EXA CORPORATION

BERTによる類似文検索と文書要約

Exa Value Forum 2020

2020年 11月26日

株式会社エクサ
テクノロジーイノベーション部
安井　由香

もくじ

- はじめに
- BERTを使った類似文検索
- BERTを応用した文書要約
- おわりに

EXA CORPORATION

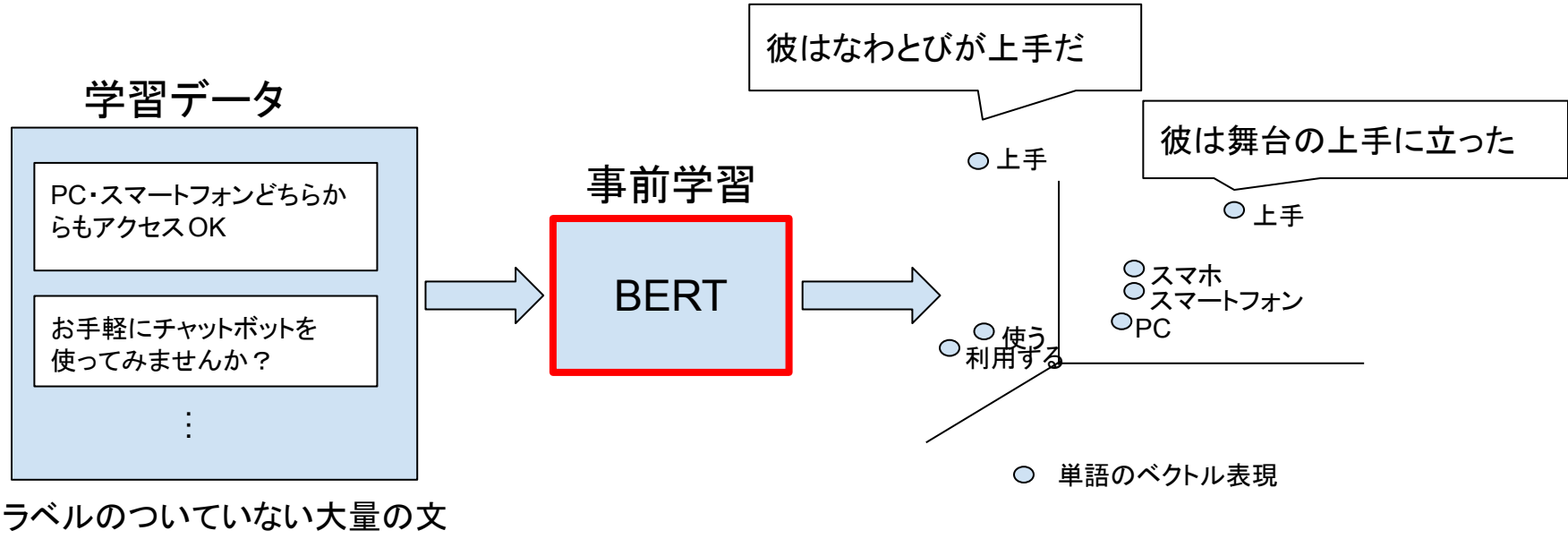
はじめに

BERTとは

- 2018年後半にGoogleが発表した自然言語のDeep Learningモデル
- 画像に比べると遅れていた自然言語のDeep Learningがブレークスルーするきっかけとなった
- BERTの最大の特徴は文脈を理解できること
  - 一部のタスクでは人間を超える性能を誇る

