

ORACLE

自然言語と機械学習

~自然言語の基本処理からBERTのファイン・チューニングまで~
2020年11月30日

園田憲一

日本オラクル株式会社

アジェンダ

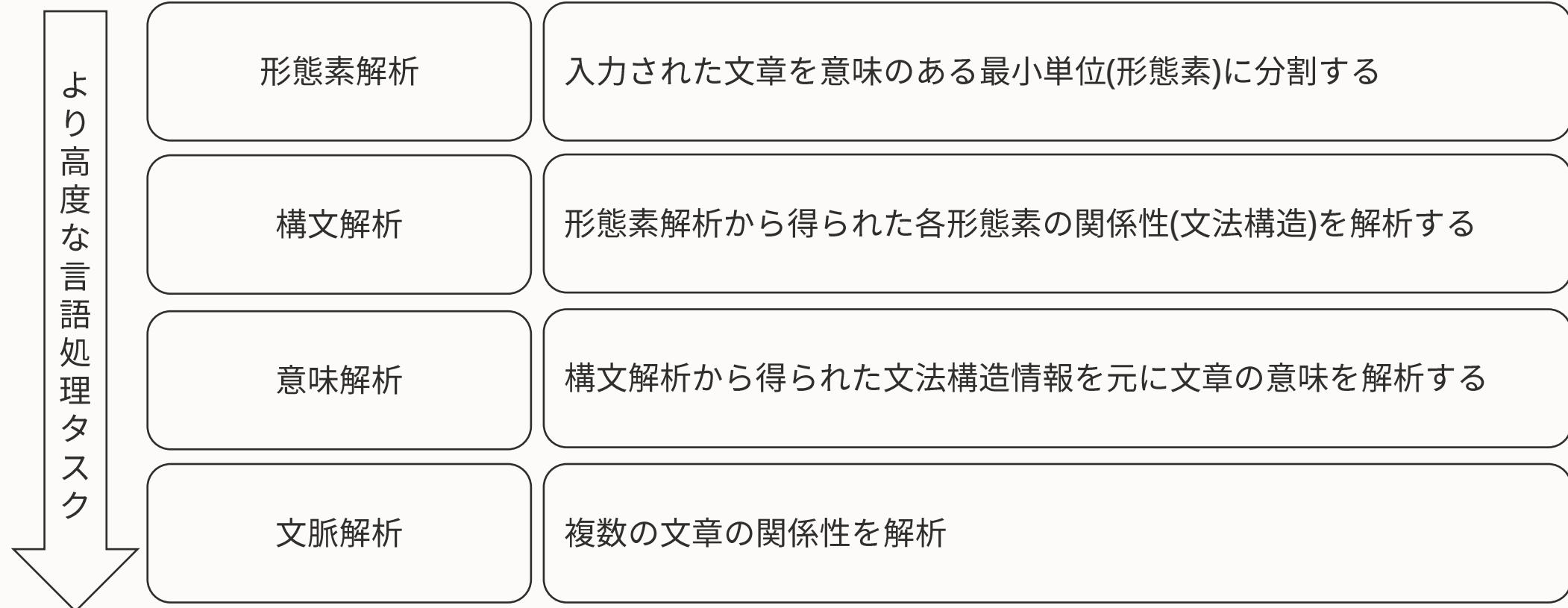
- ユースケース
- 自然言語処理の技術要素
- 自然言語の機械学習ワークフロー
- BERT(概要、ファインチューニング、GPUの効果)
- まとめ

自然言語処理のユースケース



自然言語処理 = コンピュータに言葉を理解させる

自然言語処理の技術要素





形態素解析

辞書や文法ルールに基づいて、文章を「意味を持つ最小の単位（＝形態素）」に分割し、各形態素に品詞などの各種情報を付与する。

例文) 太郎は花子と珍しい動物を見るために動物園に行きました。

形態素解析ライブラリ
(MeCab/Juman/Janome/etc.)

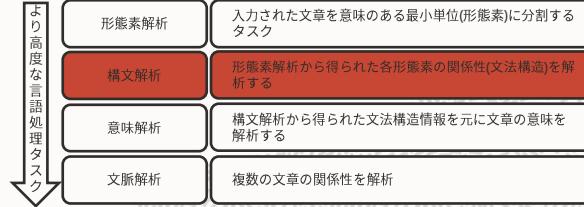
単語分割

品詞情報付与

太郎	は	花子	と	珍しい	動物	を	見る	ため	に	動物園	に	行き	まし	た
名詞	助詞	名詞	助詞	形容詞	名詞	助詞	動詞	名詞	助詞	名詞	助詞	動詞	助動詞	助動詞

