Ruri: 日本語に特化した汎用テキスト埋め込みモデルの開発

塚越 駿, 笹野 遼平　　名古屋大学

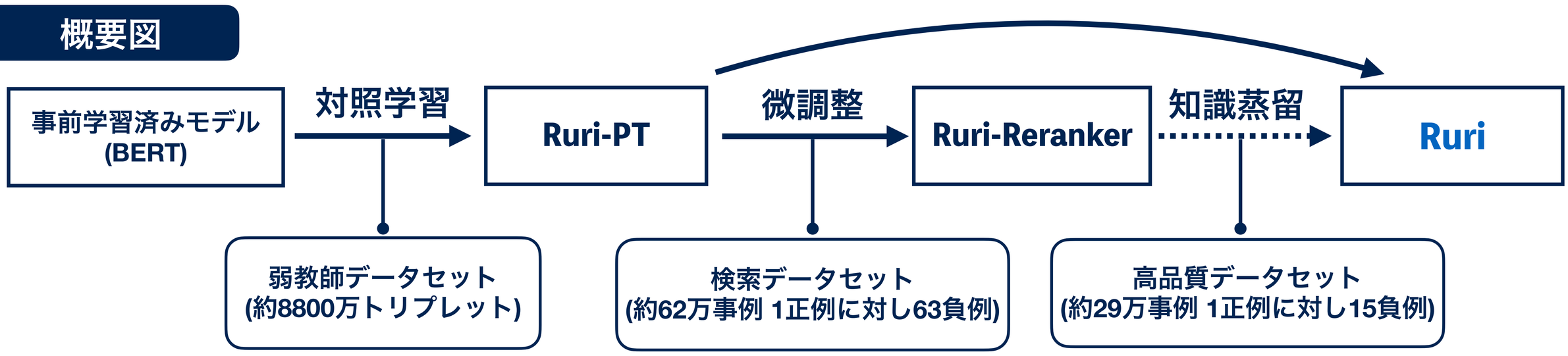
まとめ

- 日本語汎用テキスト埋め込みモデルの開発
  - 埋め込みモデル訓練のための人工データセット作成
  - 大規模な対照事前学習によるベースモデルの構築
  - 日本語評価において最高性能のRerankerの構築
  - Rerankerからの知識蒸留を用いた微調整モデルの開発
- モデル・訓練データを商用利用可能なライセンスで公開

背景

- テキスト埋め込み: テキストを表現する密ベクトル表現
  - 検索拡張生成・クラスタリング・文書検索などで応用
- 英語や多言語では対照事前学習と微調整によるモデル構築が主流に
- 日本語では大きく分けて二つの問題
  - データセットの不足 (特に検索/QA)
  - 構築ノウハウの不足

概要図



既存データセットの整備

- 合計21の前処理・負例マイニング済みのデータセットを公開
  - https://huggingface.co/datasets/hpprc/emb
- データフォーマットの統一、前処理と正規化
  - BM25, mE5-largeを利用した負例マイニング
- QA・クイズデータセットに対し正例・負例文書を自動付与
  - マイニングした事例のうち回答を含む文書を正例に

人工データセットの構築

- QAデータセット (AutoWikiQA)
  - 日本語Wikipedia記事中のパラグラフに対し質問と回答を生成
  - Swallow-MXで230万件, Nemotron-4-340Bで15万件
- NLIデータセット (AutoWikiNLI)
  - 日本語Wikipediaから文を抽出し含意・矛盾にあたる文を生成
  - Nemotron-4-340B Rewardモデルを利用してフィルタリング

Contrastive Pre-training

- テキスト埋め込みモデルのための事前学習として弱教師データセットでの大規模対照事前学習を実施
  - Web上のQAデータセットやWikipediaを中心に8800万事例を収集して訓練用データセットを構築
  - batch size 8192、系列長を256 or 192に設定
- 対照事前学習の時点でmE5-largeを超える性能を達成
  - 人工検索データセットの性能向上への寄与を確認

データセット	量
Wikipedia	3800万
MQA	2500万
CC News	900万
AutoWikiQA	1200万
論文コーパス	10万
言い換え	400万
合計	8800万

Model	JMTEB Avg.
mE5-large	70.98
Ruri-PT-small	70.41
Ruri-PT-base	70.80
Ruri-PT-large w/o retrieval	71.11
Ruri-PT-large	**72.46**

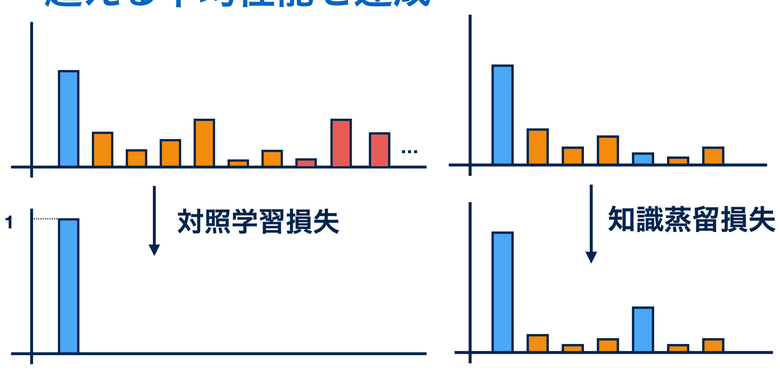
Rerankerの構築

- 埋め込みモデルの文書検索性能向上のためRerankerからの知識蒸留を実施
  - Reranker: クエリと文章を入力として関連度スコアを出力するモデル
  - 知識蒸留: Rerankerと埋め込みモデルのスコア分布を近づける学習手法
- 人工データセットを有効利用するため、2段階の学習手法を採用
  - 1st. stage: 人工データ含むnoisyだが多様なデータセットで学習
  - 2nd. stage: 主に人手データによる高品質なデータセットで学習
- Rerankerとして日本語評価データにおいて最も高い性能を達成

Model	JQaRA nDCG@10	JaCWIR MAP@10	MIRACL Recall@30
hotchpotch/japanese-reranker-cross-encoder-large-v1	71.0	93.6	91.5
bge-reranker	67.3	93.4	94.9
Ruri-Reranker-small	64.5	92.6	92.3
Ruri-Reranker-base	74.3	93.5	95.6
Ruri-Reranker-large	**77.1**	**94.1**	**96.1**

Fine-tuning

- 対照事前学習により構築したベースモデルを微調整
- 構築したRerankerによりQAデータセットのフィルタリングとスコアづけ → 対照学習損失+知識蒸留損失
- 既存の多言語汎用テキスト埋め込みモデルE5 (mE5)やGLuCoSEを超える平均性能を達成



データセット	量
検索・QA	9万
自然言語推論	20万
合計	29万

Model	Avg.	Retrieval	STS	Class.	Reranking	Clus.	Pair.
GLuCoSE	70.44	59.02	78.71	76.82	91.90	49.78	66.39
mE5-small	69.52	67.27	80.07	67.62	93.03	46.91	62.19
mE5-base	70.12	68.21	79.84	69.30	92.85	48.26	62.26
mE5-large	71.65	70.98	79.70	72.89	92.96	51.24	62.15
Ruri-small	**71.53**	69.41	82.79	76.22	93.00	51.19	62.11
Ruri-base	**71.91**	69.82	82.87	75.58	92.91	54.16	62.38
Ruri-large	**73.31**	73.02	83.13	77.43	92.99	51.82	62.29

JMTEBの性能