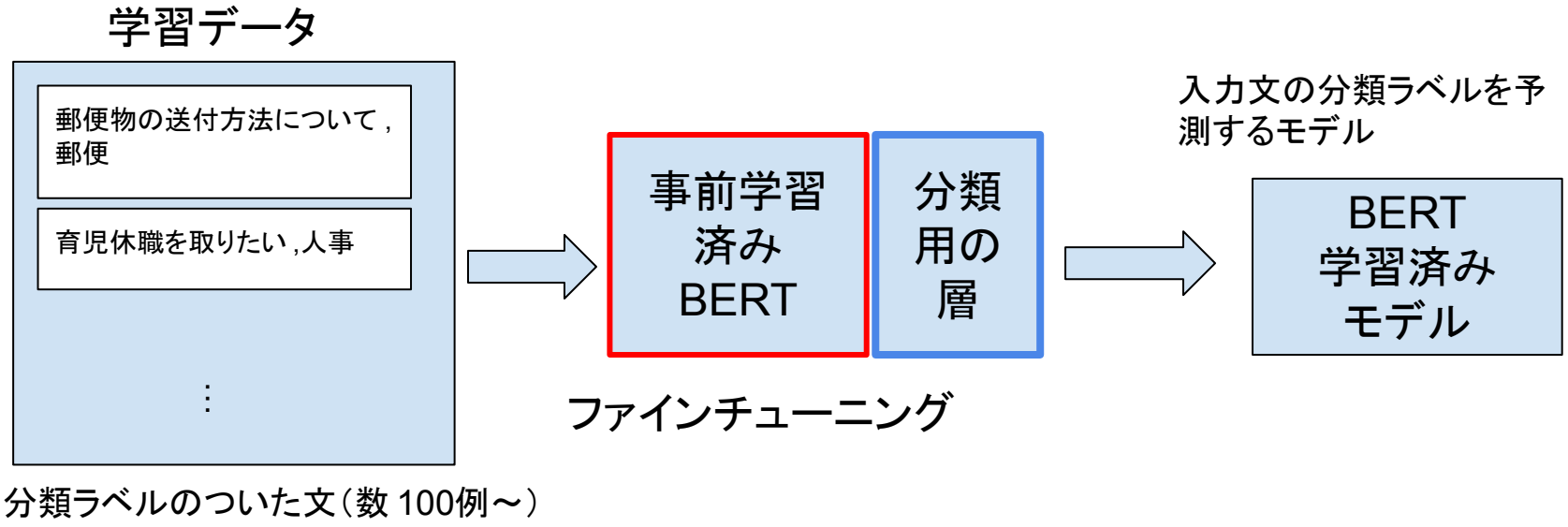
BERTの学習手順1:事前学習

- 事前学習したBERTでは単語の「文脈に応じたベクトル表現」を得られる
- 事前学習は大量の文章が必要
  - 日本語Wikipediaで事前学習したものが公開されている



BERTの学習手順2:ファインチューニング（微調整）

- ファインチューニングとはある領域で学習済みの知識を他の領域に適用すること
- 事前学習済みモデルを少量のデータでファインチューニングして分類や質問応答などを行うことができる



