自然言語処理
―準備、サブワード―

https://satoyoshiharu.github.io/nlp/

サブワードの位置づけ

- 日本語を処理する場合、処理単位へ分割する「語分ち」が必須となります（語分かちした結果を分かち書きといいます）。主に形態素解析がその役割を果たしますが、サブワードという異なるアプローチもあります。以下で、サブワードの技術に触れます。

自然言語処理：形態素解析
サブワード

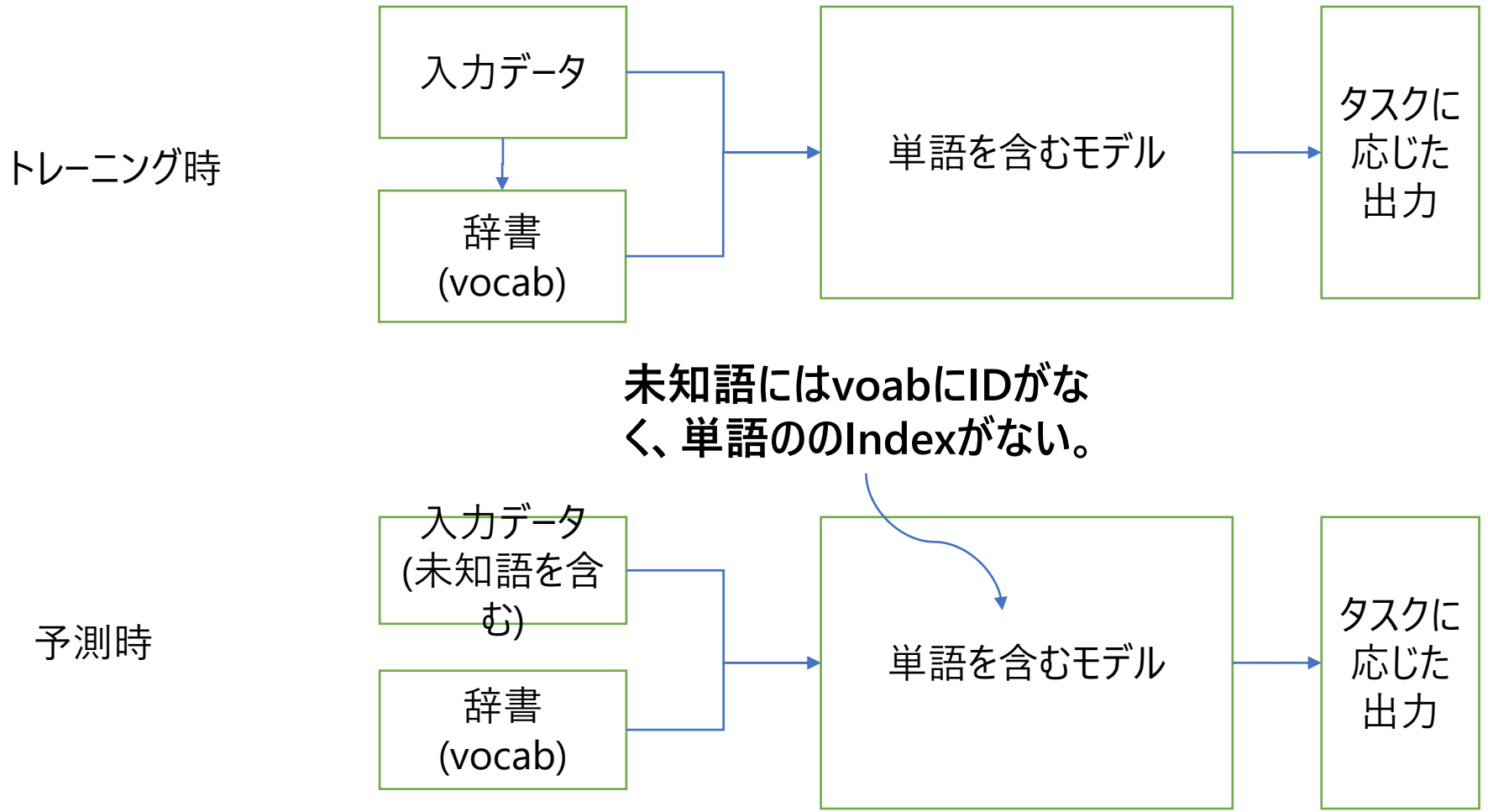
解説動画

https://yo-sato.com/

形態素解析（分かち書き）の問題点

- 大規模語彙を扱うのは計算負荷が大きい。
- 新しいデータには、これまで処理して作成した辞書に未登録の**未知語**がある。UNK (Unknown word)トークンと置き換えるなどの特殊処理が必要だ。

