自然言語処理
―準備、形態素解析―

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

形態素解析および100本ノック第4章の位置づけ

- 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）。形態素解析は、その語分ちを提供してくれます。
- 以下で、形態素解析とニューラルネットとの関係を説明します。スライドだけだとわかりにくいので動画をリンクしています。動画のしゃべり原稿はpptのノートに入れてあります。
- 形態素解析はすでに優秀なエンジンがいくつもフリーで利用でき、プログラミング的にはそれらを呼ぶだけです。100本ノックの第4章は、形態素解析エンジンの出力をいろいろ加工してみるという課題になっています。

自然言語処理

形態素解析とは？

解説動画

https://yo-sato.com/

形態素解析について説明します。

形態素とは？

表層の音：からすがきた

形態素列　　表層のテキスト：カラスが来た。

形態素
単語は表層で多様な形態をとる

からす
カラス
烏

来[ク]る
来[コ]
来[クレ]
来[キ]
来[コ]い

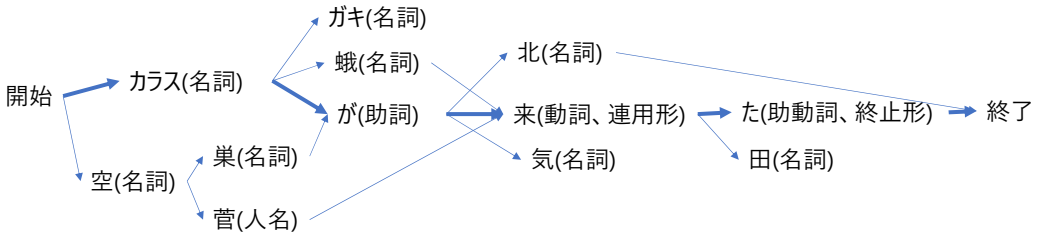
単語　　“カラス”　　“来る”

形態素というのは、単語が実際に使われるときに表層のあらわれた形態です。
カラスがきたという表層の文を例に見てみます。
カラスは、カタカナだったり、漢字だったりの形態をとります。
来るという単語は、使われるコンテキストによって活用し、クるだったり、コだったり、…
このように、我々が見聞く表層の言語は、単語がいろいろな形態をとって、つながっています。
これらの単語がいろいろな形態をとっているものを、形態素といいます。
表層の文を分析するときは、まずこの形態素を相手にします。

形態素解析とは？

形態素解析は、表層の形態素の列から、単語の活用、表記、送り仮名などの、単語の派生タイプを明らかにしつつ、最も確からしい単語列を推定すること。

表層の音：からすがきた

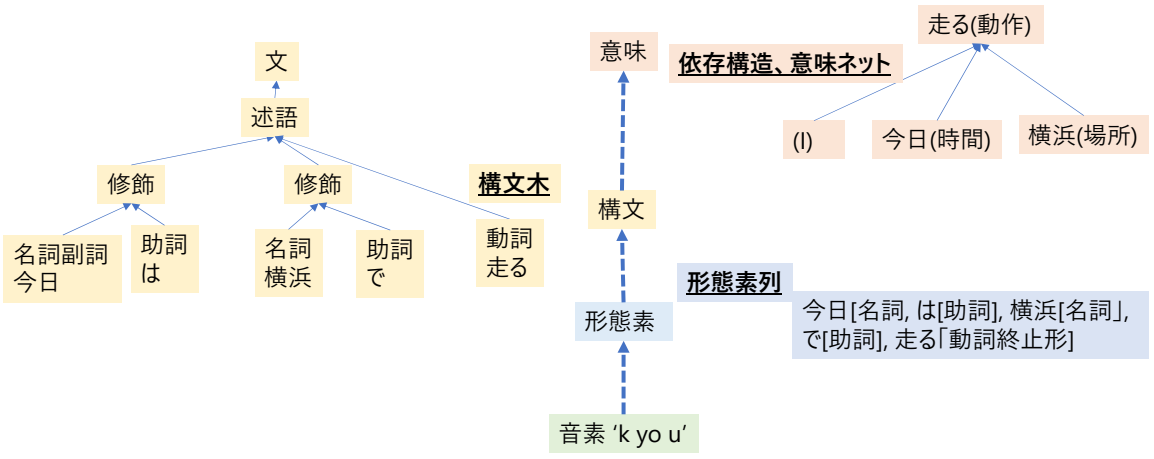


表層の文は形態素の列です。
形態素解析というのは、表層の形態素の列から、単語の活用、表記、送り仮名などの、形態素の派生タイプを明らかにしつつ、最も確からしい単語列を推定することです。