研究の目的

BERTに関する疑問

日本語ではどの程度の精度が出る？
研究用データではなく独自のデータをBERTで学習するには？
どういったシステムに使える？

BERTを使った2種類のデモ作成を通じて技術習得
①類似文検索
②文書要約

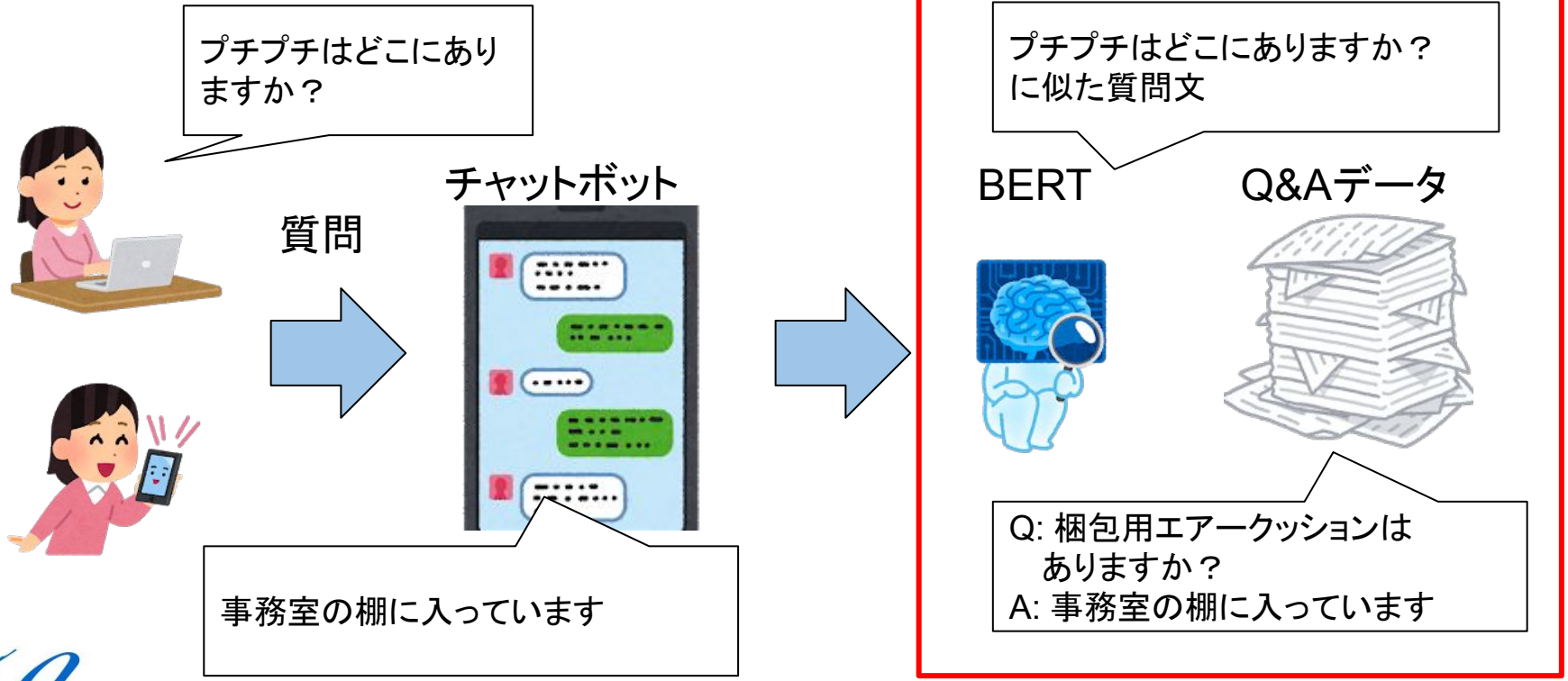
BERTを使った独自ソリューション開発

EXA CORPORATION

BERTを使った類似文検索

類似文検索のユースケース

パソコンや音声を通じた質問応答チャットボットにBERTを適用



利用したデータ

- エクサの質問応答チャットボット「社内サービスナビ」のデータを利用
- エクサ社内の事務手続きや契約関連のQ&Aデータ
- 営業管理を除く619件のデータを使って検証した
  - 営業管理は専門用語が多く難しい
  - 営業管理だけで全体の半数を占める

No	分類	件数
1	Concierge Desk	32
2	慶弔	31
3	支払い	26
4	総務	163
5	配布・送付	23
6	人事	119
7	旅費	64
8	文房具・事務用品	121
9	情報システム	28
10	その他　挨拶等	12
11	営業管理	608
	合計	1227
	営業管理除く合計	619

