My Book

published by ReVIEW

hsm_ai と学ぶはじめての人工知能

hsm_hx　著

2018-11-10 版　　発行

第 1 章

hsm_ai とは ~はじめての人工知能~

1.1 この本について

hsm_ai は、プログラミング言語 Ruby により制作された人工知能[*1]です。
この本では、hsm_ai を制作するにあたって用いた以下の技術について取り上げます。

- 形態素解析
- マルコフ連鎖
- 係り受け解析
- 特徴語抽出
- 感情極性辞書による感情解析
- VRoidStudio による 3D モデル製作
- ゲームエンジン Unity

1.2 hsm_ai とは

hsm_ai は、開発者である私、hsm_hx の発言を学習し文章を自動生成するマルコフ連鎖による bot です。

私の Twitter[*2]からツイートを取得し、形態素解析という手法を用いてそのツイートを単語ごとに分割します。単語ごとに分割した文章を数単語ずつの塊にし、そうしてできたたくさんの単語の塊を規則に従って組み替えることで日本語らしい文章を生成します (詳しくは第 1 章にてお話します)。

百聞は一見にしかずとも言うので、まずは hsm_ai の生成した文章をいくつかご紹介します。以下に示す文章は、全て hsm_ai というシステムによって生成されたものです。

> ひえ～8 時間で自動的に目を覚ました
>
> フォロワー 37 人もいるの人間になる
>
> ミーン！（閃いた顔）

[*1] 正確には人工知能と呼べるものではありませんが、それについては第 n 章にて詳しく記述します。
[*2] @hsm_hx: https://twitter.com/hsm_hx

ミーン！ で爆笑してるでしょ

頭を使うことを学んだ

あぁ＾〜単位落ちる

声出して笑ってもらえるって嬉しいかもしれない

iTunes で素敵に心を購入

お前はやりたいことやるのが大事

いかがでしょうか？ 確かになんとなく不自然に見える文章もありますが、かなりの割合で日本語として解読が可能な文章が生成されていると思います。それどころか、人間には書けないような独創的（という表現が正しいのかはわかりませんが）な文も散見されます。

この本では、この hsm_ai が生み出され対話型 bot として高専祭で展示されるまでの成長の過程をひとつずつ追っていきます。

1.3 開発環境

hsm_ai を開発するにあたって利用した言語やライブラリ、ツールは以下の通りです。

Antergos

Windows, MacOSX と並ぶ OS である Linux の一種。Arch Linux というディストリビューションを使いやすくするため予め必要なパッケージを揃えたもの。Arch Linux は最高です。

Ruby 2.5.1

日本人によって開発されたスクリプト言語。web サービスを作るのによく使われています。

MeCab 0.996

日本語の文章を形態素解析 (第 2 章にて解説) するためのエンジン。

CaboCha 0.69

日本語の文章を係り受け解析 (第 4 章にて解説) するためのエンジン。

Unity

言わずとしれた超有名ゲームエンジン。3D モデルを動かすために使用します。

VRoidStudio

最近流行りの 3D モデル作成ソフト。絵を描くように直感的な操作で本格的な 3D モデルが作れます。

第 2 章

マルコフ連鎖による文章生成

2.1　マルコフ連鎖とは

マルコフ連鎖とは、物理や統計、強化学習など様々な分野において事象をモデル化するためにしばしば用いられる確率論の考え方の一種です。

この本ではマルコフ連鎖について詳しく踏み込むことはしませんが、ざっくりとその性質を説明すると、ある事象についてその未来を考えるとき、「その事象の未来は過去に関わりなく、現在の状態のみによって定まる」という特性を持つときの未来予測に用いられるアルゴリズムです。

この後で具体的に例を挙げて解説しますが、hsm_ai ではこのような理論を応用することで日本語らしい文章を機械生成しています。

2.2　マルコフ連鎖による文章生成アルゴリズム

では、実際にマルコフ連鎖を用いて文章を生成していきます。hsm_ai の文章生成には、以下のようなアルゴリズムを採用しています。実際に例を挙げながら、機械的に文章が生成される過程を追っていきましょう。

まず、学習元になる複数の文章を用意します。ここでは、例として以下の 2 つの文を用意しました。

ここにりんごがあります。その箱にはぶどうが 2 つ入っています。

この文章をマルコフ連鎖するために、まずは下準備として、それぞれの文を単語ごとに分割します。これを分かち書きといいます。上の文を分かち書きすると、下のようになります。

ここ / に / りんご / が / あり / ます / 。その / 箱 / に / は / ぶどう / が / 2 / つ / 入っ / て / い / ます / 。

さて、次に、この分かち書きされた文章から、3 つの連続する単語をひとまとめにしたブロックを作ります。ここでポイントになるのは、各文章のはじめとおわりにそれぞれ「ここが文のはじまり (おわり) です」という目印をつけることです。実際にブロックを作ってみます。ここでは、はじめとおわりを表す印として「*」という記号を使うことにします。

[*, ここ, に], [ここ, に, りんご], [に, りんご, が], [りんご, が, あり], [が, あり, ます], [あり, ます, 。], [ます, 。, *] [*, その, 箱], [その, 箱, に], [箱, に, は], [に, は, ぶどう], [は, ぶどう, が], [ぶどう, が, 2], [が, 2, つ], [2, つ, 入っ], [つ, 入っ, て], [入っ, て, い], [て, い, ます], [い, ます, 。], [ます, 。,

*]

[] の中に、3 つの連続する単語をコンマ区切りで並べています。hsm_ai の文章生成には、プログラムにより生成された何千、何万もの単語ブロックが使われています。