形態素解析は、自然言語処理の中で、どう利用されているか？

ここで、形態素解析は、自然言語処理の中で、どのように利用されているのかを見ます。

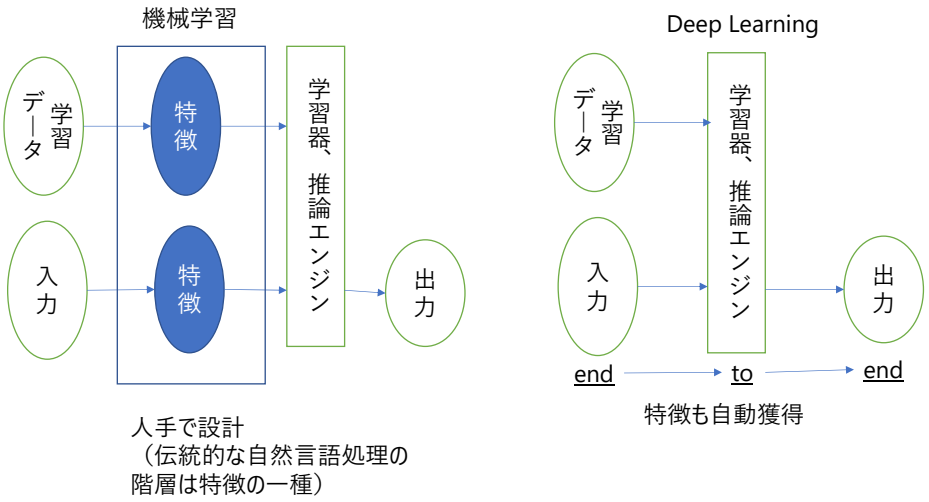
伝統的な自然言語処理



特徴(Feature)の階層

伝統的な自然言語処理は、処理単位が下位から上位へと階層をなしています。音声認識の場合は、最初に音素があります。それが組み合わさって形態素となります。テキストから処理をする場合は、最下層は形態素です。形態素が組み合わさって、構文となります。そして、構文が意味的な構造を構成します。
このように、形態素は階層の下のほうにあって、解析の初期段階に処理されます。

ニューラルネットで、人が特徴を設定するのでなく、それも学習するEnd-to-Endへ

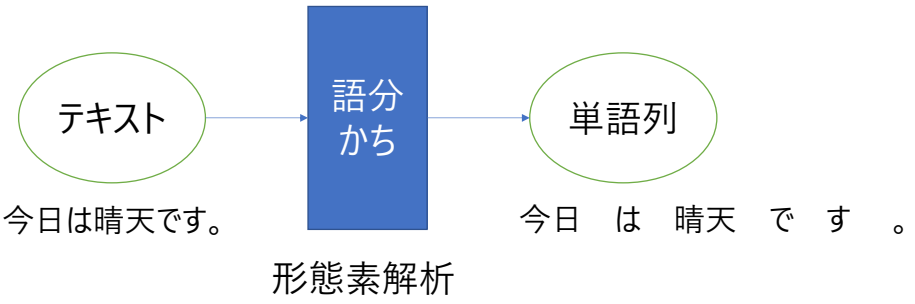


一方、機械学習やニューラルネットが登場しました。

機械学習では、画像を特徴づける特徴量を人が設計します。ニューラルネットでは、それを機械が決めることが、特色とされます。

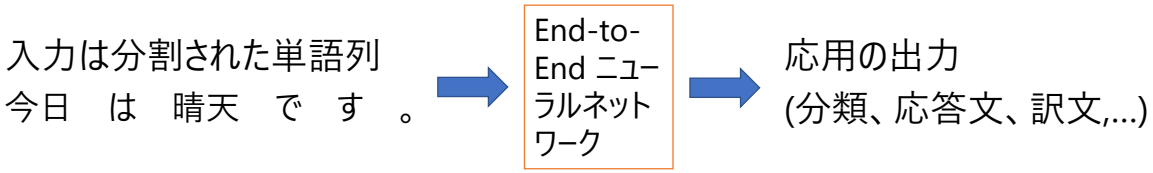
自然言語処理にも、ニューラルネットが応用されるようになりました。ここで、ニューラルネットでは、入力から、望む出力を一気に計算します。End-to-Endと呼ばれます。入力から形態素、構文・意味とレイヤを上ってあるいは下って、情報を変換するのでなく、入力から出力へ一気にニューラルネットで変換します。

