1 イントロダクション

1.1 TL;DR

- ▶ 石川啄木の短歌に含まれる単語を意味ごとに分類し特徴量化
- ▶ 特徴化された短歌をナイーブベイズで学習し分類器に
- ▶ 完成した分類器をマルコフ連鎖で生成した短歌に適用し、最も「石川啄木らしい」短歌を選択する

1.2 対象読者

- ▶ とりあえず短歌や石川啄木に興味がある人
- ▶ 機械学習ってよく聞くけど何ができるのかよくわからない人
- ▶ 「偶然短歌bot」よりももうちょっとちゃんとした短歌生成をしたい人
- ▶ 機械学習やってるけど最近の深層学習一辺倒の流れに飽き飽きしてきた人（マサカリ投げないで＞＜）

1.3 能書き

おはようございます。タイトルに機械学習って入れればだいたいバズるだろみたいな甘い考えで書き始めた博多市 (hakatashi) です。いちおうこの SunPro の会長ということになっています。この会誌の編集長もしています。自分で設定した締切当日にこれを書いてます。つらい。

さて、今回は技術島としてのコミケ初参加ですが、せっかくコミケに参加するんだから、いつものガチガチの技術的なことではなく、少し趣を変えて文芸的なことをやりたいと思いました。そこで、以前から暖めていたアイデアである「石川啄木の未完の短歌」と、最近よく見る「機械学習」についての研究を記そうと思います。

機械学習といっても、結論から言うと、流行りの深層学習とかそういういうのではなく、また人工知能やらにも応用しにくい内容ですが、短歌に興味がある方も機械学習に興味がある方も、どうか最後までお付き合いください。

この記事では、基本的にやったことを中心に述べ、プログラムのソースコードは掲載しません。使用したスクリプトなどは https://github.com/hakatashi/takuboku-last-tanka に掲載するので、あわせて参照してください。

2 石川啄木について

まずは機械学習云々の前に、ご存じの方も多いでしょうが、今回の研究対象である石川啄木について簡単に触れておきます。

2.1 略歴

図2.1 石川啄木の肖像

石川啄木（本名 石川一）は、明治期に活躍した日本の歌人です。この頃知られている歌人の中でも、特に若くしてこの世を去った歌人であり、明治 45 年、満 26 歳にて病死しました。彼の晩年の友人である土岐善麿が、昭和 55 年、満 94 歳にして生を全うしたことを考えると、そのあまりにも早すぎる死が際立ちます。

石川啄木は、岩手県に生まれ、宮沢賢治や金田一京助などの数々の文人を輩出した旧制盛岡中学校に在籍しました。現在知られている彼の作品は、主に新聞や同人誌などの歌壇に掲載された物が多く、そのうち半分以上が歌集に収録されています。啄木が生涯で著した歌集は、わずかに『一握の砂』と『悲しき玩具』の二集のみであり、『一握の砂』は明治 43 年、『悲しき玩具』は明治 45 年、先述した土岐善麿の手によって啄木の死の二ヶ月後に、それぞれ出版されました。

2.2 啄木の作風

彼の歌風は極めてストレートで、特に歌集『一握の砂』以降の歌作は、自然主義的傾向から日常に去来する感情を素直な言葉で綴ったものが多く、卑近ながらにして読む者をハッとさせるような驚きと共感を与えてくれます。啄木といえば、現代的には彼の書いた赤裸々でスキャンダラスな「ローマ字日記」が（いい意味でも悪い意味でも）通俗に知られていますが、これも自らの感情を包み隠さず歌った啄木ならではの文章だと言えるでしょう。

僕が特に気に入っている作品をいくつか紹介しましょう。

> 治(おさ)まれる世の事無(ことな)さに
> 飽(あ)きたりといひし頃こそ
> かなしかりけれ
>
> 「さばかりの事に死ぬるや」
> 「さばかりの事に生くるや」
> 止(よ)せ止せ問答
>
> 今までのことを
> みな嘘にしてみれど、
> 心すこしも慰(なぐさ)まざりき。
>
> 古新聞！
> おやここにおれの歌の事を賞(ほ)めて書いてあり、
> 二三行(ぎやう)なれど。