文章中にある形態素をデータとして抽出する

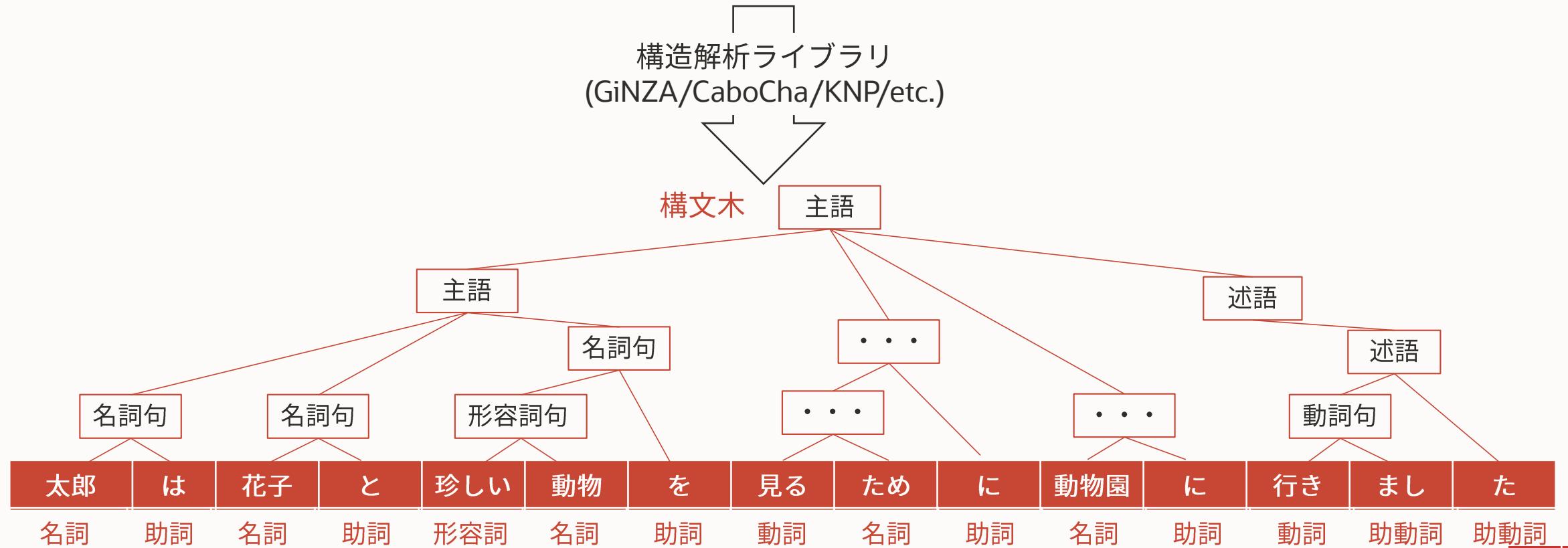




構文解析

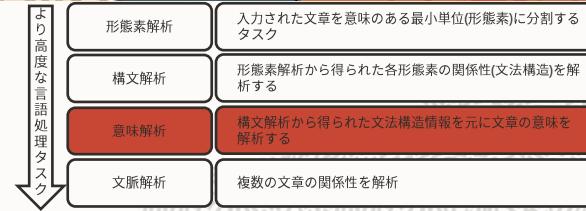
抽出された形態素間の関係性を解析し、文章の構造(主に文法構造)を明確にする。

例文) 太郎は花子と珍しい動物を見るために動物園に行きました。



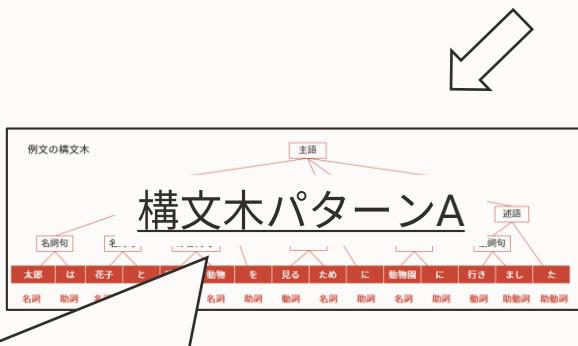
デモンストレーション 形態素、構文解析の基本処理

意味解析

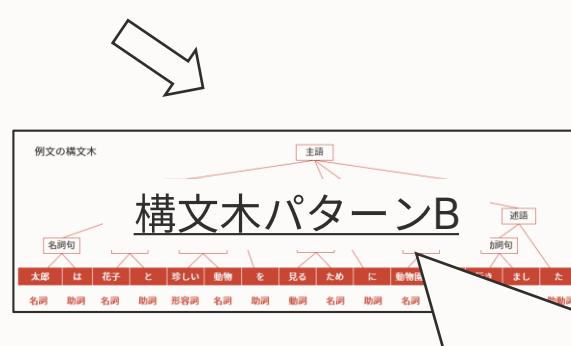


辞書データをもとに、単語と単語の関連性の強弱により、正しい構文木を選択する

例文) 太郎は花子と珍しい動物を見るために動物園に行きました。



太郎と花子は「珍しい動物」を見るために一緒に動物園に行った

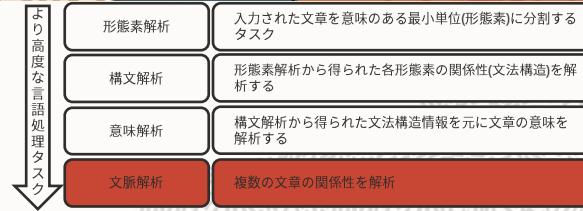


太郎は、「花子」と「珍しい動物」の両方を見るために動物園に行った

- ・ 「太郎」と「花子」は関連性が強い(両方ともに名詞かつ人名)
 - ・ 「花子」と「珍しい」は関連性が弱い(不適切な形容詞)
 - ・ 「珍しい」と「動物」は関連性が強い(適切な形容詞)

