Wikipediaを用いた自動校閲機能の検討

大阪大学大学院情報科学研究科

舌　達也

目次

- 会社紹介
- インターンスケジュール
- 背景
- 取り組み
- 感想

会社紹介

会社名	株式会社エクサウィザーズ
所在地	105-0021 東京都港区東新橋1丁目9-2 汐留住友ビル 21階
設立	2016年2月
資本金	22億円（2022年6月時点）
従業員数	383名（2022年6月時点/正社員）
事業内容	AIを利活用したサービス開発による 産業革新と社会問題の解決
グループ会社	エクスウェア株式会社 株式会社エクサホームケア 株式会社VisionWiz
海外現地法人	EXAWIZARDS INDIA LLP（インド） EXAWIZARDS LLC（アメリカ）
株式公開市場	東京証券取引所マザーズ市場（現グロース市場） 証券コード：4259　上場日：2021年12月23日

会社紹介

DX AIプロダクト

exaBase DXアセスメント&ラーニング
exaBase ウェルビーイング
exaBase 1on1
exaBase 予測・分析
exaBase FAQ
exaBase 企業検索

exaBase DXアセスメント&ラーニング

DX AIプロダクト提供により
社会的価値を実現、
約400社に導入済

✓ 公的機関、大企業、中小企業含む
社会全体の効率性を改善

日本のDXをリード可能なIT人材不足*

430,000人

2025年以降に予想される経済的損失*

年間12兆円

ソーシャルAIプロダクト

トルト　ハナスト　とりんく

ハナスト

介護スタッフの労働環境を改善

✓ 話すだけで介護内容を記録
✓ 1日当たり平均40分の労働時間を削減(1)

主要KPI

- 介護施設あたり月26,000円(3)
  ハナスト → CARE KARTE 12,000施設(2)
- 将来的に約27,000施設(4)まで拡大する可能性

*注記：(1)当社実施のユーザー調査における利用者評価より推定 (2) 2021年3月時点。提携先のケアコネクトジャパンの有する施設数 (3) 介護施設当たりの想定入居者数に基づく当社推計 (4) 厚生労働省の「介護サービス施設・事業所調査」における2019年時点の介護関連施設数に基づく当社推計***
出所： 経済産業省「DXレポート ～ITシステム「2025年の崖」克服とDXの本格的な展開～」（2019年9月） **厚生労働省「令和元年介護サービス施設・事業所調査の概況」（2019年10月）*

インターンスケジュール

機械学習エンジニア (自然言語処理) のインターンに参加

- 8/15
  テーマ説明、PCセットアップなど (本社)
- 8/16〜9/15
  インターン業務 、ミーティングなど(リモート)
- 9/16
  成果報告会、交流会 (本社)

背景

- テーマ
  テキスト作成業務を補助する自動校閲アプリケーションの開発・検討
- 校閲とは？
  文章に書いてある内容の事実誤認、矛盾、無許可の引用や不適切な表現が無いかなどを確認して修正する作業
- 校正との違い
  - 校正は誤字脱字や英語のスペルミス、表記のゆれを修正する
  - 校閲は文章の内容にまで踏み込む
  - 校閲を行うアプリケーションは少ない

背景

- 例
  - 痛みの緩和・万病の予防をします (承認範囲外の表現)
  - 最高級のたんぱく質 (誇大広告、最大級表現)
  - 内臓脂肪や皮下脂肪を減らすのを助ける (優良誤認表示)
  - 他社商品の2 倍の内容量 (有利誤認表示)