2.3 形式的特徴

啄木の歌には、いくつかの形式的な特徴があります。これは機械学習によって「啄木の短歌」を生成する際に大いに考慮するべき特徴なので、詳しく述べていきます。

まず何より重要なのは、それまでの短歌や和歌には無かった「一首三行書き」のスタイルです。啄木は新聞や同人誌などの文壇に投稿する際は、スタンダードな一行書きの形式で掲載されることが多かったのですが、歌集に収録される短歌はすべて、この「三行書き」のスタイルで記されています。これは土岐善麿の第一歌集『NAKIWARAI』の「ローマ字一首三行書き」の影響を受けてのものだと言われており、改行によって一息つくことによって伝統的な寂(さび)にも似た美観が生まれるとともに、見た目にもコンパクトにまとまった印象を受けます。

次に、三十一文字の定型に囚われない、大胆な字余りと句跨がりが特徴として挙げられます。明治以降においてはこのような破調を得意とする歌人が多いのも事実ですが、その中でも啄木は特に過激な部類に属し、一句に二音以上、歌全体で六音以上の字余りというのもざらに見られます。

最後に、句読点などの約物の多用と字下げによる表現が挙げられます。この特徴は主に第二歌集『悲しき玩具』に見られ、句読点の他に感嘆符、疑問符、ダッシュ、括弧類が散りばめられて、短歌であるにもかかわらず、まるで散文のような読み応えを感じます。字下げは『悲しき玩具』の後半百首程度から特に顕著になり、三行書きと相まって読む者に極めて特殊な印象を与えます。

全体として、内容的にも形式的にも、これまでの短歌のスタイルから大きく外れながら、優れた名歌を詠み続けた歌人として、彼の死から百年以上経った現代においても石川啄木は知られています。僕も詩歌や短歌の類をこよなく愛していますが、その中でも啄木は歌人とくれば第一に挙げられるほど愛好しており、昔から彼の全集や歌集を書棚に収集しています。僕が好んでいるのは、何よりも彼の歌に見られる彼の繊細な美意識と、それを時として過激な言葉で綴ったストレートな感情の吐露、そして三行書きという新たな枠組みによって確立されたリズムの心地よさです。完全に余談ですが、僕も文芸を愛するものとして、ときおり歌屑など詠んでみるものですが、その際は彼に倣って「三行書き」のスタイルを順守するようにしています。

2.4 啄木、最後の一首

さて、本篇の題にある「未完の短歌」という言葉に訝しんだ方も多いでしょう。「未完の小説」「未完の大作」という言葉はあれど、短歌において未完ということはほとんど無いはずです。僅かに三十一文字の短歌を途中で書いたまま残しておくということが考えにくいため当然ですが、こと啄木においてはこの「未完の短歌」というものが存在します。

26 という歳で惜しまれつつも結核で夭逝した石川啄木ですが、先述したとおり、彼の最後の歌集である『悲しき玩具』は彼の死の直後に出版されています。これは啄木が死の直前に、「一握の砂以降」と題された直筆の大学ノートなどを歌稿として託していたもので、このノートが彼の遺稿であると言えるでしょう。

手元にこのノートの複製原稿があるため参照しますが、このノートには規則正しく一ページに三行書きの短歌が四首記載され、余計な書き込みもない綺麗な原稿です。そして、最後のページには次のように記されています。