文脈解析

複数文章の文章同士の関係性を解析する



太郎は花子と珍しい動物を見るために動物園に行きました。

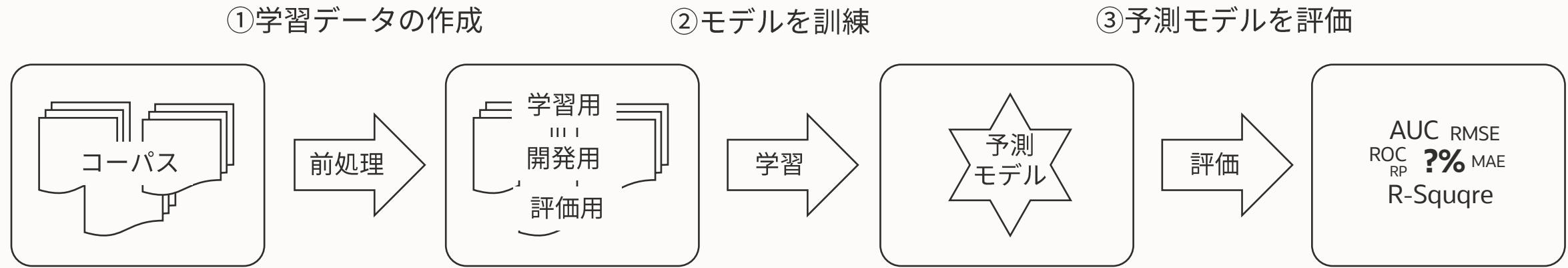
彼らはそこでウナギイヌを見つけました。

2つの文章の
関係性は？

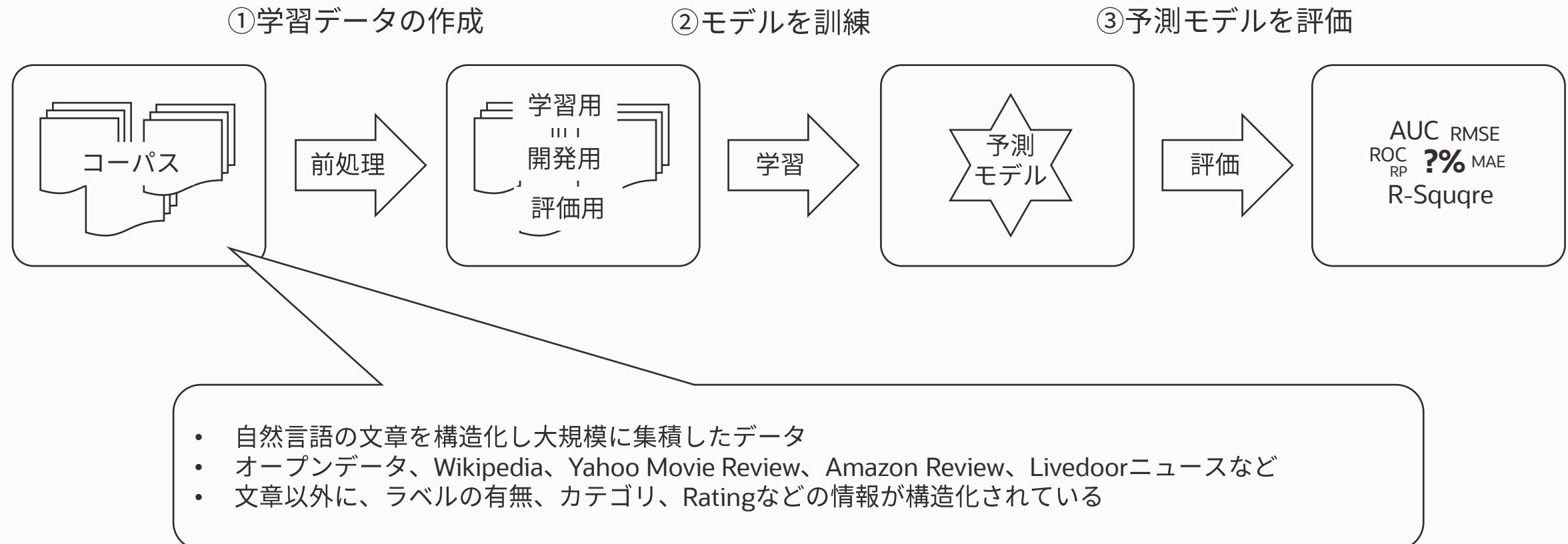
「彼ら」とは
誰？

「そこ」とは
何処？

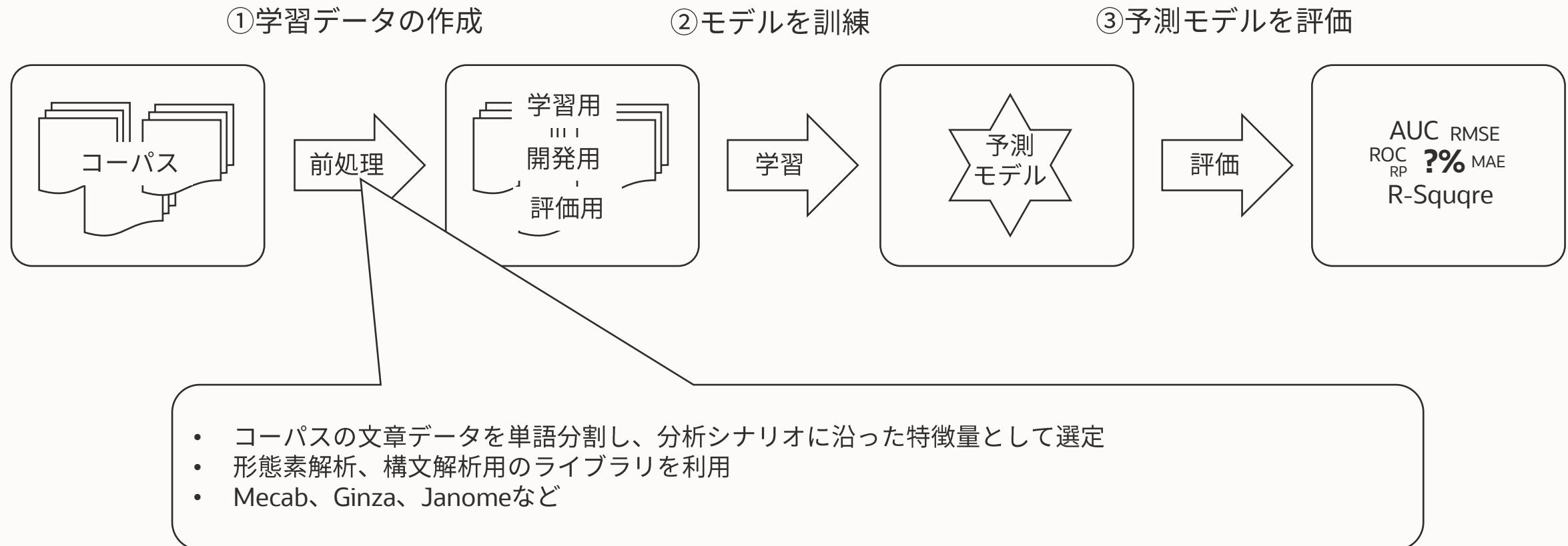
自然言語処理における機械学習のワークフロー



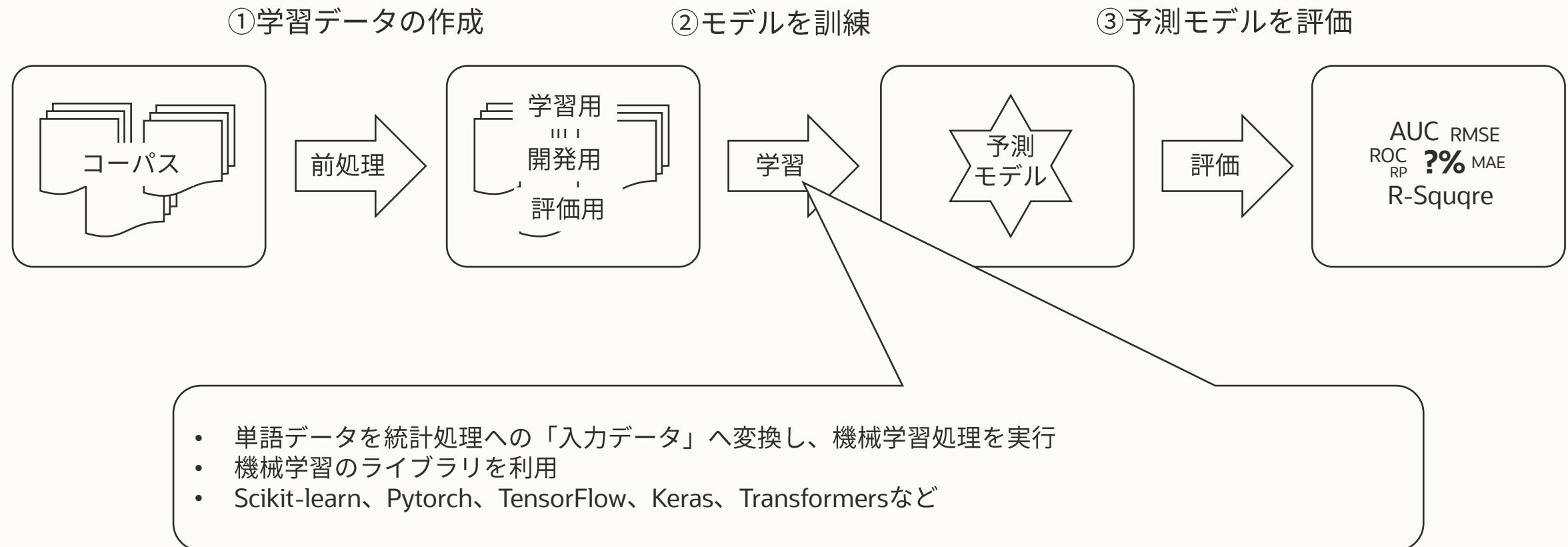
自然言語処理における機械学習のワークフロー



自然言語処理における機械学習のワークフロー

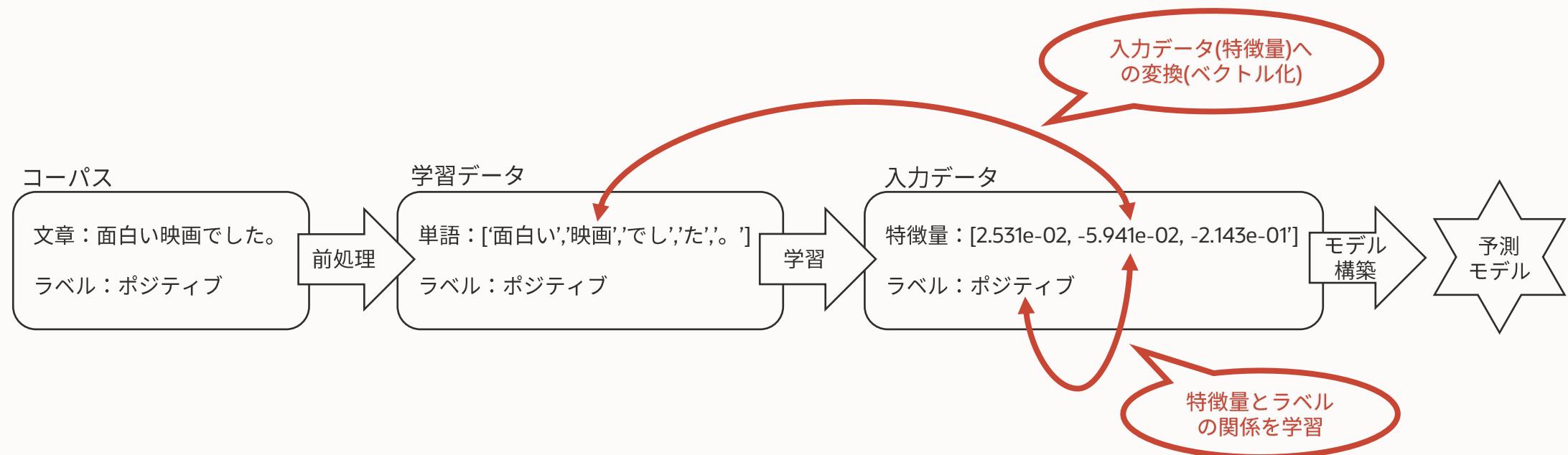


自然言語処理における機械学習のワークフロー



学習処理の概要

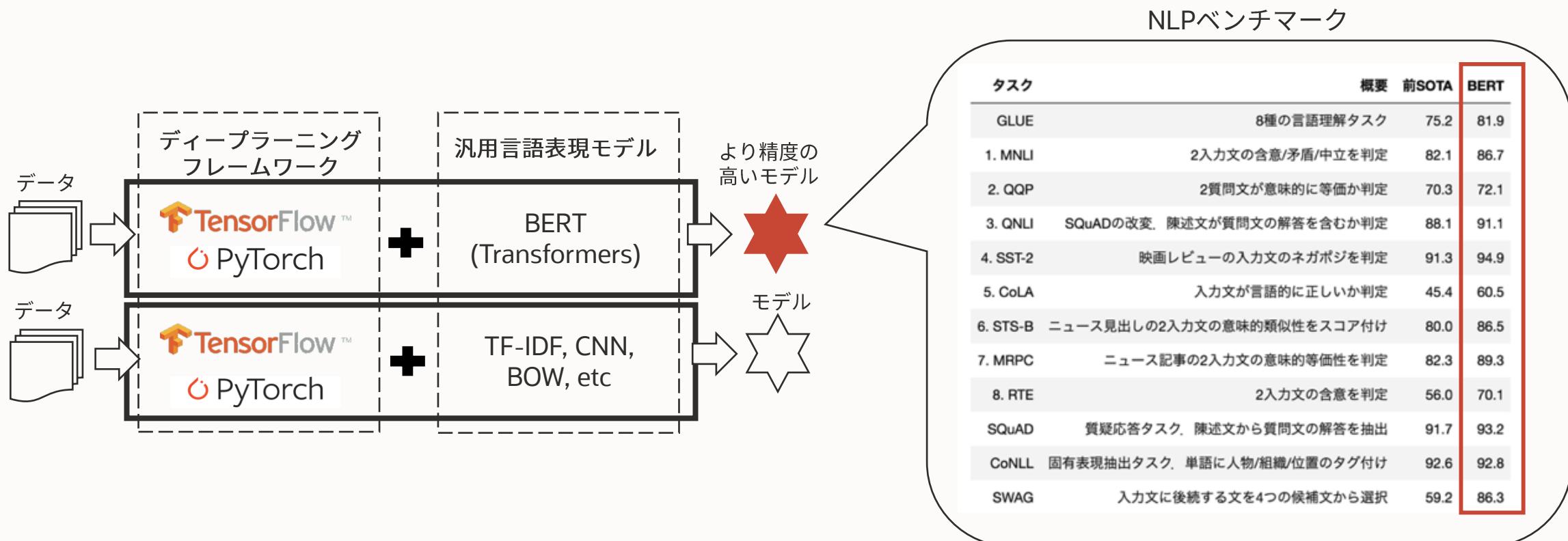
映画レビューのコメントがポジティブかネガティブかを判別する分類問題の場合



特徴量のベクトル化および学習時に一般的な手法(TF-IDFなど)ではなく、
