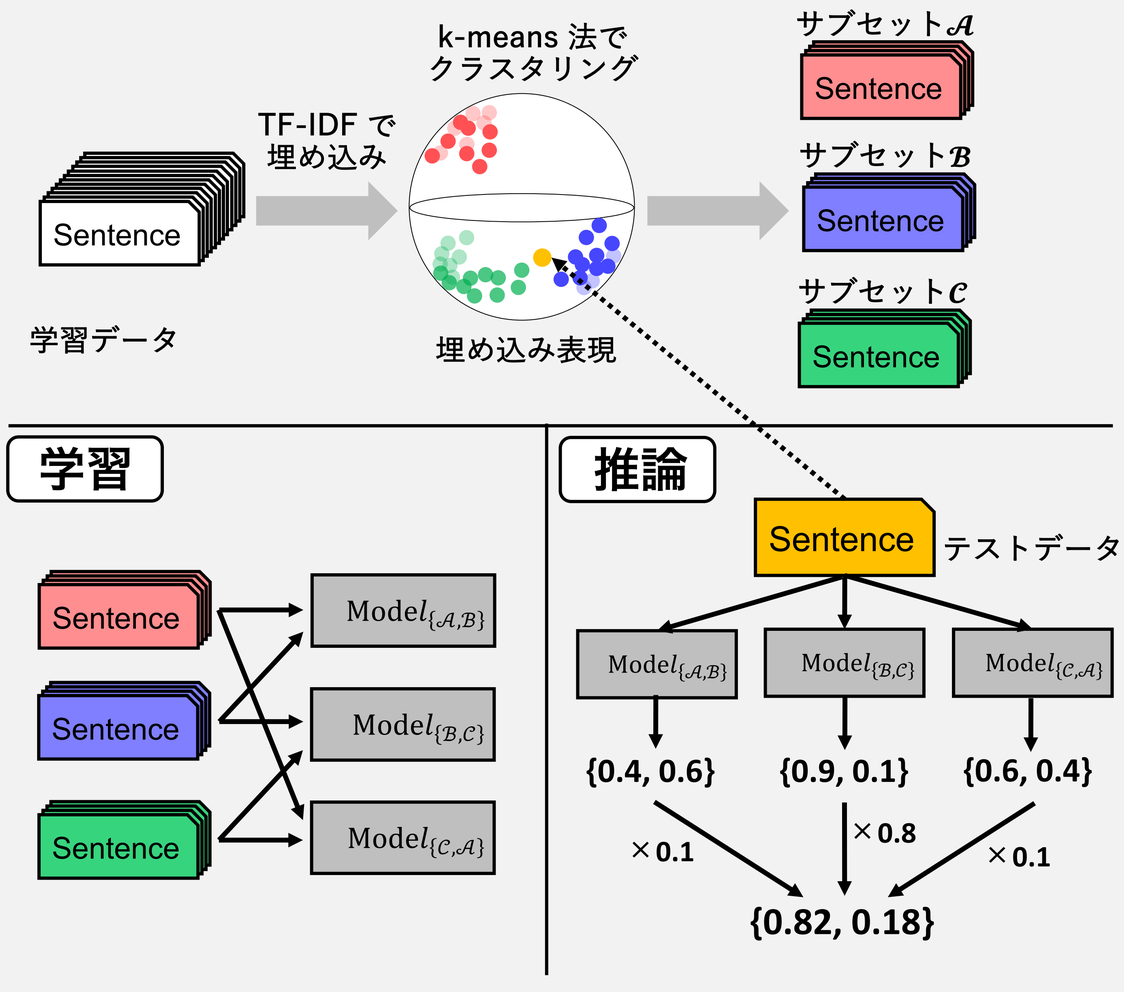
- k個のドメインコーパスでk個の Expert LM (ELM)を事前学習
- これらをマージしたものを初期値として再起的に学習

Cluster-BTM (c-BTM) [3]

- コーパスを k-means 法でクラスタリング
- クラスタ毎にELMを事前学習 → 入力埋め込みと各クラスタ中心からの距離でlogitを重み付ける

手法

- CBTM の考え方をもとに，データセットをクラスタリングし，事前学習済みモデルの Fine-tuning に利用する



- それぞれのモデルにサブセットを二種類学習させる
  → 境界部分の問題に対処
- テストデータも同様に埋め込み，得意なモデルの重みが大きくなるようにアンサンブル

実験・結果

実験

- モデル：RoBERTa-base（125M）
- データセット：SST-2 (感情分析タスク）
- ベースライン
  - 全てのデータで Fine-tuning したモデルのアンサンブル [2]
- 文殊モデル
  - k = 3 でクラスタリング
  - 重みの計算は以下の式で行う

$$\alpha_{AB} = \frac{d_C}{d_A + d_B + d_C}, \alpha_{BC} = \frac{d_A}{d_A + d_B + d_C}, \alpha_{CA} = \frac{d_B}{d_A + d_B + d_C}$$

（α_{ij} は Model{ij} の重み，d_i はクラスタ i の中心からのユークリッド距離）

結果

Training Data	Acc. (%)	Ensemble	Acc. (%)
$\{\mathcal{A},\mathcal{B},\mathcal{C}\}_{seed=0}$	93.23	Baseline	94.15
$\{\mathcal{A},\mathcal{B},\mathcal{C}\}_{seed=1}$	93.69		
$\{\mathcal{A},\mathcal{B},\mathcal{C}\}_{seed=2}$	93.35		
$\{\mathcal{A},\mathcal{B}\}$	93.12	Monju	94.04
$\{\mathcal{B},\mathcal{C}\}$	93.46		
$\{\mathcal{C},\mathcal{A}\}$	92.66		

- 少ない学習量でベースライン相当の性能を達成
- アンサンブルした際の性能の上がり幅はベースラインより大きい

今後の展望

“3人寄れば…”の3人をどうやって作るか問題

- 1つのデータセットから複数のモデルを作成
  - それぞれ異なるデータの学習順序を用いる
  - 埋め込み手法に SentenceBERT 等のニューラルモデルを用いる
  - GMM 等のクラスタリング手法を用いる
  - 各モデルに異なるアーキテクチャを用いる
- 複数データセットから複数のモデルを作成
  - 転移学習を利用

他にも…

- デコーダモデルで文殊の知恵を創出
- アンサンブル以外にもマージ・蒸留等

[1] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. Neural computation, 3(1):79–87.
[2] Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. Branch-train-merge: Embarrassingly parallel training of expert language models. In First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022.
[3] Suchin Gururangan, Margaret Li, Mike Lewis, Weijia Shi, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. Scaling expert language models with unsupervised domain discovery. ArXiv, 2023

[1] goo 辞書 “三人寄れば文殊の知恵（さんにんよればもんじゅのちえ）”から引用
[2] シード値を変えることで 3 種類作成したモデルを同じ重み(1/3)でアンサンブルした