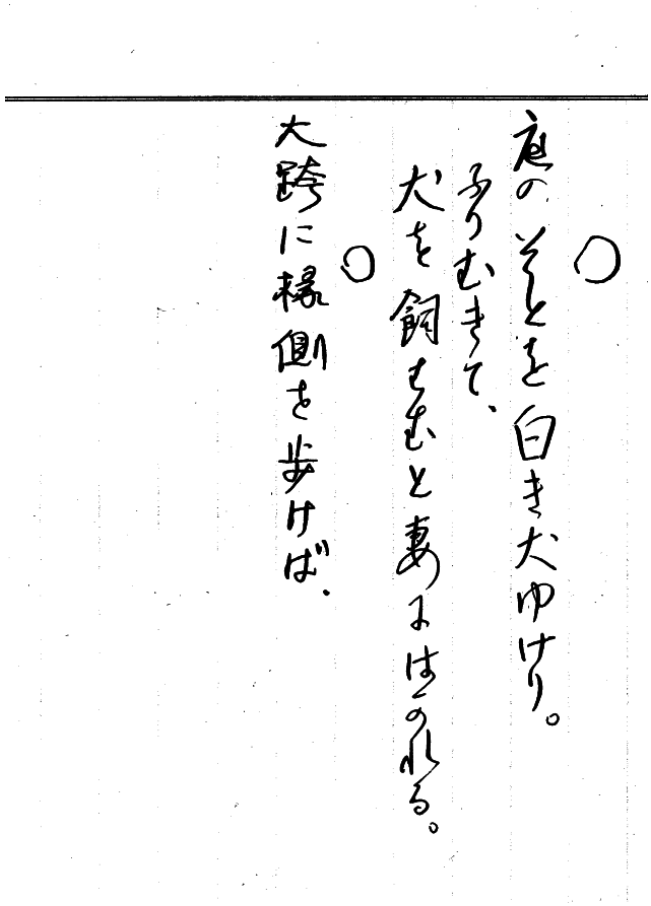


図2.2　『悲しき玩具』最後のページ

ご覧のとおり、一ページ四首であるところが二首しか記されておらず、最後の一首は初二句が一行に書かれているのみで、続きが記されていません。通し番号で言えば 193 首目、何らかの理由で啄木はここで筆を断ち、最後の一句は未完成のまま遺されたと言っていいでしょう。

この未完成の一首は、善麿の発行した『悲しき玩具』や、のちに出版された啄木の全集や歌集では単に無視され、啄木研究家の間でも深く顧みられることはありませんでした。そのため一般にはあまり知られていない事実ですが、啄木にはこのように絶筆となった「未完の短歌」が存在します。

『悲しき玩具』の筆跡

ちなみに、この未完の短歌を含むノートの最後の数ページですが、明らかにそれ以前の部分と筆跡が異なります。それまでのページは、全体的にやや扁平の行書体で書かれており、仮名も現代のものとほぼ変わらず、今から読んでも非常に読みやすいものになっているのに対して、こちらの数ページのほうは連綿が多く、変体仮名が用いられた女性的な筆跡であり、多少の古文書学の知識がないと読みにくいものになっています。

この数ページに含まれる 17 首の短歌が、誰か別人が代筆したものか、それとも啄木の自筆によるものか、もしそうだとしたらなぜ筆跡を変えたのか、という問題は啄木研究者の間でも意見が分かれており、特に定まった説明がありません。いずれにせよこれらの短歌が啄木の作であるということに異論はないので、今回はこれ以上深く触れないことにします。

これについては、特に最近の論文で、同筆説を掲げる湯澤比呂子の論文 [3] に詳しく記されているので、参照してください。

3 「最後の一首」を復元する

さて、このように謎の多い未完成の一首ですが、気になりはしないでしょうか。欠損しているのは五・七・七のわずかに 19 音です。啄木がここに何を記したかったのか、何を詠みたかったのか。わずか 19 音なら復元できる気がしませんか？

今回は機械学習などの技術を駆使してこの 19 音を復元する試みを記します。復元には様々な手法が考えられますが、今回は以下の手順で復元を試みました。

- ▶ 啄木の既存の短歌を形態素解析器にかける
- ▶ マルコフ連鎖で「大跨に縁側を歩けば」で始まる適当な文章を生成する
- ▶ 生成した文章のうち、文節に区切った時に五・七・七の条件を満たしている物を抽出し、これらを「最後の一首」の候補歌とする
- ▶ これとは別に啄木の短歌から一首ずつ特徴量を抽出してナイーブベイズによる分類器を作る
- ▶ 「最後の一首」の候補歌を分類器にかけ、最もスコアが高いものを復元された「最後の一首」とする

それぞれのステップを解説していきます。

4 字句解析

まずは、啄木の短歌にどのような単語がどのように使用されているかを調査します。

4.1 原本について

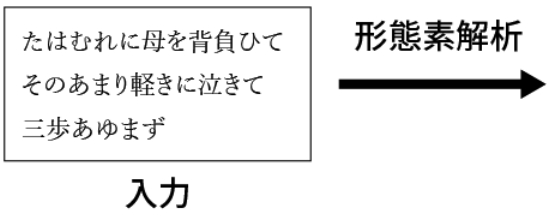
今回は、「最後の一首」推定に使用する「啄木の短歌」として、『一握の砂』所収の 551 首と、

『悲しき玩具』所収の 194 首、合わせて 745 首を使用しました。啄木にはこのほかに歌集未収録の歌が千首以上知られていますが、啄木の短歌に特徴的な三行書きの形式が見られないこと、および信頼できるデジタルデータが見つからないことから入力として使用しませんでした。745 個というのは機械学習の学習データとしてはかなり少ないほうですが、まあ、こればかりは仕方がないでしょう。

入力データは、青空文庫のデータを使用しました。([1], [2])