BERTを利用することで予測モデルの精度(言語認識の精度)が飛躍的に向上

Bidirectional Encoder Representations from Transformers(BERT)

- 最先端の汎用言語表現モデル
- TensorFlow、PyTorch用の「Transformers」に実装、既存のMLライブラリと変わらないプログラミングモデル
- 2018年10月、当時のNLPベンチマークにおいてSoTAを大幅に更新



BERTモデル特有の言語学習法

下記2つのタスクに着目して学習データから言語学習を行い汎用表現言語モデルを構築

Next Sentence Prediction(NSP)

2つの文章が連続した(関係のある)文章かどうか？

Masked Language Model(MLM)

文章中の複数箇所をマスクし、その部分に入る単語が何になるか？



BERTモデル特有の言語学習法

下記2つのタスクに着目して学習データから言語学習を行い汎用表現言語モデルを構築

Next Sentence Prediction(NSP)

2つの文章が連続した(関係のある)文章かどうか？

太郎は花子と珍しい動物を見るために動物園に行きました。

彼らはそこでウナギイヌを見つけました。

2つの文章の関
係性は？

Masked Language Model(MLM)

文章中の複数箇所をマスクし、その部分に入る単語が何
になるか？

太郎は花子と珍しい動物を見るために動物園に行きました。

[Mask]は[Mask]でウナギイヌを見つけました。

「彼ら」とは
誰？

「そこ」とは何
処？

単語が文章中に現れる前後のコンテキストに応じてモデルが単語の意味を決定する、
双方向性(Bidirectional)の学習手法



BERT: デモンストレーション

BERT事前学習済モデルを使った単語予測

太郎は動物園に行った。そこで珍しい動物を見つけた。

形態素解析
↖ ↗

太郎 | は | 動物 | 園 | に | 行っ | た | 。 | そこ | で | 珍しい | 動物 | を | 見つけ | た | 。

特定の単語をマスク
↖ ↗

太郎 | は | 動物 | 園 | に | 行っ | た | 。 | [Mask] | で | 珍しい | 動物 | を | 見つけ | た | 。

学習済みモデル(BERT)でマスクされた単語を予測
↖ ↗

[Mask] = [そこ]