さて、このたくさんのブロックをどのように使うのかというと、ここでマルコフ連鎖というものを使います。条件に合うブロックを探し、その中からランダムに 1 つを選び後ろにつなげていくことで文章を生成していきます。具体的に上のブロックを使ってマルコフ連鎖によって文章が生成される過程を追ってみましょう。

まず、文章のはじめは「*」としてありました。なので、「*」からはじまるブロックを探します。[*, ここ, に], [*, その, 箱] の 2 種類です。

この 3 つの中から、ランダムに 1 つを選びます。今回は [*, ここ, に] を選んだとします。

次に、[*, ここ, に] につながるブロックを探します。つまり、「に」からはじまるブロックを選べば良いというわけです。ここでは、[に, りんご, が], [に, は, ぶどう] の 2 つが考えられます。

この 2 つの中から、ランダムに 1 つを選びます。次は [に, は, ぶどう] が選ばれたとします。

こうして、[*, ここ, に], [に, は, ぶどう] という 2 つの接続可能なブロックが選ばれました。これを「*」で終わるブロックに到達するまで続けます。その様子を表したものが下の図です。

このように、「ここにりんごがあります。」,「その箱にはぶどうが 2 つ入っています。」という 2 つの文から新しく、「ここにはぶどうが 2 つ入っています。」という意味の違う文章が生成されました。これが、マルコフ連鎖による文章生成です。

2.3 形態素解析エンジン MeCab

さて、マルコフ連鎖による文章生成アルゴリズムについてはなんとなくイメージを掴んでもらえたかと思います。しかし、先ほど登場した「分かち書き」という処理をプログラミングで実現するにはどうすればいいでしょうか？ 3 つの単語をブロックにして連鎖させる…といった箇所については、ある程度プログラミングの経験がある人であれば愚直にコードに起こすことができるでしょう。ですが、ある文章を単語ごとに分割してその品詞を特定する、という処理はどうやって書けばいいのでしょうか？ おそらく、それを実現するには膨大な時間と研究が必要です。

そこで役に立つのが形態素解析エンジン MeCab[*1]です。

MeCab は、京都大学と NTT 株式会社の共同研究プロジェクトによって開発された形態素解析エンジンです。形態素解析というのは、日本語や英語など、私達が普段から使う言語（自然言語）の文を単語に分割し、その品詞などを判別する解析作業のことを指します。

MeCab を使用することで、文章を簡単に形態素解析し分かち書きされた状態にすることができます。MeCab は GitHub 上で公開されているオープンソースソフトウェアなので、Git が導入されている環境であればリポジトリ[*2]を clone しビルドすることですぐ使えるようになります。

```
$ git clone https://github.com/taku910/mecab.git
$ cd mecab/mecab
```

[*1] http://taku910.github.io/mecab/
[*2] https://github.com/taku910/mecab

```
$ ./configure --with-charset=utf8
# make install
$ cd ../mecab-ipadic
$ ./configure --with-charset=utf8
# make install
```

以上の手順で、MeCab 本体と MeCab を動かすための辞書データをインストールします。また、必要に応じて、新語やネット用語などに特化した辞書データである mecab-ipadic-neologd[*3]も追加で導入します。

ここまでできたら、MeCab を実際に動かしてみます。コマンド上で ‘mecab‘コマンドを実行し、続けて好きな文章を入力することで動作を確認することができます。

また、MeCab は各種プログラミング言語からスムーズに使用するためのバインディングを標準で提供しています（Perl, Ruby, Python, Java, C#）。その他、MeCab をより快適に利用するためのライブラリも各種言語で充実しています。例えば、Ruby では natto という Gem が配布されており、hsm_ai は natto を採用しています。

2.4 Twitter の bot としてリリースする

さて、理論を抑えたところで、実際に Twitter からツイートの情報を取得し、そのデータを元に生成した文を Twitter に投稿してみます。

Twitter からデータを取得したり、ツイートやいいねなどをプログラムから行うには、Twitter が公式に提供している API を使います。API とは Application Program Interface の略で、プログラムから何らかのアプリケーションを利用するための決まった形式のことを指します。具体的には、今回使う TwitterAPI の他にも Google の提供する Google Maps API や Microsoft の提供する Face API など、様々なものがあります。このような API を利用することで、プログラミングを始めたての初心者でも既存のサービスの機能をプログラムから利用したり、顔認識や機械学習などの複雑な処理を自分で実装することなく自分のプログラムに組み込むことができます。

今回はツイートの取得、自動ツイートを実現したいので、TwitterAPI を使います。TwitterAPI の利用には Twitter 開発者登録が必要です。Twitter のデベロッパー向けページにアクセスし、開発者として認証してもらうために数点の質問に答えます。開発者として認証されると、プログラムから TwitterAPI を利用するために必要なトークンが発行されるので、それを使って TwitterAPI を利用するプログラムを書きます。

TwitterAPI の利用の仕方については、インターネットで検索するとたくさんヒットするのでここでは割愛しますが、多くのプログラミング言語では TwitterAPI を簡単に利用するためのライブラリが公開されています。

このようにして作られたのが、現在 Twitter 上で動作している hsm_ai です。ソースコードは GitHub[*4]上で公開しているので、興味がある人は参考にしてみると良いかもしれません。

[*3] https://github.com/neologd/mecab-ipadic-neologd
[*4] https://github.com/hsm-hx/hsm_ai

第 3 章

hsm_ai と学ぶはじめての人工知能

2018 年 11 月 10 日　初版第 1 刷　発行

著　者　hsm_hx