4.2 形態素解析

形態素解析とは、特殊な処理を施していない普通の文章 (自然言語) を、形態素という単位に分割する解析処理のことです。



形態素	読み	品詞	原形
たはむれ	タワムレ	名詞	戯れ
に	ニ	助詞	に
母	ハハ	名詞	母
を	オ	助詞	を
背負ひ	セオイ	動詞	背負う
て	テ	助詞	て
その	ソノ	連体詞	其の
あまり	アマリ	副詞	余り
軽き	カルキ	形容詞	軽い
に	ニ	助詞	に
泣き	ナキ	動詞	泣く
て	テ	助詞	て
三	サン	名詞	三
歩	ポ	名詞	歩
あゆま	アユマ	動詞	歩む
ず	ズ	助動詞	ず

図4.1　形態素解析の実例

形態素は言語学の用語で、一般に言われる「単語」と概ね一致しますが、接尾辞や接頭辞など、単独で用いられない語素も含まれます。自然言語の文章を解析してこの形態素に正しく分解し、品詞推定などを行う形態素解析は、日本語の自然言語処理を行う上での最初の一歩であり、いわば料理で言うところの下ごしらえみたいなものです。このあたり英語やドイツ語などの西洋諸語ではだいぶ楽なんですが、いかんせん日本語は分かち書きを行わないのでまず単語の境界部分から推定しなければいけません。

この形態素解析を行う形態素解析器は、ChaSen や MeCab や kuromoji など、すでに多くの実装が存在し、それぞれに長所や短所がありますが、今回はあとに紹介する近代文語 UniDic を使用するために MeCab を選択しました。

4.3 辞書データ

形態素解析を行うには、解析対象の文章で用いられている単語を収録した辞書データが必要です。一般的には IPA 辞書などの現代語コーパスを用いた辞書データを使用するのが普通なのですが、今回解析する文章は明治末期の短歌であり、当然歴史的仮名遣いで記述されています。普通の辞書では太刀打ちできません。

そこで、今回使用したのは、国立国語研究所が公開している近代文語 UniDic[1]です。これは、同研究所が開発した現代語版の形態素辞書である UniDic に近代日本語（特に明治期）の語彙を追加したもので、現在知られているほとんど唯一の近代語彙の形態素辞書と言っていいでしょう。

これを MeCab にインストールして、先ほどの文章を食わせるとこうなります。

リスト4.1　近代文語UniDic解析例

```
$ echo "たはむれに母を背負ひて そのあまり軽きに泣きて 三歩あゆまず" | mecab
たはむれ        名詞,普通名詞,一般,*,*,*,タワムレ,戯れ,たはむれ,タワムレ,タハムレ,和,たはむれ,タワムレ,タハムレ,タワムレ,*,*,*,*,*,*,0,C2,*
に      助詞,格助詞,*,*,*,*,ニ,に,に,ニ,ニ,和,に,ニ,ニ,ニ,*,*,*,*,*,*,*,名詞%F1,*
母      名詞,普通名詞,一般,*,*,*,ハハ,母,母,ハハ,ハハ,和,母,ハハ,ハハ,ハハ,ハハ,*,*,*,*,*,*,1,C3,*
を      助詞,格助詞,*,*,*,*,ヲ,を,を,オ,ヲ,和,を,オ,ヲ,ヲ,*,*,*,*,*,*,*,"動詞%F2@0,名詞%F1,形容詞%F2@-1",*
背負ひ  動詞,一般,*,*,文語四段-ハ行,連用形-一般,セオウ,背負う,背負ひ,セオイ,セオヒ,和,背負ふ,セオウ,セオフ,セオウ,*,*,*,*,*,*,*,2,C1,*
て       助詞,接続助詞,*,*,*,*,テ,て,て,テ,テ,和,て,テ,テ,テ,*,*,*,*,*,*,*,"動詞%F1,形容詞%F2@-1",*
その    連体詞,*,*,*,*,*,ソノ,其の,その,ソノ,ソノ,和,その,ソノ,ソノ,ソノ,*,*,*,*,*,*,0,*,*
あまり  副詞,*,*,*,*,*,アマリ,余り,あまり,アマリ,アマリ,和,あまり,アマリ,アマリ,アマリ,*,*,*,*,*,*,*,0,*,*
軽き    形容詞,一般,*,*,文語形容詞-ク,連体形-一般,カルイ,軽い,軽き,カルキ,カルキ,和,軽し,カルシ,カルシ,カルシ,*,*,*,*,*,*,*,1,C1,*
に      助詞,格助詞,*,*,*,*,ニ,に,に,ニ,ニ,和,に,ニ,ニ,ニ,*,*,*,*,*,*,*,名詞%F1,*
泣き    動詞,一般,*,*,文語四段-カ行,連用形-一般,ナク,泣く,泣き,ナキ,ナキ,和,泣く,ナク,ナク,ナク,*,*,*,*,*,*,*,0,C2,*
て       助詞,接続助詞,*,*,*,*,テ,て,て,テ,テ,和,て,テ,テ,テ,*,*,*,*,*,*,*,"動詞%F1,形容詞%F2@-1",*
三      名詞,数詞,*,*,*,*,サン,三,三,サン,サン,漢,三,サン,サン,サン,サン,*,*,N3,*,*,*,*,0,C3,*
歩      名詞,普通名詞,助数詞可能,*,*,*,ホ,歩,歩,ポ,ポ,漢,歩,ポ,ポ,ホ,ホ半濁,半濁音形,*,*,*,B1S6SjShS,1,C3,*
あゆま  動詞,一般,*,*,文語四段-マ行,未然形-一般,アユム,歩む,あゆま,アユマ,アユマ,和,あゆむ,アユム,アユム,アユム,*,*,*,*,*,*,*,2,C1,*
ず      助動詞,*,*,*,文語助動詞-ズ,終止形-一般,ズ,ず,ず,ズ,ズ,和,ず,ズ,ズ,ズ,*,*,*,*,*,*,*,*,"形容詞%F4@-1,動詞%F3@0",*
EOS
```