デモンストレーション BERT事前学習済モデルを使った単語予測

BERT: ファインチューニング

概要

- 事前学習済みモデルを流用し、追加学習のみで精度の高いモデルを生成する手法

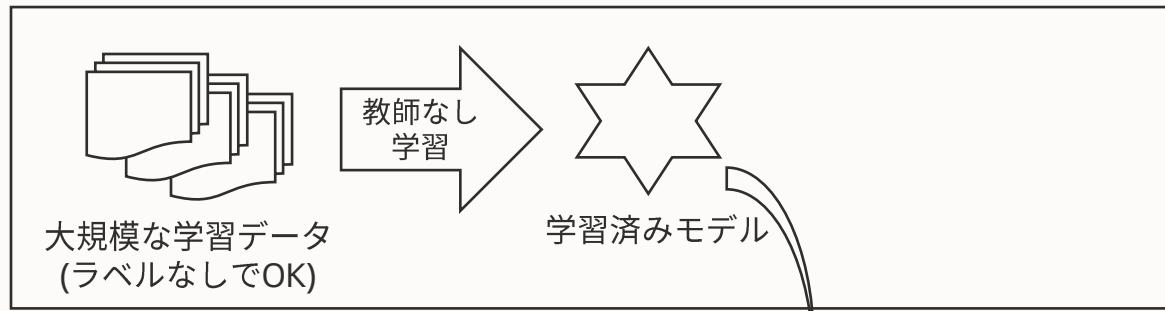
メリット

- 事前学習済みモデルを流用できるため、短時間の追加学習のみで最終的なモデルが生成できる
- コンピューティング・リソースのコストも削減できる
- 追加学習のラベルありデータは少量でもモデルの精度確保が期待できる

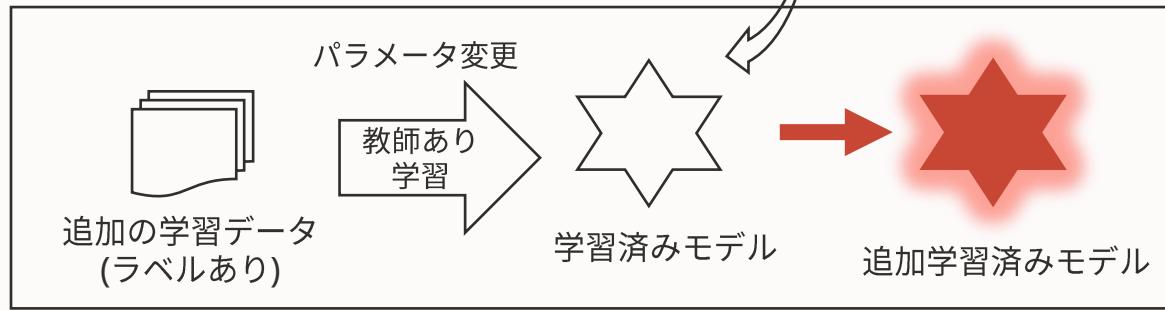
デメリット

- ファインチューニングをしない場合と比較して、モデルの精度が落ちる場合もある

事前学習(教師なし)



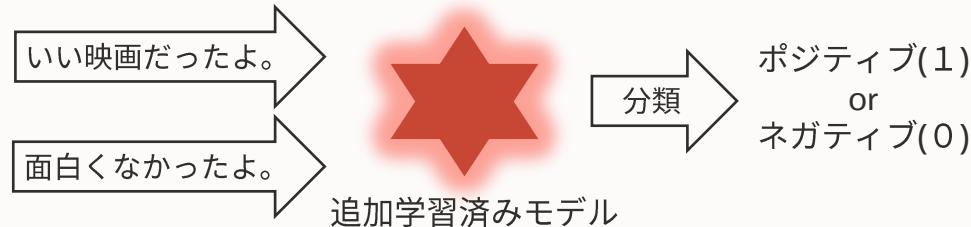
ファイン・チューニング(教師あり)



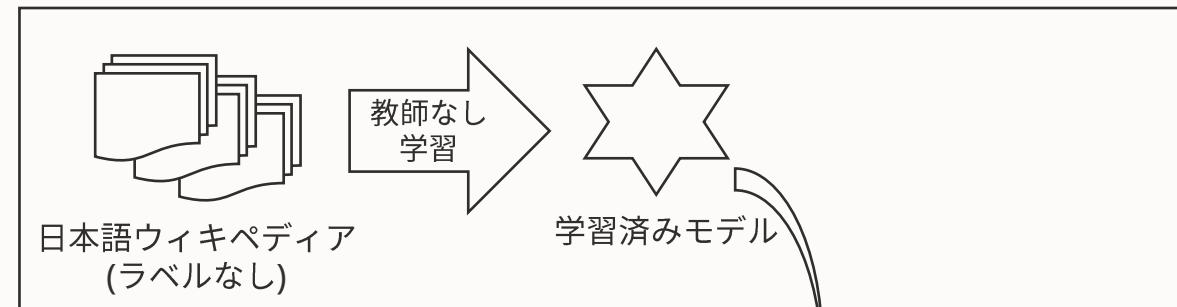
BERT: フайн・チューニングのデモンストレーション

概要

映画レビューコメントのポジネガ分析(バイナリ分類)