形態素解析の結果、単語の区切りが得られる



ニューラルネットで自然言語処理をする場合でも、形態素解析を利用します。
形態素解析の結果、単語の区切りが得られます。

語分ちは、End-to-end のニューラルネットでも利用する



その単語に区切られたものを、分かち書きと言い、単語に分割する処理を語分ち処理と言います。
形態素解析で、語分ちを行います。それが、End-to-Endニューラルネットへの入力となります。

形態素解析のほかの用途

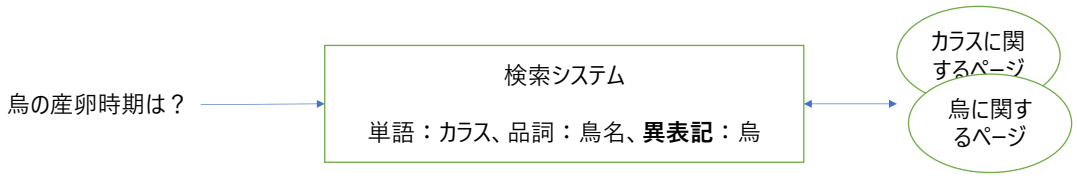
ここで、形態素解析のほかの用途を見ておきます。

エンティティの抽出

形態素解析の結果、品詞という構文的・意味的な情報を得られる。

例えば、WEBで対話アプリを組むとき、地名は地名として認識する、人名は人名として認識する、動詞から意図を解釈する、などを行います。形態素解析の結果、品詞が得られるので、そこに役立ちます。品詞判定でわかる単語のうち、人名、地名など、世界に実在するものを、エンティティといいます。それを抽出することを　NER（Named　Entity　Recognition）などといいます。音声対話システムを組むときに、エンティティという用語が出てくるので、覚えておきましょう。

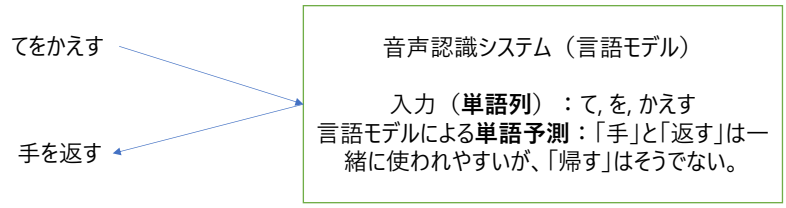
異表記の認識



形態素解析の結果、辞書に異表記情報があれば、異表記単語の処理結果を統合できる。

また、検索では、「物事」と「ものごと」は、表記が異なる同単語であるといったことを踏まえた処理をします。そこで、形態素解析で、ある単語が表層にどのような形態で現れたかを認識する形態素解析が利用されることがあります。

言語モデルによる単語予測



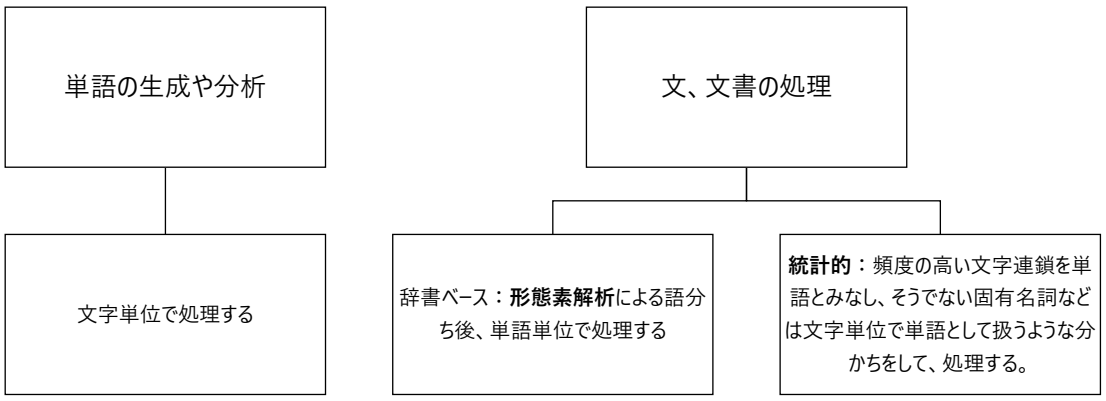
分かち書きをした結果、単語列をニューラルネットに覚えさせれば、コンテキストから単語予測ができるようになる。