この近代文語 UniDic は非常に優秀で、例えばこの例における「軽き」の部分ですが、通常の文語における原形である「軽し」のほかにも、対応する現代語の原形である「軽い」が情報として与えられます。これによって、現代語でないにもかかわらず、意味としてはほとんど現代語と統一的に処理することができます。このほかにも、「和語」「漢語」といった語種の情報など、言語解析においてとても有用なデータが揃っているため、応用がしやすい辞書です。[2]

まず、この形態素解析器を用いて、745 首の短歌を形態素に分解しました。精度を求めるなら、この程度の量は人間によるチェックを入れたほうが良いのですが、プログラマである僕にそんな体力はなかったのでチェックはしていません。そういうのはもっと暇な人にお願いしましょう。

1　配布 URL: http://www2.ninjal.ac.jp/lrc/index.php?UniDic%2F%B6%E1%C2%E5%CA%B8%B8%ECUniDic

2　一つ不満があるとすれば、現代語版の UniDic がオープンソースで BSD などのパーミッシブなライセンスで配布されているのに対し、近代文語 UniDic がバイナリでの配布で、二次利用が制限されているという点でしょうか……

5 マルコフ連鎖

解析した短歌をもとに、マルコフ連鎖というテクニックを用いて「啄木の短歌」っぽい文章を無限に生成していきます。

5.1 マルコフ連鎖について

プログラミングや自然言語処理について詳しく知らない方でも、マルコフ連鎖という単語は耳にしたことがあるのではないでしょうか。近年は Twitter 上のツイートをもとにマルコフ連鎖で適当なことを言わせる bot が多く知られており、しゅうまい君[3]などは特に有名だと思います。

マルコフ連鎖は、本来は状態遷移する確率モデルの解析に用いられる用語で、特に言語処理特有の用語ではなく、強化学習などの分野でも用いられます。これは解析対象のモデルがマルコフ性[4]を持つという仮定のもとで、存在しうる状態遷移を列挙するパターンの一つです。自然言語処理におけるマルコフ連鎖は、これを形態素間の遷移関係に応用したもので、与えられたコーパスをもとに新規に文章を生成する能力を持ちます。が、当然ながら本来の自然言語がそんな単純なパターンで生成されるはずもなく、マルコフ連鎖が生成するのはあくまで擬似的な自然言語にとどまります。

5.2 マルコフ連鎖の仕組み

では、具体的にマルコフ連鎖がどのように文章を生成するのか見てみます。今回の解析対象の一つである『一握の砂』所収の一首、「うるみたる目と／目の下の黒子のみ／いつも目につく友の妻かな」を例にとります。