- したいこと

  入力された文章中の不適切な可能性のある部分を検知するモデルの開発

取り組み

取り組み - Wikipedia上の記事の編集履歴を取得

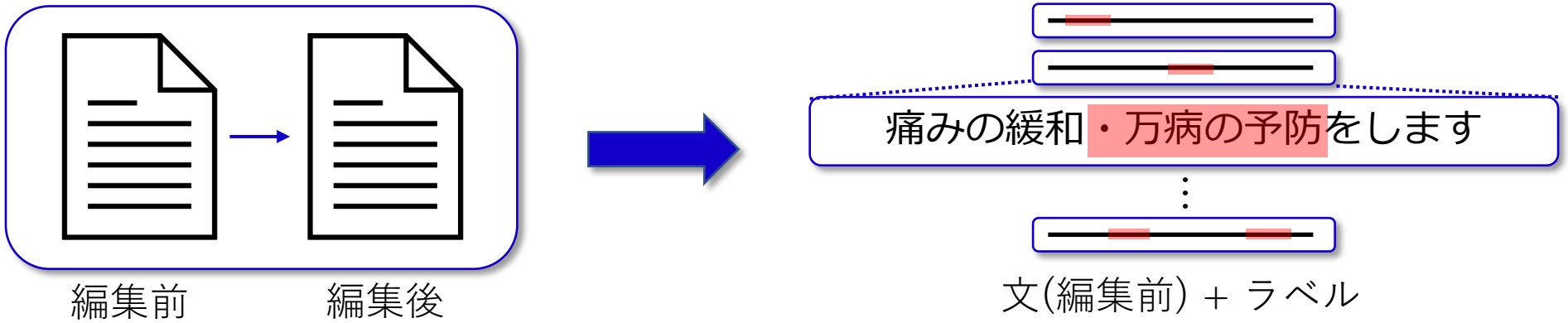
文に校閲にあたる編集が行われているラベルをつけたデータが必要

- Wikipediaの記事における編集履歴の例

- Wikipediaの記事の特性
  - 中立的な視点、検証可能性、独自研究は載せない
  - これらに基づいた編集には、本タスクにとって望ましい編集を含む

取り組み - データセットの作成

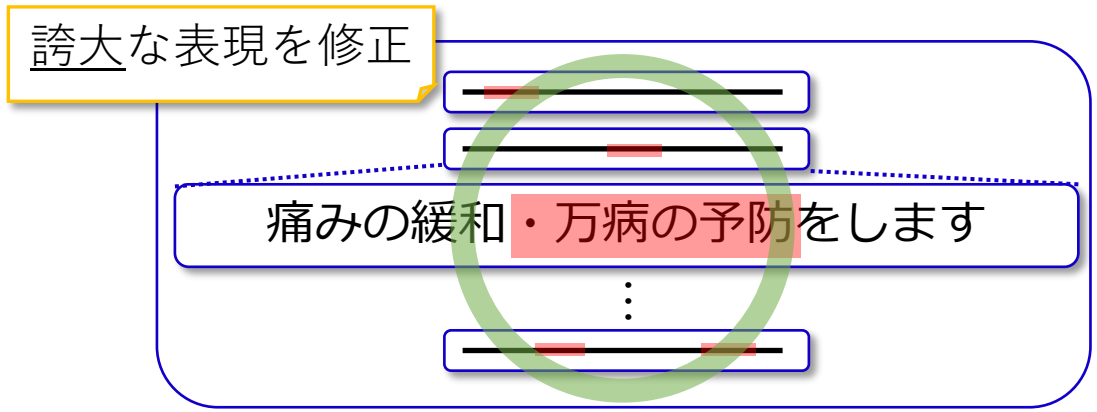
- 記事の情報
  - 記事id、タイトル、カテゴリ、編集情報 (日時、コメント)、本文など
- 編集前後の記事対からラベルつきデータを作成
  - 編集前後の各文に対して類似度を測定
  - 程よい編集が行われている文の編集部分を抽出

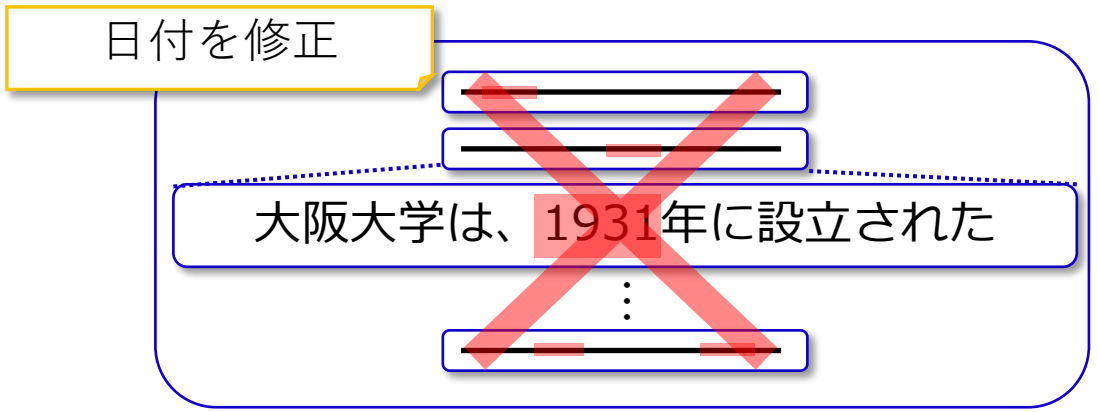


取り組み -フィルタリング

Wikipedia上の編集が校閲にあたる編集とは限らない

- 編集コメント (なぜその編集をしたか)を用いたフィルタリング
- 「誇大」「独自研究」「検証可能性」等のキーワードが編集コメントに含まれているかどうか
- その他、漢字→ひらがな変換のみの編集は削除など