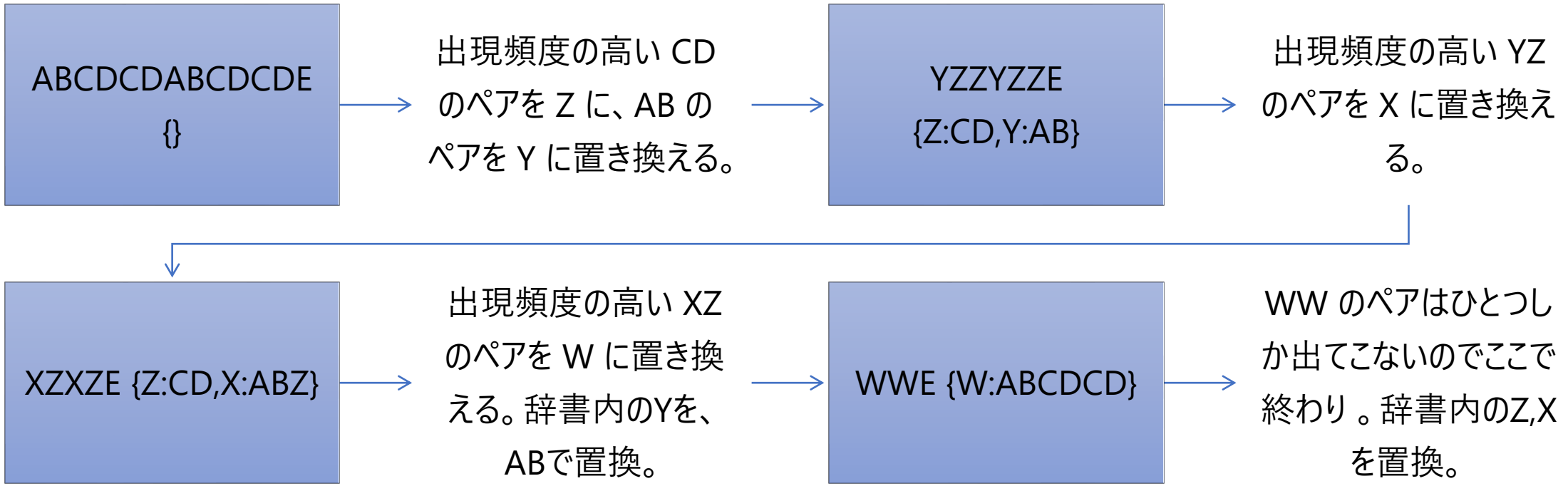
未知語があるとエラーになる



サブワードの基本的な考え方

- 高頻度語は、単語として扱う。
- 低頻度語は、文字や部分文字列を、単語であるかのように扱う。

実装の原理：Byte Pair Encoding
共通パターンを辞書登録して、情報圧縮



BPEの応用

- 英語など単語区切りが明確な言語
  - 単語やその部分をベースに、高頻出パターンを拾い、単位としてまとめていく。残りは文字を単位とする。
- 日本語など単語区切りがない言語
  - 文（文字列連鎖）から、高頻出パターンを拾い、単位としてまとめていく。残りは文字を単位とする。(SentencePiece)
    - 参考、https://www.pytry3g.com/entry/how-to-use-sentencepiece, https://qiita.com/taku910/items/7e52f1e58d0ea6e7859c

メリットによる使い分け

形態素解析

- 品詞や読みなどの負荷情報が得られる
- 分かちの長さが予測しやすい
- 未知語に配慮する必要がある。

サブワード

- 未知語処理に悩まされない
- 単語数上限を決められる

注

- サブワードで、分ちの長さが見通しにくいという点は、意外と悪さします。
- ニューラルネットアプリが、ある種の表現に弱いので、トレーニングデータで類似の表現を追加したとします（データAugmentation）。
- サブワードは高頻出パターンを単語とみなしますので、その種類の表現が全体で1単語とみなされてしまいます。
- ニューラルネットでは、単語ベクトルというものに、その単語が使われる文脈を仕込むのですが、長い単位で一単語とみなされるため、その単語ベクトルにコンテキストの学習が十分行われなくなります。
- 結果、トレーニングデータを追加しても効果が上がらず、逆効果となることも。

課題：サブワード

- sentencepiece.ipynb に、サブワードを実装したSentencePiece を動かすコードサンプル（https://www.pytry3g.com/entry/how-to-use-sentencepiece のまんまです）を入れました。コードを読んで、動かしてみましょう。最後のところに、別の文章を設定する行を追加し、実行してください。

確認クイズ

- ありません。