3 https://twitter.com/shuumai

4 すべての過程の将来条件の確率分布が現在状態にのみ依存し、そこに至る状態によって左右されないこと

入力　うるみ/たる/目/と/目/の/下/の/黒子/のみ/いつ/も/目/に/つく/友/の/妻/かな

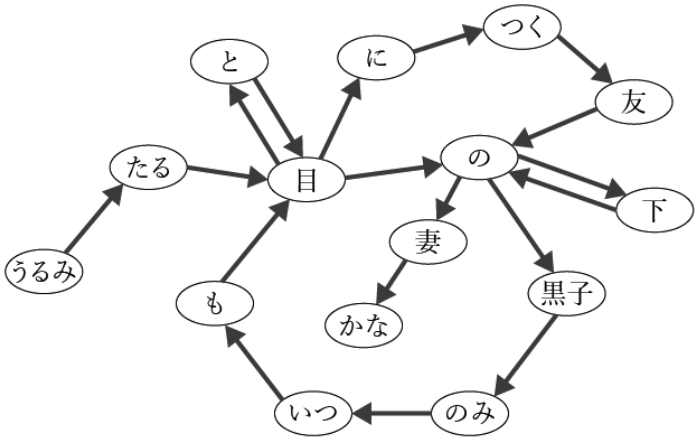
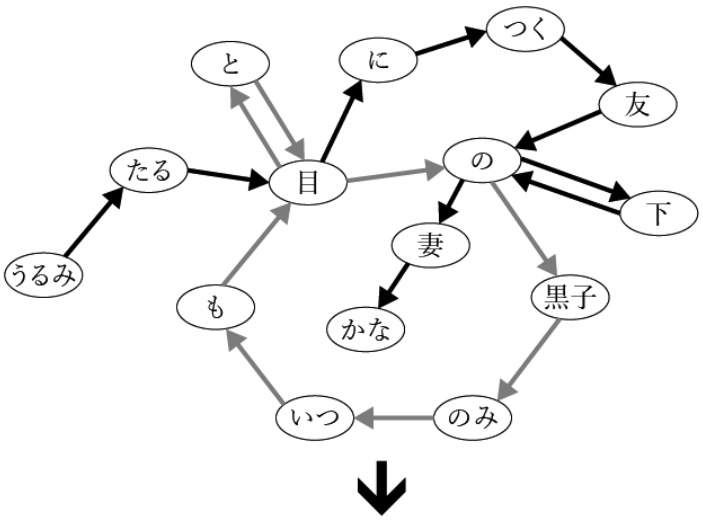


図5.1　マルコフ過程によるオートマトン生成

まず、マルコフ過程は、文章における形態素を状態とみなし、形態その並びを状態間の遷移関係を表現しているものとみなします。そして、この遷移関係をもとに、オートマトンと呼ばれる状態遷移図を作成します。

図では出現する形態素を丸印で、形態素どうしの遷移を矢印で表しています。左端の「うるみ」から矢印をうまく辿って行くと、もとの短歌を復元できることが分かるかと思います。



出力　うるみ/たる/目/に/つく/友/の/下/の/妻/かな

図5.2　マルコフ連鎖による文章生成

次に、このオートマトン上を、もとの短歌における順番を無視して適当な方向に辿っていきます。

上図では、実際に辿った遷移を濃く塗ってあります。辿った形態素を前から順番に繋げて読んでみると、「うるみたる目につく友の下の妻かな」となります。まあ、これだけではあまり実際的な意味を持つ文章になりませんし、短歌の形式にもなっていないのですが、日本語として文法の正しい文章にはなっています。

このようにマルコフ連鎖は、与えられたコーパスをもとに、それに似た文章をいくらでも生成することができます。逆に言うと、与えられたコーパスに含まれる単語からしか文章を生成できないという欠点を持ち、学習データが小さいと生成できるパターンも少なく、無理のある文章になりがちです。

今回の学習データが啄木の生涯の歌作の半分近くを網羅しているとはいえ、さすがに 745 首では啄木の語彙の全てをカバーしているとは言いがたいです。実際、今回復元しようとしている歌の初二句に含まれる「大跨」「縁側」と言った単語は、これらの学習データには含まれていません。もっと復元の精度を上げるためには、何らかの方法で辞書データから啄木の使いそうな単語を新規に選択することが求められますが、そのあたりはなかなか難しいところです。

実際に 745 首のデータをもとに「大跨に縁側を歩けば」からマルコフ連鎖を伸ばしていくと、下のようになります。(なお句読点や約物のたぐいは、短歌の意味を取る上で本質的でないという理由から除去してあります)

