# 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/

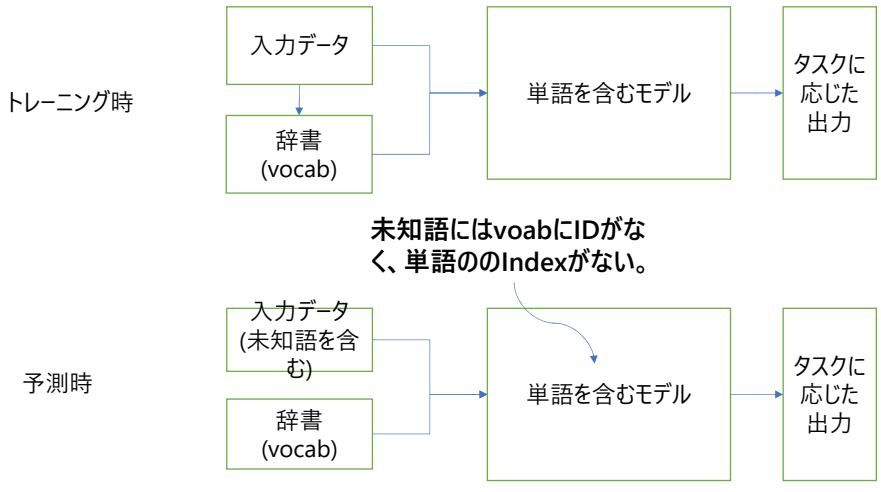
ここでは、形態素解析の発展テーマとして、サブワードという考え方を紹介します。

# 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。

辞書を使った形態素解析には、二つの問題点があります。
一つ目は、単語数が数百万オーダーなので、計算負荷・メモリ負荷が大きいこと。
二つ目は、未知語の問題です。

# 未知語があるとエラーになる



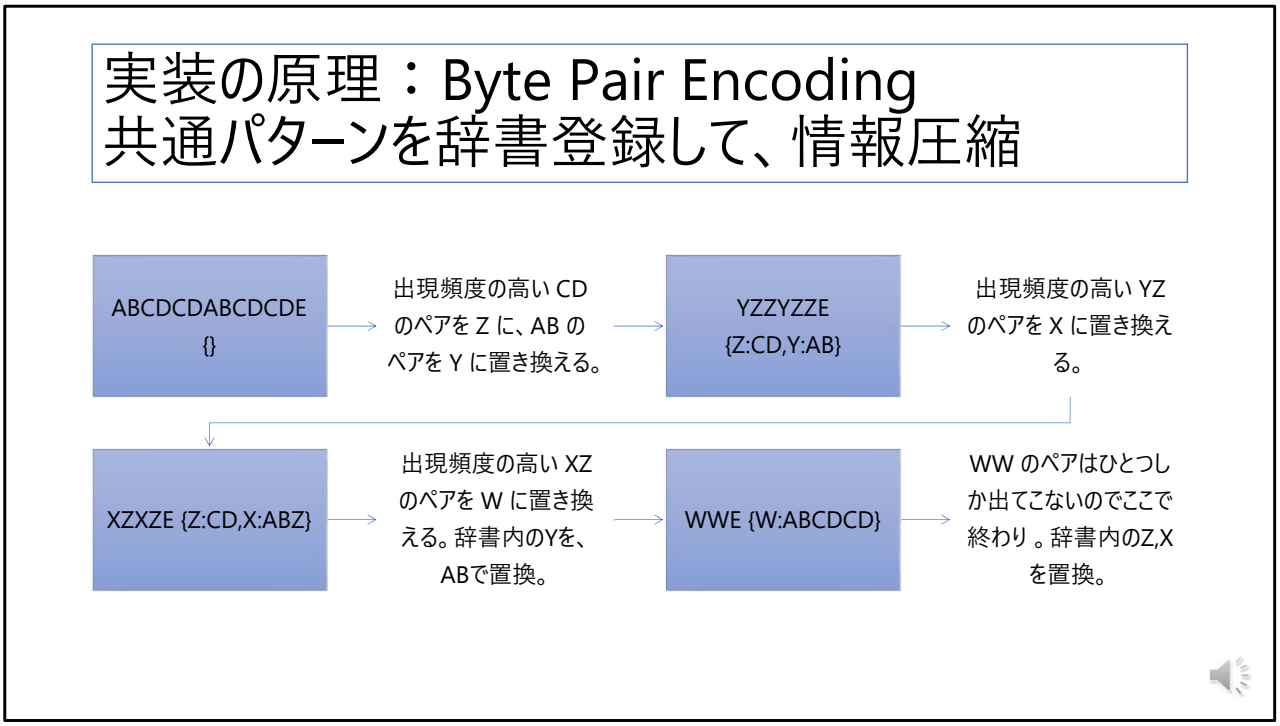

システムをトレーニングする際、トレーニングに使ったテキストに含まれる単語、あるいは形態素解析エンジンの使う辞書に登録された単語、は、運用時でも問題なく処理されます。
しかし、運用時にトレーニング時になかった単語、あるいは形態素解析エンジンが知らない単語が、入力データにあると、特殊な処理を入れておかないとエラーになったりします。
日本語の場合、特に人名などの固有名詞はすべてを網羅して辞書に持つわけにいきません。そのため、特殊な未知語処理が必要となります。

# 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。

単語に比べて、文字ならば、日本語の場合3000くらいで、抑えることができます。
単語の処理時に道後の問題が存在するので、文字を併用しようというアプローチがあります。辞書に頼らずに、高頻度な文字列を単語とし。低頻度な文字列は文字ごとに単語扱いしてしまいます。

/*