また、音声認識や、機械翻訳では、ある単語がその周辺の単語コンテキストでどれだけ現れやすいかという情報（言語モデル）を使って、同じ発音なのに異なる単語である同音語を識別したり、次に来る単語を決めたりします。言語モデルは、形態素解析した単語列をニューラルネットに入力して作ります。

多様な処理単位

形態素解析は、End-to-endニューラルネットでも必要な処理だという説明をしましたが、実は、いつもそうするとは限りません。

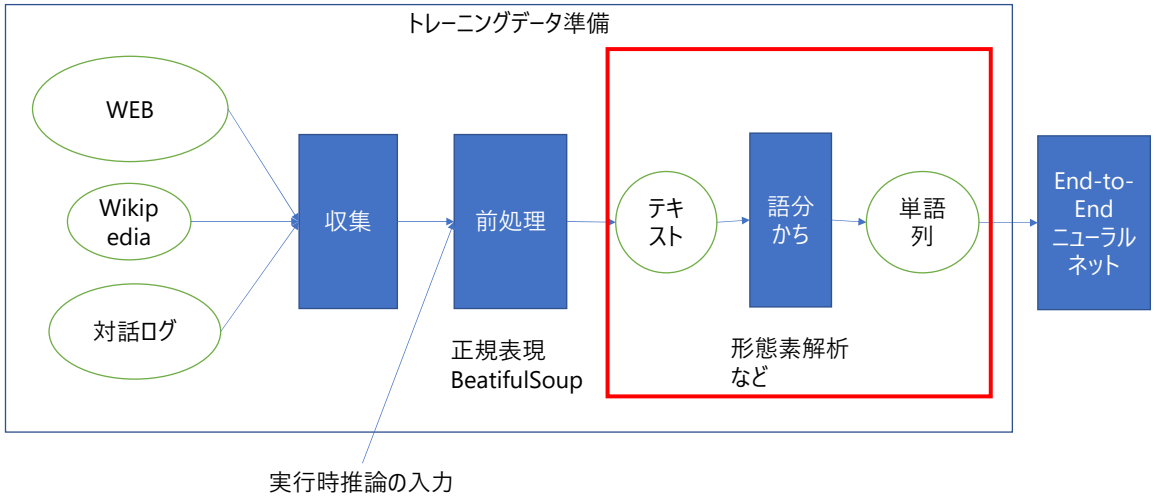
形態素解析は、必須ではなく、応用による



処理目的によっては、人名自動生成など、文字単位で処理すれば済むケースがあります。

また、単語単位で処理をする場合、分かちをするために、文字づらの統計処理、あるいは情報理論的な処理で（頻度の高い文字連鎖を単語とみなし、そうでない固有名詞などは文字単位で扱う）分かちをすることもあります。しかし、たいていは、辞書に基づいた形態素解析を行って、単語区切りを設定します。

通常のニューラルネット開発プロセス内での形態素解析の位置づけ



最後に、ニューラルネット開発プロセス内での形態素解析の位置づけを見ておきます。
最初にデータを収集します。それらはHTMLタグを含んでいたりしますので、前処理して、きれいなテキストにします。そのプレーンなテキストに対し、語分ち処理をして、単語列に変換します。この単語列が、End-to-endニューラルネットへの入力となります。