リスト5.1　近代文語UniDic解析例

```
$ lsc marcov.ls
大跨に縁側を歩けば襟に頤をぢっとしわが泣くをこはす気持のよさは逃げ去れり病ある眼このごろ思ふ
大いなる敵目の酢のかと医者にかの浜薔薇よ墓ぞ子よかなしくもめとらず泣きたくて酒のむが出づるや
はこの世さびしくも長く書きさしてふと思ふ存分叱りつくる人が小さくなればかなそれをも暮したりと
をのべて笑める日も来にみしかな昔小半日堅き皮をばむしりてあそびきそれにそを嗅ぐ
```

これだけではまだまだ短歌とはいえません。力技でどうにかしましょう。

6　短歌を抽出する

前章では啄木の短歌から「文章」を生成しました。次はここから「短歌」を生成しましょう。

6.1　短歌形式

今回の短歌復元の試みでは、短歌の形式をターゲットにしてそれを目標に文章を生成するのではなく、ランダムに生成した文章から、偶然短歌の形式になっているものを抽出するという形式を取っています。(マシンパワーが許すならそっちのほうが手っ取り早い)

御存知の通り、和歌短歌というのは五七五七七の三十一文字によって構成されます。つまり適当な場所で区切った場合にこの形式（プラス字余り）に収まればいいのですが、この区切りはどこで区切ってもいいわけではありません。

例えば、「アーモンドよりココナッツが食べたい」といった文章を綺麗な定型の俳句だと認識する人はいないでしょう。これは「アーモンド／より」や「ココナッツ／が」のような単語と単語の境界でも、文節を構成している語群に切れ目を入れることに無理があるためです。

実際、どこに切れ目を入れるべき（入れていい）か、というのは難しい問題です。古典的な短歌では、通常の句と句の間の切れ目のほかに、句切れと呼ばれる大きな内容の切れ目を入れるのが綺麗だとされています。このあたりは歌風も大きく影響してくることですが、こと啄木においては、第二章で述べたように句跨がりの多い歌人であることを踏まえ、単純に文節の切れ目が五七五七七を形成している場合に「短歌」を成しているとみなすことにしました。

6.2 文節分け

形態素解析した文章を文節に分け、さらなる解析を行うためのソフトウェアとして、CaboChaなどの係り受け解析エンジンなどが知られていますが、今回扱うのは通常の現代文と違う近代文語文なので使いづらいですし、今回は単に文節に分けたいだけなので、より単純なアルゴリズムを採用することにしました。

一般に、一つの文節には一つの自立語が対応します。自立語とは付属語とともに単語の二大分類のひとつであり、名詞、動詞、形容詞、接続詞などが含まれます。つまり、文中の自立語を見つければ、自然に文節の切れ目も見えてきます。

基本的に、見つけた自立語に後続する付属語を追い込むことで文節を形成することができます。しかし、いくつかの例外があります。

- 接頭辞は後ろの自立語に追い込む
- 動詞が連続する場合は連立ごとみなし、ひとつの文節にする
- 文初の付属語は単独で文節を形成する

これらのルールは、いわゆる単語と形態素の違いに起因するものです。このルールに従って文節分けを行うと、図 6.1のように、入力された短歌から五七五七七の句分けの構造を大観することができます。

