

森羅プロジェクト 取り組み リンクングタスク

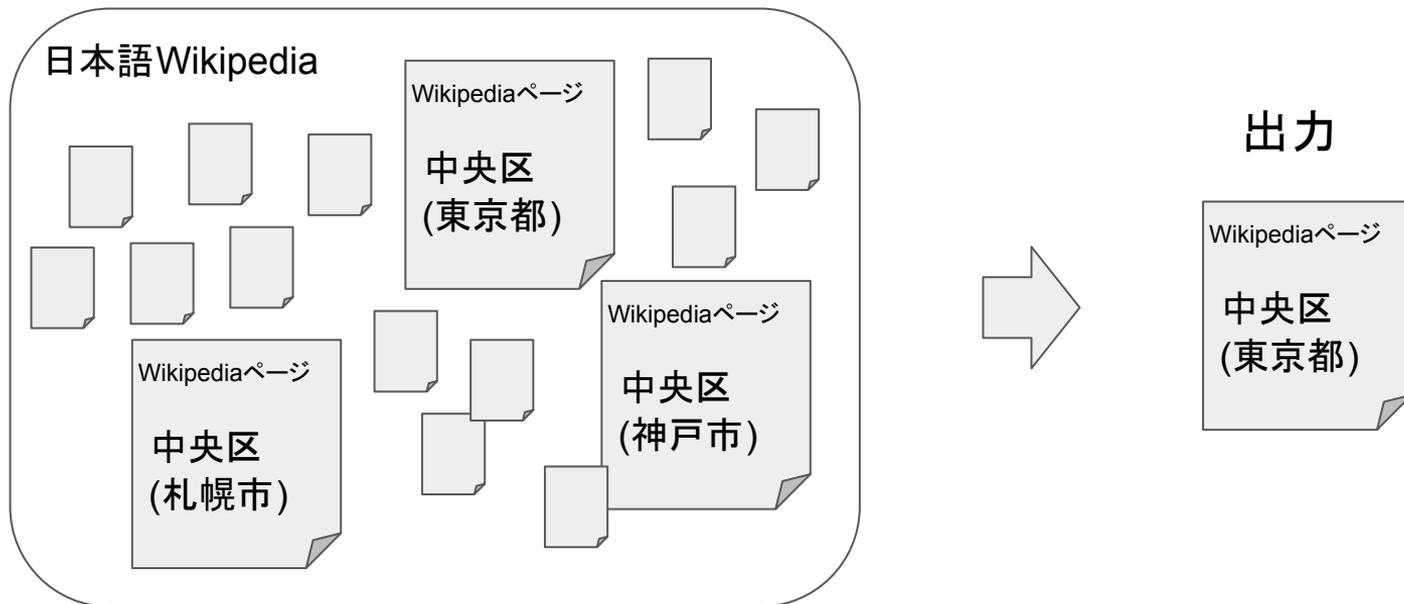
筑波大学 陰山 宗一

目次

1. タスクの確認
2. 手法
 - a. Bi-Encoder
 - b. Cross-Encoder
3. 評価結果
 - a. 自前testデータ(教師データの一部)
 - b. リーダーボード
4. リーダーボードでの難しい点
5. 試したが上手くいかなかったこと

タスクの確認（リンクング）

- 文章の単語とリンクするWikipediaページを予測
 - 入力：中央区の人口は千代田区に次いで2番目に少ない

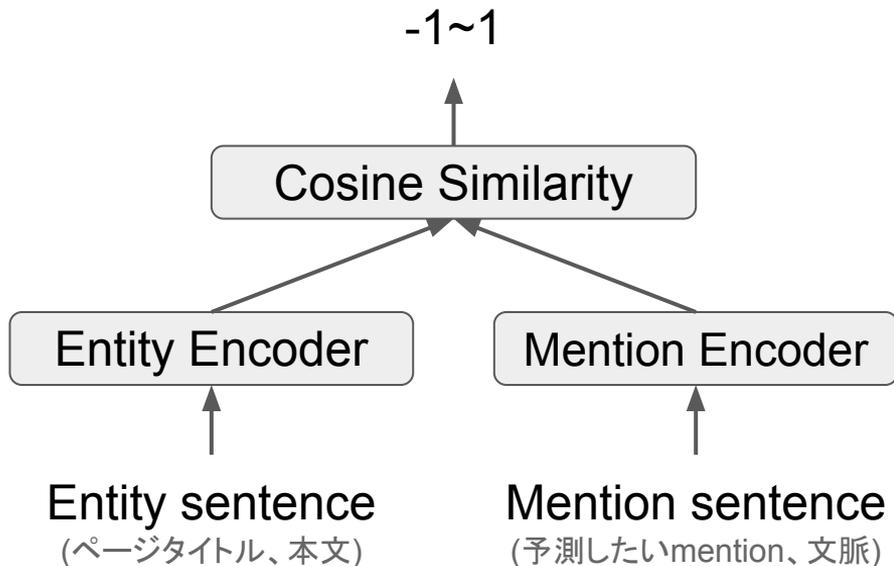


[1] 予測したいmentionが複数あるときの話。
Bi-Encoderは事前に全てのEntityをEncodeしておけば効率的に計算できるが、
Cross-EncoderはMentionを予測するごとにEntityの数だけEncodeする必要。