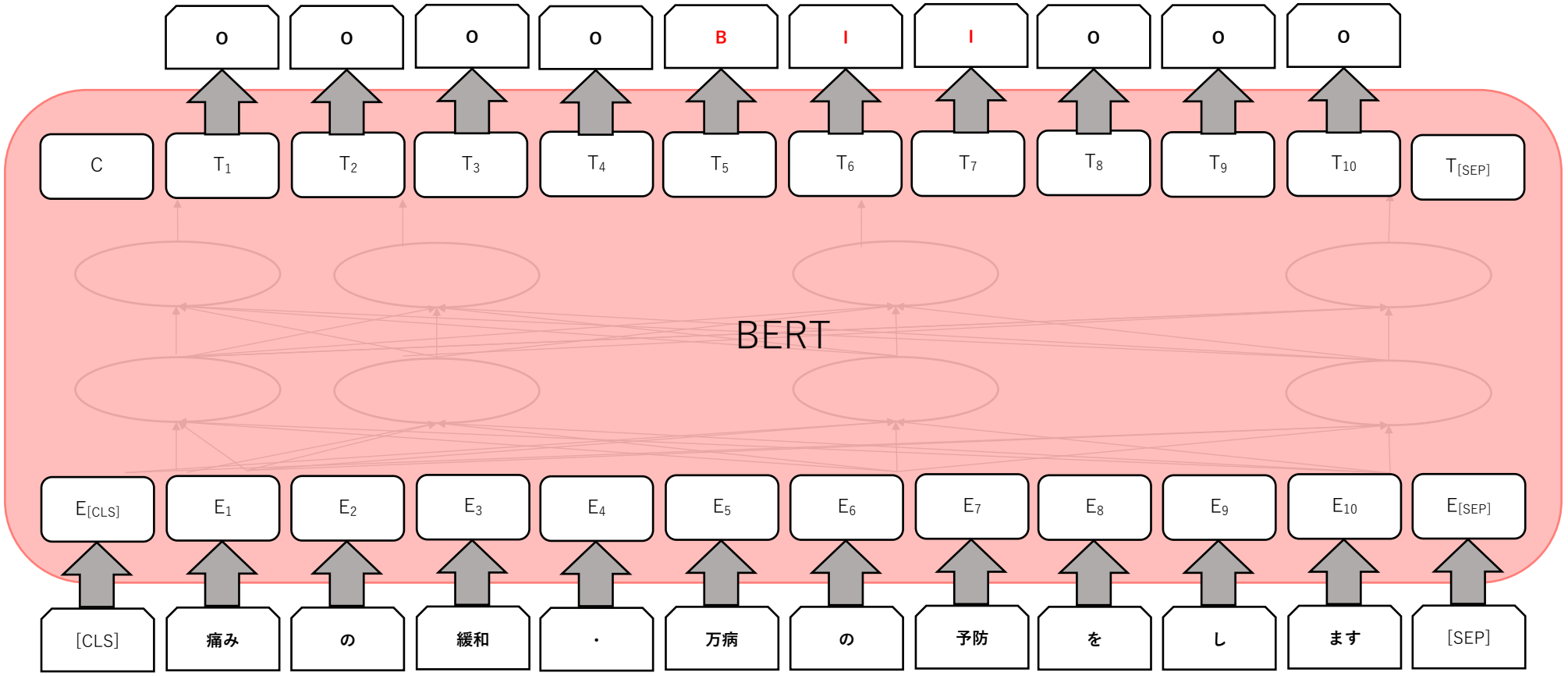




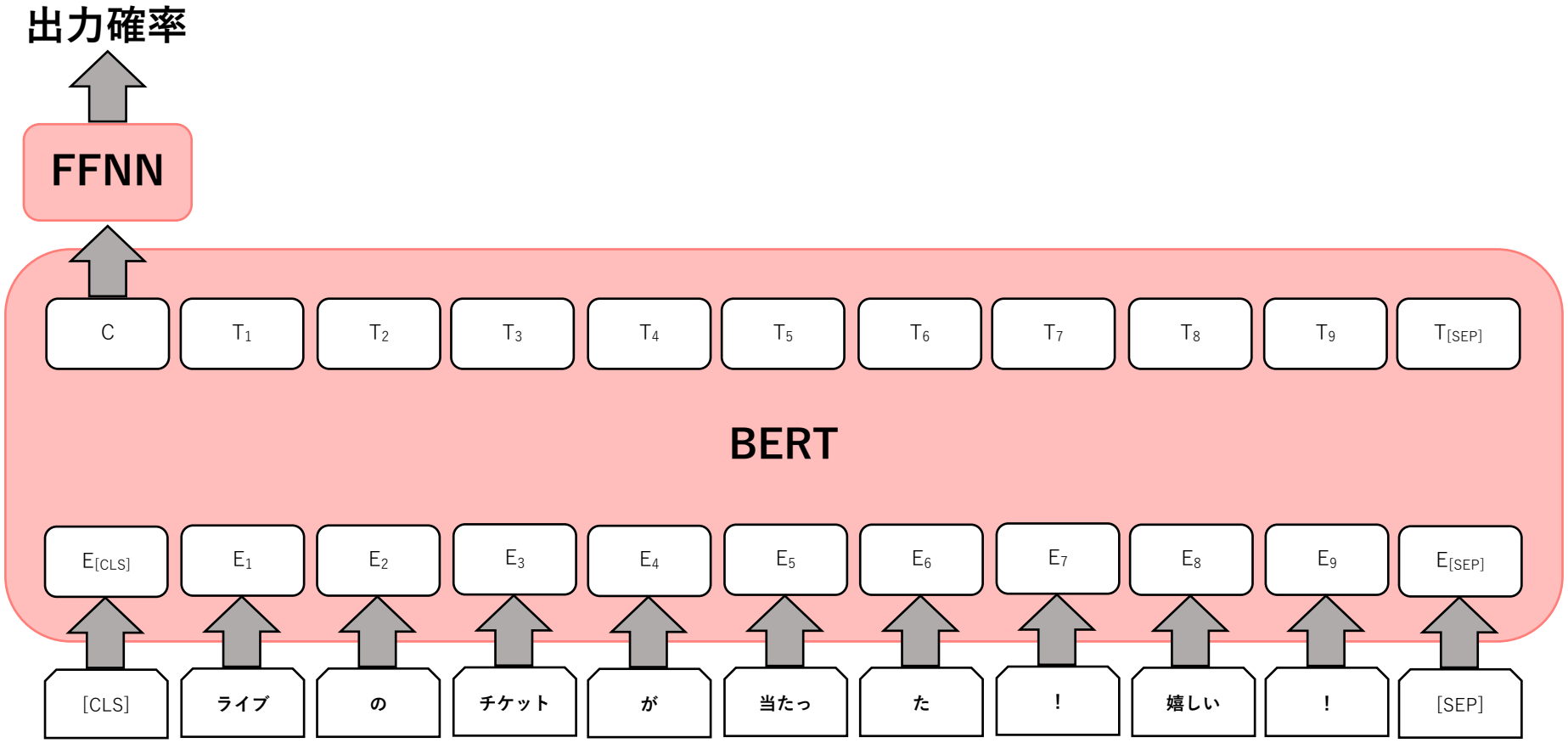
取り組み – モデルの学習・検証

- データ数: 約7万件
- モデル: BERT
  - 自然言語処理タスクのための汎用的な事前学習済みモデル
  - 事前学習後の重みを初期値とし、今回のデータでファインチューニング
- 学習したモデルの検証
  - 作成データ内のテストセット: 約65%ほどの正解率
  - 独自の薬剤の添付文書: 約30%ほどの正解率
- その他の実験
  - 検知理由も提示するモデルの訓練
  - 検知のみならず、検知した箇所の修正候補を提示するモデルの訓練

取り組み – モデルの学習・検証



BERTによるラベル予測


出力確率
FFNN
C
T1
T2
T3
T4
T5
T6
T7
T8
T9
T[SEP]
BERT
E[CLS]
E1
E2
E3
E4
E5
E6
E7
E8
E9
E[SEP]
[CLS]
ライブ
の
チケット
が
当たっ
た
！
嬉しい
！
[SEP]

取り組み – まとめ

- 校閲にあたる箇所の検知機能の開発・検討のために、データセット作成、モデルの訓練を行なった
- ある程度機能したが、ノイズデータの除去が不十分
- 特定のドメインに対してはさらなる工夫が必要
- 検知理由、修正候補の推論も今後の課題

感想

- メンター以外の方とのミーティング
  - ビジネスサイドの専門家 (医療系文書)
  - 他分野のエンジニア
- ノイズデータ除去の困難さ
- 普段の研究生活では用いないツール
- 成果報告会での他分野の学生との交流