事前学習(教師なし)



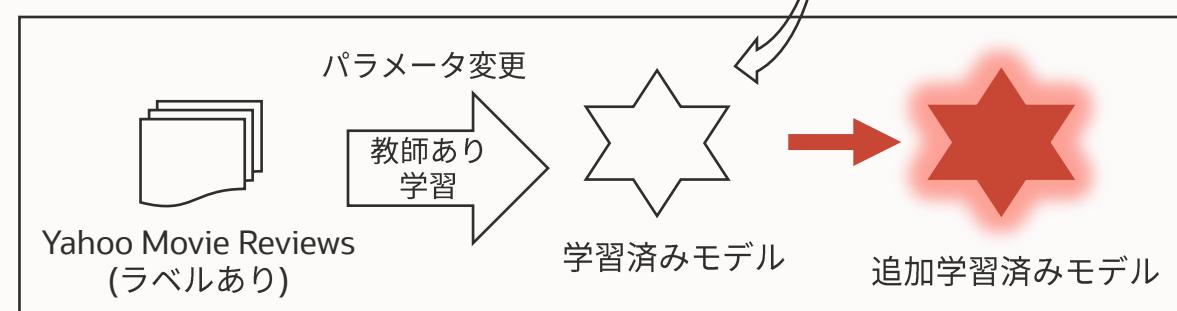
- 事前学習済みモデル

- 東北大学幹・鈴木研究室(<https://github.com/cl-tohoku/bert-Japanese>)
- BERT分類モデル(bert-base-japanese-whole-word-masking)
- コーパス：日本語ウィキペディア

- ファイン・チューニング(追加学習)

- コーパス：Yahoo Movie Reviews
- レビュー数：10000
- レビューの文章：平均約300文字/コメント
- ラベルの割合：ポジティブ5038件、ネガティブ4962件

ファイン・チューニング(教師あり)



BERT: フайн・チューニングのデモンストレーション

追加学習のコード概説

```
#ライブラリのimport
from toiro import classifiers
from toiro import datadownloader

#追加学習のコーパスのダウンロードして定義
corpus = 'yahoo_movie_reviews'
datadownloader.download_corpus(corpus)

#学習データ、開発データ、テストデータに分割
train_df, dev_df, test_df = datadownloader.load_corpus(corpus, n=12500)

#追加学習の分類モデルを定義
model = classifiers.BERTClassificationModel()

#追加学習の分類モデルを訓練
model.fit(train_df, dev_df, verbose=True)

#予測
text = "いい映画だったよ"
pred_y = model.predict(text)
print(pred_y)
1
```

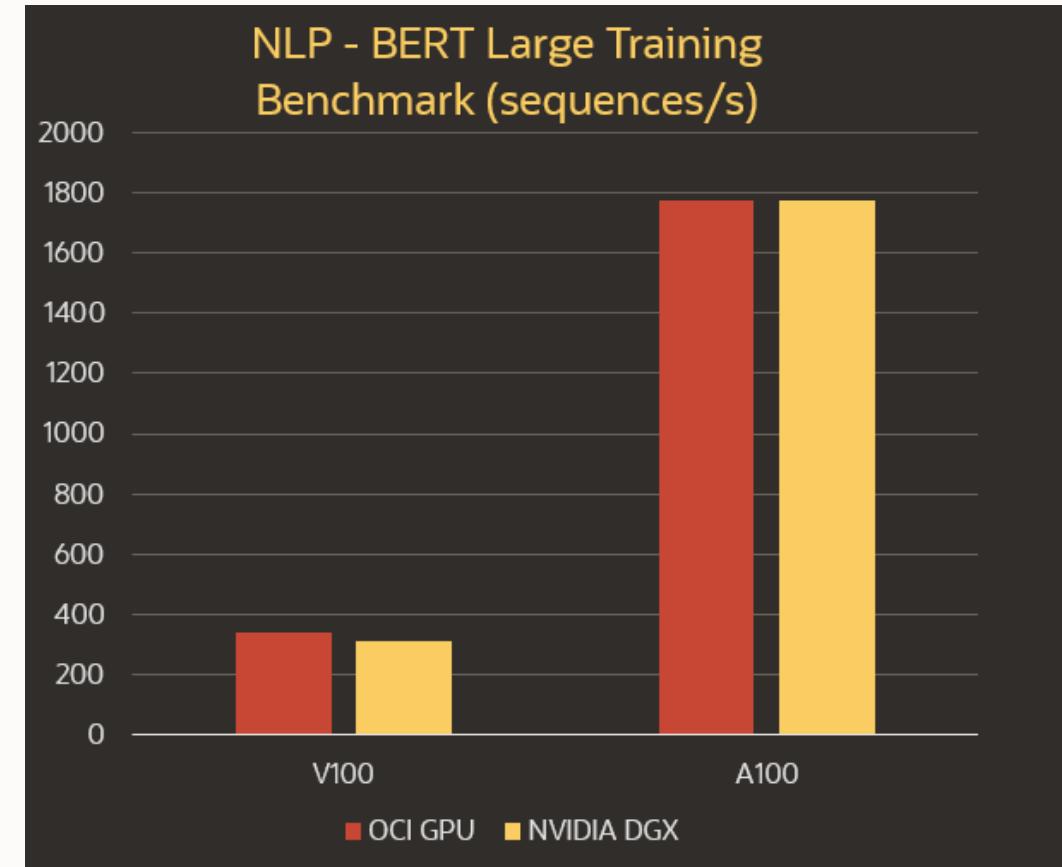
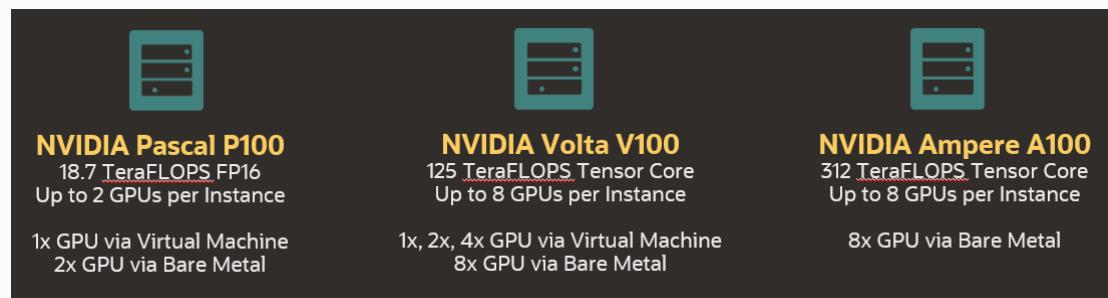
<div style="border-radius: 50%; width