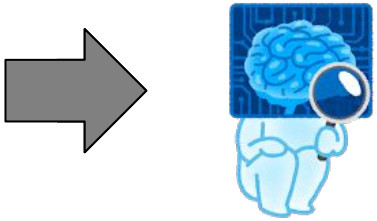
＊2019年7月時点のデータ

EXA CORPORATION

BERTで2つの文章の関係性判定

文A：猫に小判

文B：猫をかぶる



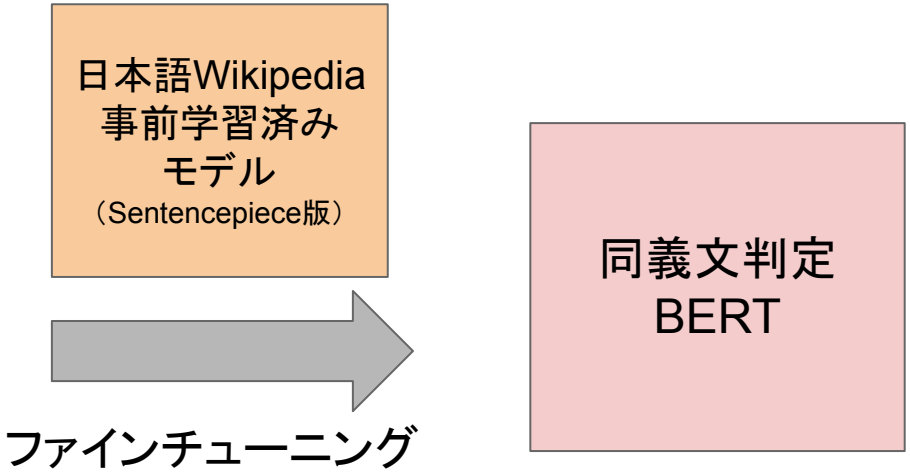
違う意味！

BERTをファインチューニングしてクエリと同じ意味のQ&Aデータを探す

社内サービスナビデータの学習

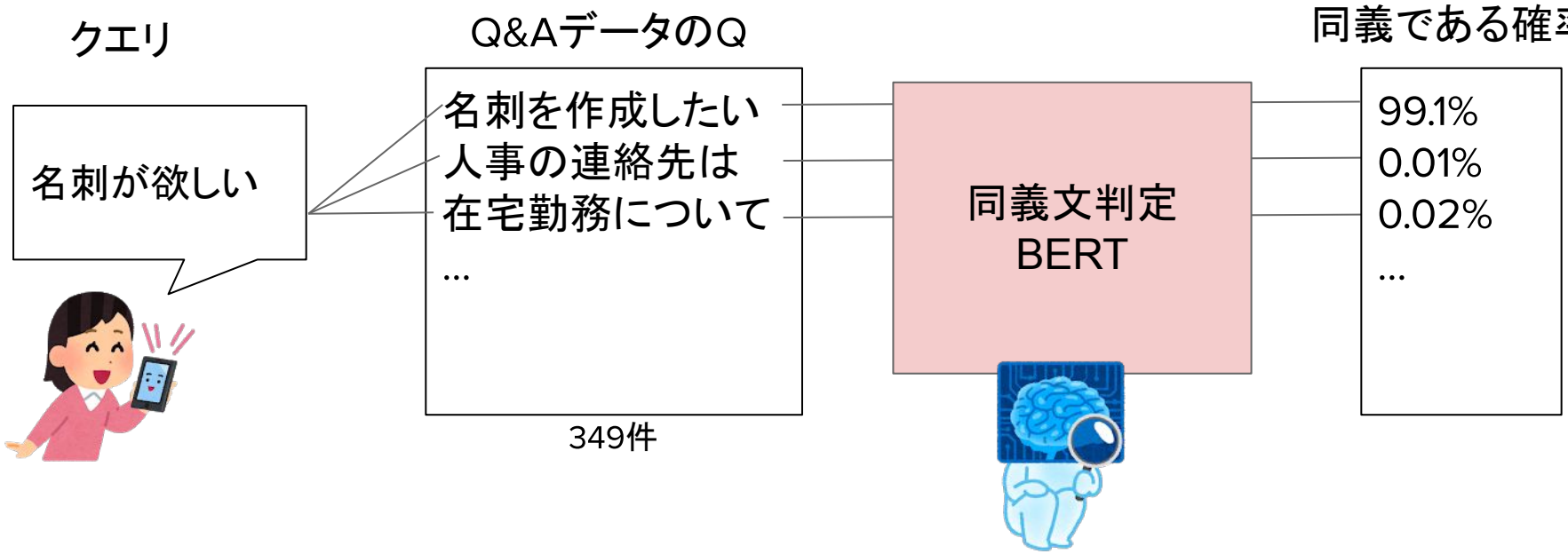
質問文のペアを用意し、同じ意味か、違う意味かのラベルをつける
今回は学習データを1898件用意

文A	文B	同義(1)or異義(0)
名刺を作成したい	名刺が欲しい	1
人事の連絡先は	テレコンを借りたい	0
花束を配達してほしい	名刺が欲しい	0



社内サービスナビデータの検索

クエリと同義の質問文を見つける



精度

No	手法	1位正解率	5位正解率	10位正解率
1	同義文判定BERT	53.3%	83.3%	86.6%
2	BERT事前学習のみ	25.0%	43.3%	46.7%
3	TF-IDF（ベースライン）	46.7%	73.3%	76.7%