100本ノック第4章課題30～39

- 「100本ノック」の4章の課題を解いてみましょう。
- 第4章の課題は、形態素解析エンジンがすでにあってそれが利用できるという前提で、その出力加工を Python でどう書くかという課題です。
- 「NLP準備、形態素解析.ipynb」というノートをコピーし、冒頭の準備、基本事項をやった後、各課題のセクション下のコードセルに解答コードを完成させ、実行ログを残してください。
- データを扱う際、pyplotというグラフ描画パッケージを重宝します。ついでに基本的な使い方をマスターします。
- 以下に、課題を解く際に参考となることを説明します。

課題を解くための参考情報

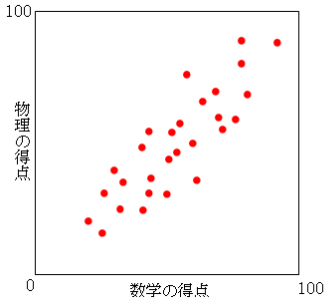
ヒストグラム

- 縦軸に度数、横軸に階級をとった統計グラフ

https://ja.wikipedia.org/wiki/%E3%83%92%E3%82%B9%E3%83%88%E3%82%B0%E3%83%A9%E3%83%A0

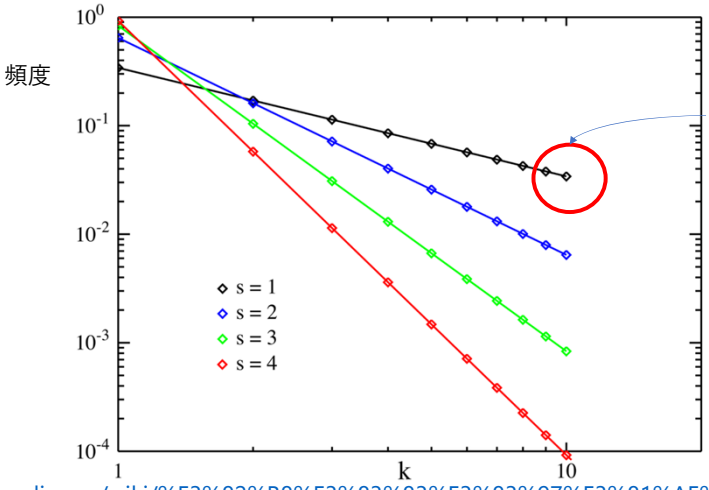
散布図

- 二つの特性を横軸と縦軸とし，観測値を打点して作るグラフ表示



https://ja.wikipedia.org/wiki/%E6%95%A3%E5%B8%83%E5%9B%B3

Zipf(ジップ)の法則：。言語や人文科学の分野で広範囲にみられる現象です。

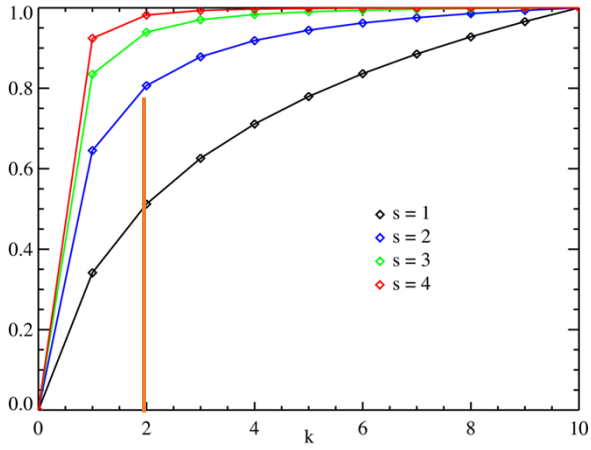


10番目の項目の頻度は、1番目の項目の頻度の1/10。

https://ja.wikipedia.org/wiki/%E3%82%B8%E3%83%83%E3%83%97%E3%81%AE%E6%B3%95%E5%89%87

Zipfの法則≒パレート(80/20)の法則：上位20%の原因が、80%の問題を引き起こす。上位20%の営業員が80%の利益を稼ぐ

累積分布



順位

https://ja.wikipedia.org/wiki/%E3%82%B8%E3%83%83%E3%83%97%E3%81%AE%E6%B3%95%E5%89%87
https://effectiviology.com/80-20-rule-pareto-principle/#Related_concept_Zipf%E2%80%99s_law

確認クイズ

- 形態素解析の確認クイズをやってください。