手法 (モデル)

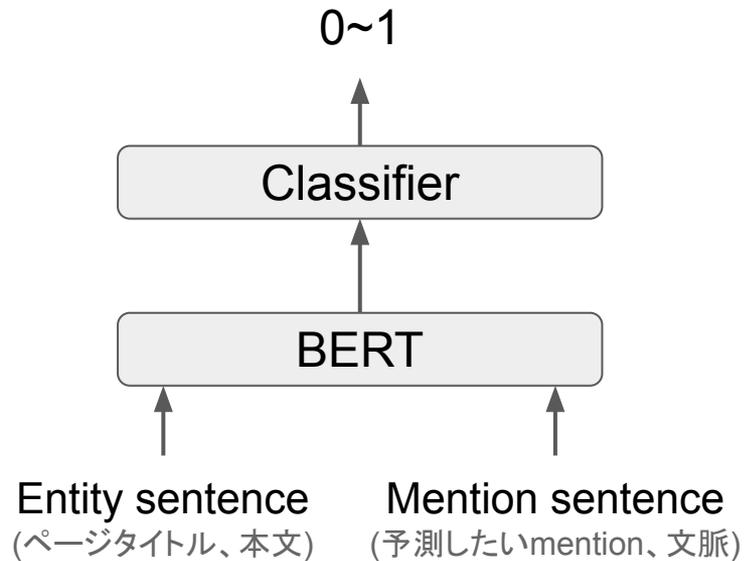
- **Bi-Encoder**

- 性能そこそこ！ 効率いい！ [1]



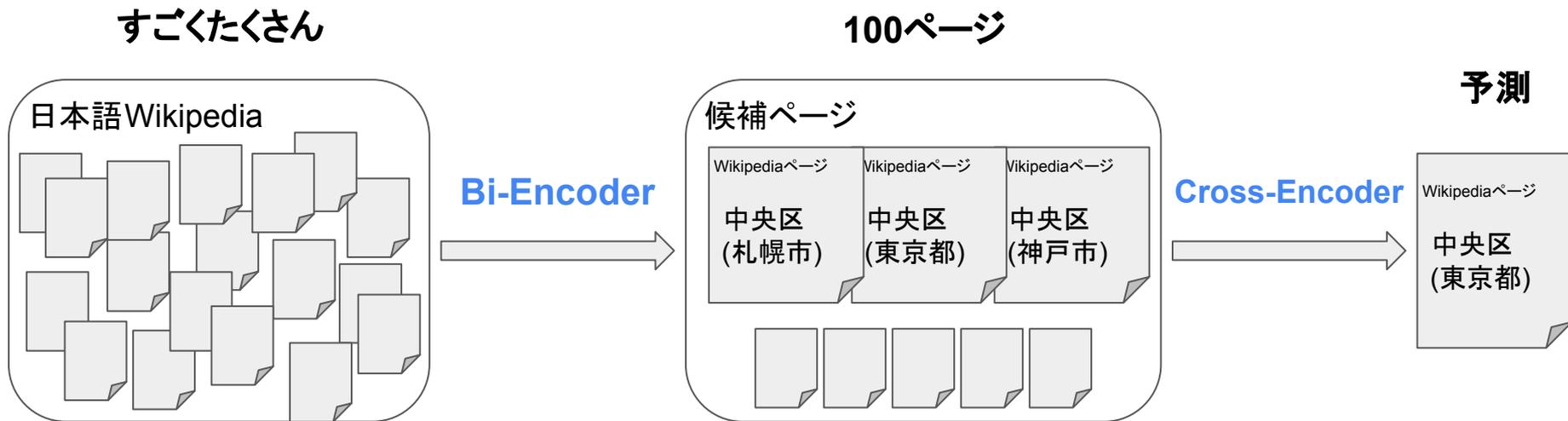
- **Cross-Encoder**

- 性能高い！ 時間かかる... [1]



手法

入力:中央区の人口は千代田区に次いで2番目に少ない



評価結果

- 自前testデータ(配布されてる教師データの一部)

| model | Recall |
|---------------------------------------|---------------------|
| Bi-Encoder (top100) 全ページ→候補100ページ | 86.38 |
| Cross-Encoder (top1) 候補100ページ→1ページ | 90.27 |
| 全体 全ページ→候補100ページ→1ページ | <u>77.97</u> |

予測対象データ:
属性値抽出ベースラインシステム出力
(Micro-F1:51.51)

- リーダーボード

| model | Micro-F1 |
|--|----------|
| Bi-Encoder (top1) 全ページ→1ページ | 9.14 |
| + Cross-Encoder (top1) 全ページ→候補100ページ→1ページ | 19.09 |
| ルールベース (非常にシンプルなアンカリンクを用いた手法) | 29.53 |
| (ベースラインシステム) | 41.44 |

リーダーボードでの難しい点

- リンキングタスクの予測は2種類
 - リンクするページのID
 - リンクするページはない ← これをやっていない性能
- 評価指標がMicro-F1、予測対象データが属性値抽出の**不正解**を含む
 - 「**予測をしない**」という選択肢
 - 不正解データに予測 → precisionが低下

試したが上手くいかなかったこと

- Wikipediaの内部リンクを用いたBi-Encoderの事前学習
 - 内部リンクとリンクングタスクの相性？
 - 今後、分析する予定