＊n位正解率：検索結果の上位 n個の候補において、1位～n位以内に正解が含まれる割合

同義文判定BERTで最も高い精度となった

EXA CORPORATION

検索結果の一例

ファインチューニングで「プチプチ＝梱包用エアークッション」ということを学習できている

クエリ：プチプチはありますか

順位	BERT事前学習のみ	スコア（クエリとの距離）
1位	クリアファイルはありますか	39.717
2位	シャーペンは、ありますか	42.246
3位	サインペンは、ありますか	42.707
4位	ボールペンはありますか	43.531
5位	電池は、ありますか	43.971

順位	同義文判定BERT	スコア（クエリと同義である確率）
1位	梱包用エアークッションは、ありますか	0.991
2位	梱包用エアークッションが欲しい	0.990
3位	-	-
4位	-	-
5位	-	-

EXA CORPORATION

BERTを使った類似文検索：まとめと課題

- まとめ
  - BERTをファインチューニングし、クエリと同義のQ&Aデータを検索することができた
- 課題
  - 検索速度の向上（GPU：20s/query、CPU：300s/query）
    - BERTの軽量化
    - 推論回数を減らす工夫
    - 事前学習による良質なベクトル取得
  - 精度の向上
    - BERT単独ではなく複数の手法と組み合わせる
    - BERT以降の最新モデルの検証

EXA CORPORATION

BERTを応用した文書要約

文書要約のユースケース

ニュース記事などの文章の中から重要な3文を抜き出す

入力（研修案内文）

BERTはGoogleが開発した高性能な自然言語処理モデルで、Google検索をはじめ様々な自然言語処理・理解に用いられています。
本研修では、BERTによる日本語文書分類モデルの作成から実際に分類を行うデモアプリケーションの作成までを体験していただきます。
目次
1.機械学習における自然言語処理の流れ
2.BERTの構造と学習
3.[演習] HuggingFace Transformersを使った学習・推論
4.[演習] Flaskを使った簡単なデモアプリ構築
3の演習では、Google ColaboratoryでBERTのファインチューニングを行います。
BERTのファインチューニングにはHuggingFace社のライブラリであるTransformersを用います。
また、4の演習では3で作成した学習済みモデルを使って実際に分類を行うアプリケーションを作成します。
4の演習は各自のPCに環境構築して実施します。
＜参加者前提＞
・必須
エクサ標準PC（情報システム部より配布されているWindows10搭載のPC） 保有者
・推奨
何らかのアプリケーション開発経験があること。
＊研修で扱う言語はPythonですが、サンプルコードを提供しますのでPythonの知識は必須ではありません。
＜研修日数＞
1日（9:30 ~ 16:30）
＜開催日時・場所＞
第1回 10月15日（木） 9:30～16:30 オンライン
第1回 10月22日（木） 9:30～16:30 オンライン
第1回 10月29日（木） 9:30～16:30 オンライン
＜講師＞
テクノロジーイノベーション部 安井 由香
＜定員＞
5名
＜申込方法＞
Gsuiteにて募集します。
ご希望の開催回のリンク先にて申込み下さい。
キャンセルは下記問い合わせ先へ連絡してください。
期日までに2人以上の申込がない場合は中止します。
募集期限は各日程の1週間前です。