デモンストレーション BERTファイン・チューニング 映画レビューコメントのポジネガ分析

Blog : A practical guide to getting started with Natural Language Processing

Oracle Cloudで利用可能！

- Nvidia GPU A100
 - V100の後継モデル
- OCI Compute Service BM.GPU4.8で利用可能
 - CPU 64コア
 - GPU A100 x8



[Blog : A practical guide to getting started with Natural Language Processing](#)

まとめ

- ・ 自然言語処理の主要タスクは？
 - ・ 形態素解析、構文解析、意味解析、文脈解析
- ・ 意味解析、文脈解析において機械学習が有効！
- ・ 機械学習の「学習」とは？
 - ・ 単語を特徴量としてベクトル化し、それを統計処理にかけること
- ・ BERTモデルはなぜ言語認識の精度が高いのか？
 - ・ 新しい学習の仕組みが入っているから
 - Next Sentence Prediction、Masked Language Model
- ・ ファインチューニングによって何が嬉しいのか？
 - ・ 少量のラベル付きデータでもモデルが作れる
 - ・ 学習時間、コンピューティングリソースなどのコスト削減

参考図書

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
<https://arxiv.org/pdf/1810.04805.pdf>
- 東北大学 幹・鈴木研究室 Pretrained Japanese BERT models released / 日本語 BERT モデル公開
<https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/>
- 京都大学 黒橋・褚・村脇研究室 BERT日本語Pretrainedモデル
http://nlp.ist.i.kyoto-u.ac.jp/index.php?ku_bert_japanese
- Hugging Face
<https://github.com/huggingface/transformers>
- 自然言語処理ライブラリtoiro
<https://github.com/taishi-i/toiro>
- A practical guide to getting started with Natural Language Processing
<https://blogs.oracle.com/cloud-infrastructure/a-practical-guide-to-getting-started-with-natural-language-processing>
- 機械学習に使える日本語データセットまとめ
<https://lionbridge.ai/ja/datasets/japanese-language-text-datasets/>



セミナーでいただいたご質問の回答(1)

Q: フайнチューニングのデモは時間がかかるとのことでしたが、どのくらいの処理時間だったのでしょうか。

A: CPU(Xeon 24コア)インスタンスで2時間以上かかっていたため、GPU(V100 x1)インスタンスに変更したところ10分程度で学習完了しました。

Q: 先ほどのデモのnotebookってシェアしていただけますでしょうか？

A: 下記Githubのリポジトリで公開させていただきます。

<https://github.com/oracle-japan/oci-datasience-nlp-demo01.git>

Q: フайнチューニング時の学習は教師ありで説明されておりましたが、教師なし学習でも可能でしょうか？

A: 可能です。

Q: BERTのデモでは穴埋め問題を解いていらっしゃいましたが、他にはどんな問題を解けるのでしょうか？

A: 文章分類や類似文章検索などユースケースのページでご紹介した様々な問題に対応できます。

Q: CloudでGPUインスタンスを選ぶ際に色々な要素 (GPU数, GPUメモリ, CPU数, 価格 等)があると思うのですが、選び方のベストプラクティスがあれば教えていただけないでしょうか？

A: 主にサイジングの話になると思いますが、これは既存の処理を実行したときの性能情報がないと難しいと思います。例えば既存でNvidia V100で実行したときの性能情報をベースにNvidia A100ではどこれくらいのコア数が必要かを推測するという方法になるかと思います。

Q: 今回BERTのデモでは、分類を見せていただきましたが、追加でQ&Aを学習させると、質問文を入力して回答文を生成するような文章生成ができるのでしょうか？

A: はい。文章生成は自然言語処理の中では頻繁に開発されているユースケースです。



セミナーでいただいたご質問の回答(2)

Q: 自然言語処理を用いて文章の要約をしてみようと思っております。おススメのモデルなどご存知でしたら教示お願い致します。

A: BERT関連ですと「BertSum」があります。英語ソースになりますが下記にURLをご紹介します。BertSumで検索すると他にも沢山情報がありますのでそちらをみていただいてもよいかと思います。

論文：<https://arxiv.org/pdf/1903.10318.pdf>

github：<https://github.com/nlpyang/BertSum>

Q: フリーでダウンロードできる、BERTのような軽量な日本語モデルはありますか？

A: BERT以外ですと、XLNet、RoBERTa(BERT改良版)、GPT2、ALBERTなどが有名です。





ORACLE