なお、高頻度語でも、複合語は分割したほうが単語数を抑えられる。「体系、体系的、本質、本質的、合理、合理的」は、それぞれを単語とみなすよりも、接尾語を分離し、「体系、的、本質、合理」という単語からなるとしたほうが、語数が抑えられる。
*/



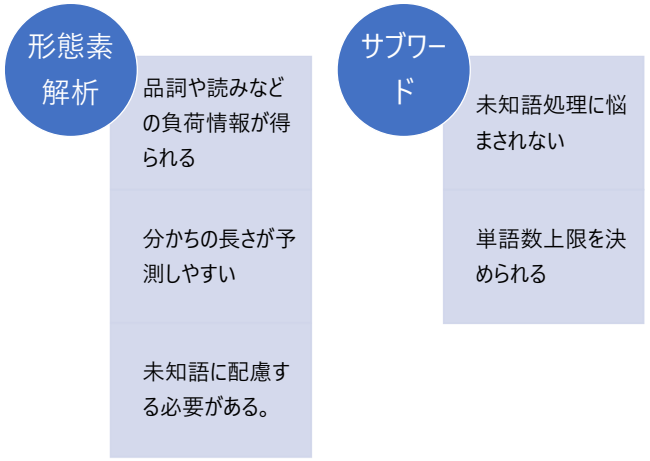

単語よりも短い文字レベルから、文字列を分割するアルゴリズムとして、BPEがあります。スライドにサンプルデータトレースを載せました。
…

# 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

このようなBPEの考えを応用し、ニューラルネット処理の処理単位にするというアプローチがあります。
英語など、単語の間に空白があるケースでは、単語の語幹にあたる部分以外にedとかingとか部分文字列にパターンがあります。それら高頻出パターンを拾っていって、残りを文字のまま扱います。
日本語など単語区切りがない場合、より単純で、文字列から高頻出パターンを拾い、単位としてまとめていき、残りは文字を単位とします。

## メリットによる使い分け




辞書を使った形態素解析は、品詞や読みなどありがたい情報が付加的に得られる、というメリットがあります。
サブワード的なアプローチには、未知語処理に悩まされることがないというメリットがあります。
また、検出する高頻出単語数を設定できるので、処理単位となる単語数を低く抑えられるというメリットもあります。
品詞が欲しいアプリでは形態素解析、そうでない場合はサブワード、という使い分けがいいかもしれません。

# 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。