BERT

BERTはGoogleが開発した高性能な自然言語処理モデルで、Google検索をはじめ様々な自然言語処理・理解に用いられています

本研修では、BERTによる日本語文書分類モデルの作成から実際に分類を行うデモアプリケーションの作成までを体験していただきます

また、4の演習では3で作成した学習済みモデルを使って実際に分類を行うアプリケーションを作成します

EXA CORPORATION

要約の種類：抽出要約と抽象要約

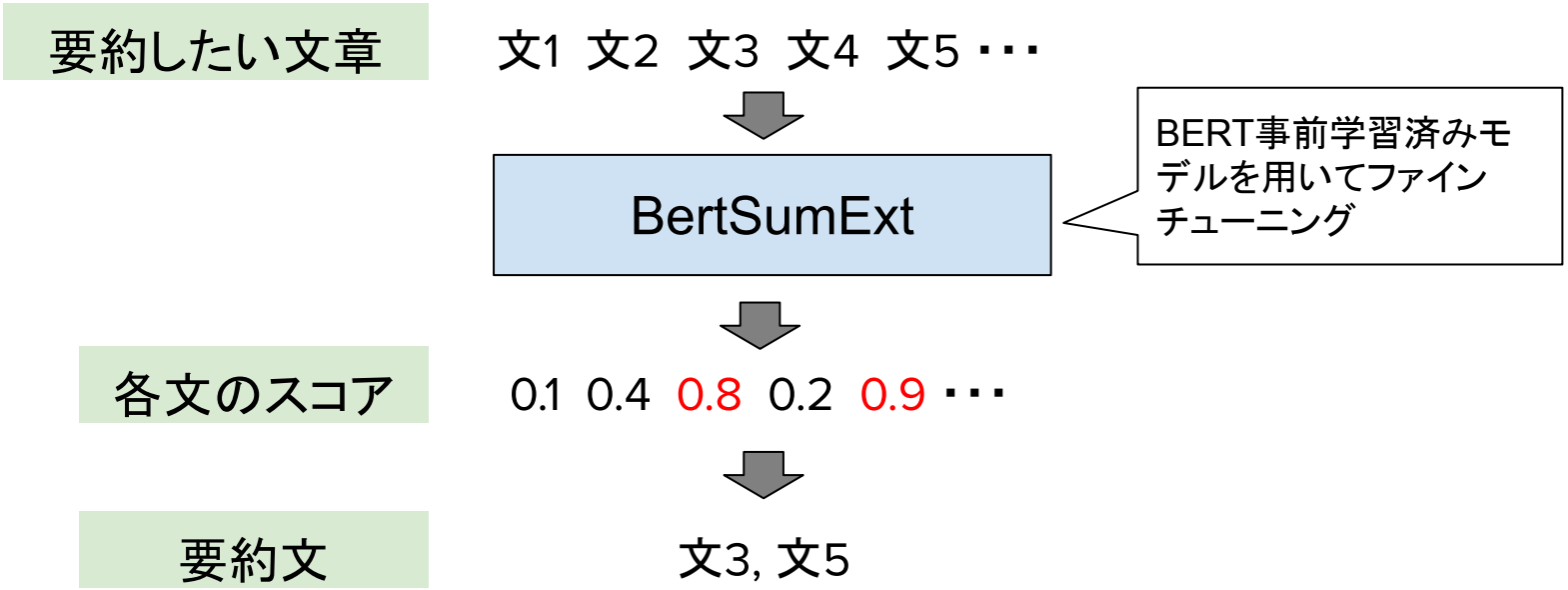
- 抽出要約
  - 元となる文章から重要な文を抜き出して要約文とする
  - 元の文をそのまま抜き出すので文法の誤りがない
  - 元の文の表現を変えることはできない
- 抽象要約
  - 元となる文章の内容に応じた要約文を作る
  - 元の文にはない表現・単語を使える
  - 自然な文を作ることが難しい

EXA CORPORATION

BERTの文書要約への応用：BertSum

抽出要約

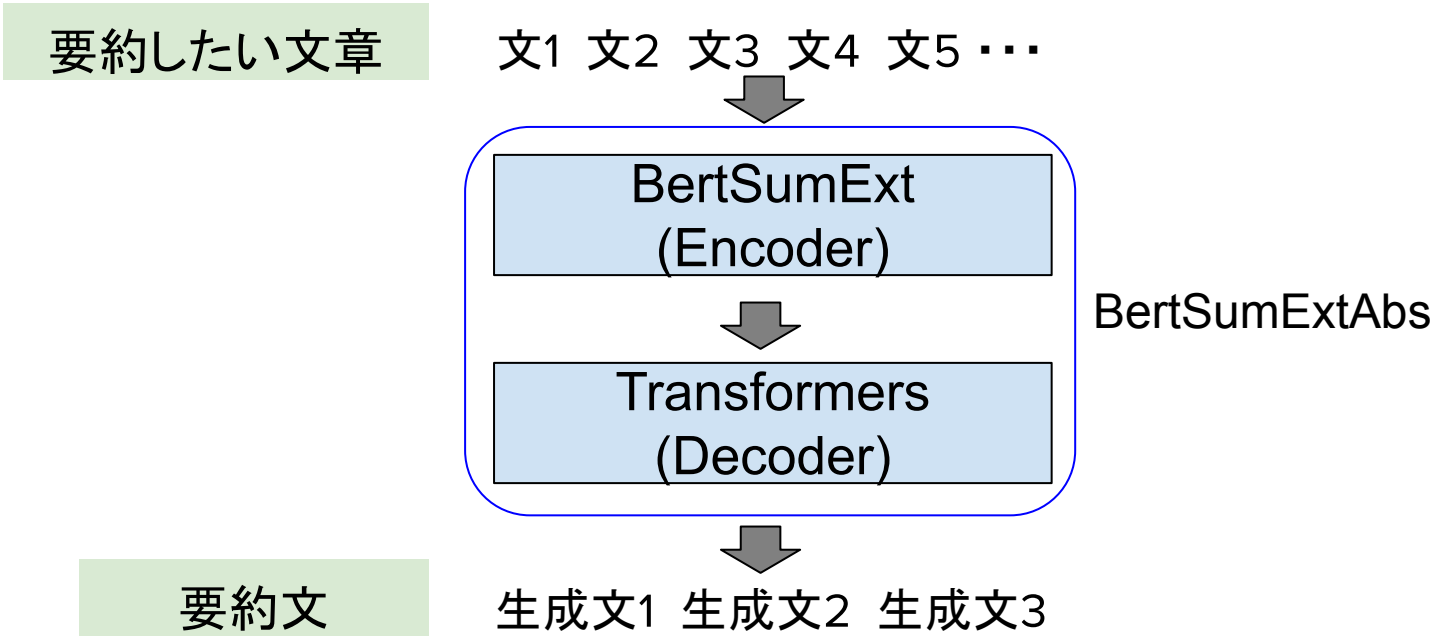
- 要約したい文章の各文に対して0～1のスコアを付ける
- スコアが1に近いほど要約文として抽出すべき文



BERTの文書要約への応用：BertSum

抽象要約

- 抽出要約で学習済みのBertSumExtをEncoderとして用いる
- Decoderは多層のTransformerで要約文を生成する



利用したデータセット

- 三行要約データセット
  - KodairaTomonori/ThreeLineSummaryDataset
  - Livedoor Newsの記事本文を三行にまとめた要約文（三行要約）が付いている記事のIDを集めたデータセット
  - 要約文は人手で書かれた「抽象要約」

	データセットに含まれる 記事IDの件数	クローリングして実際に 集まった件数（2020年2月時点）
学習	210,000	105,923
テスト	1,200	888
バリデーション	1,200	870

EXA CORPORATION

結果

- 正解の要約文（抽象要約）とBertSumの要約文（抽出要約/抽象要約）の一致度をROUGEスコアで計算

No	要約文の選択方法	ROUGE-1	ROUGE-2	ROUGE-L	説明
1	正解ラベルの付いた文	59.32	40.39	48.60	元の記事中で正解の要約文に最も近い文（正解ラベルの付いた文）と正解の要約文の比較。抽出要約ですべて正解すればこの値になる。
2	**BertSum抽出要約**	**46.14**	**24.28**	**34.29**	**BertSumExtによる抽出要約**
3	**BertSum抽象要約**	**42.10**	**17.10**	**30.01**	**BertSumExtAbsによる抽象要約**
4	冒頭3文	41.95	19.14	29.29	記事の冒頭3文との比較。冒頭に重要な文があることが多いため要約評価の際にリファレンスとして用いられる。
5	LexRank	39.78	17.22	27.26	他の手法。文をグラフ構造に置き換えて重要度を算出。

EXA CORPORATION

BERTを使った文書要約：まとめと課題

- まとめ
  - BERTをファインチューニングし、抽出要約と抽象要約を試行した
    - 抽出要約では高精度なモデルを作成できた
    - 抽象要約では出力を1文に限定すると文生成の精度が向上した
- 課題
  - 長文データへの対応
    - BERT以外の長文対応モデルの試行
    - 長文を複数ブロックに分けてBERTで処理する方法の検討
  - 精度向上
    - 特に抽象要約については他のモデルの試行も必要

EXA CORPORATION

おわりに

おわりに

- 本セッションではBERTを使った類似文検索、文書要約の試行についてご紹介しました
- 今後もBERTをはじめAI分野の知見収集・ソリューション開発にチャレンジしていきます
  - 疑問・アイデアなど何でもご相談ください！
  - 一緒にチャレンジしましょう！
  - BERTハンズオン研修もご用意していますのでぜひご活用ください！

EXA CORPORATION