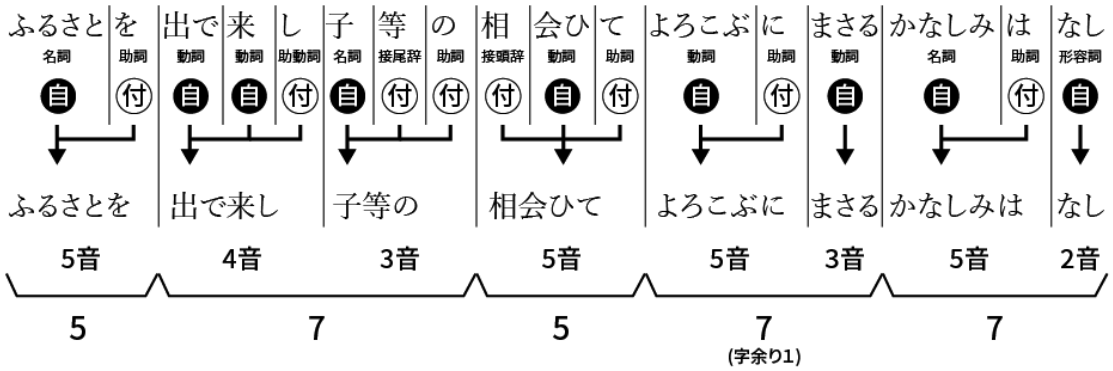


図6.1　文節分けの一例

6.3　短歌フィルターの実装

今回は、啄木の短歌としてふさわしい形式になっているものを抽出するために、これらの文節の構造に以下の制約を課しました。

- 原則として五音・七音・七音の部分で区切れなければならない
- 字余りは各句について一音まで、全体で二音まで許可する

字余りは初二句ですでに二音の余りがあるため、少なめに設定しました。啄木の特徴を考えるともう少し増やしてもいいかもしれませんが、少し増やすと字余りの短歌ばかりが大量に生成されてしまうので、計算リソースの観点から今回はこの程度にとどめました。

これにより、以下のような短歌が生成されます。

リスト6.1　マルコフ連鎖から抽出された短歌の一例

```
大跨に 縁側を歩けば いつもいふが 才に過ぎたる 煉瓦遠くより
大跨に 縁側を歩けば とある日に 酒肆の かなしさばかりの
大跨に 縁側を歩けば 湯気が夜の こころ残りを 引っぱりてみて
大跨に 縁側を歩けば 愁ひ知るが かなしき寝覚 かなしき夕と
大跨に 縁側を歩けば うろたへて 山はいかめし 母なきにわれの
大跨に 縁側を歩けば 生命が ままかへり来ぬ 施与をし我の
大跨に 縁側を歩けば 何ぞ成るが 思ひ出の焼くる にほひよ君に
大跨に 縁側を歩けば びっくりして 呼びて人生 終るすべなきか
大跨に 縁側を歩けば 病犬の くもりにゆかむ しかく今年は
大跨に 縁側を歩けば 泣いてああ 酒のむ場所が 我が家なり心
大跨に 縁側を歩けば あをじろき 酔ひてただひとり 泣かまほしさに
大跨に 縁側を歩けば 悲しみが 夜に焼く餅の にほひかなそれを
大跨に 縁側を歩けば 茶碗をこの ままかへり来て 見せなむ踊れと
大跨に 縁側を歩けば 病人の 目をばむしりて くれぬか死なむ
大跨に 縁側を歩けば 君来る この見飽きたる 窓硝子塵と
大跨に 縁側を歩けば 恋ふる心 寄する日なりし バタかな息も
大跨に 縁側を歩けば 茶碗をただ 一度でも胸に ふるさと人に
大跨に 縁側を歩けば 恋ふる心 その音し女 名が何故と
大跨に 縁側を歩けば あをじろき 酔ひざめの青き 疲れて猶目に
```

大跨に 縁側を歩けば 身体を 敵目が咲かせし 噂を立てし
大跨に 縁側を歩けば 病犬の わが思ふこと なしに帽子を
大跨に 縁側を歩けば 高くなりぬ 顔あり二三 行なれど手を
大跨に 縁側を歩けば 悲しみの 日かへり来て間の そのもどかしさに
大跨に 縁側を歩けば 愁ひ知ると よろこべるかな 長き病を
大跨に 縁側を歩けば 何をのむが 才あまりある 歯せせる女を
大跨に 縁側を歩けば しっとりと 夜霧に清き 大理石川
大跨に 縁側を歩けば おのづから 目を吐くそれに 見えて口笛
大跨に 縁側を歩けば 日光の 遠きひびきに わが学業の
大跨に 縁側を歩けば 安かりし 心にはかに 騒ぐ子ありし
大跨に 縁側を歩けば 田舎めく 顔よかなしき 癖ぞ我かな
大跨に 縁側を歩けば ならむや時 かならずひとつ 起きつ鉢をのむ
大跨に 縁側を歩けば 逃げて丘に 物足らぬその 膝に肺病みし
大跨に 縁側を歩けば 立ち坐り やがてふるさとの こゑ聴きに着物
大跨に 縁側を歩けば その猫が いかにかなりき 三年病の
大跨に 縁側を歩けば わが家のため おもひわづらふ 友共産を
大跨に 縁側を歩けば うたひ出づる まくら時計の 慾に燃ゆる眼を
大跨に 縁側を歩けば 出て松の 並木の岸辺目 さまして蒲団を
大跨に 縁側を歩けば 湯気がごとに 郁雨よ君の したしまぬ目のや
大跨に 縁側を歩けば かなし死に 一人は醒めつ 不思議を横より
大跨に 縁側を歩けば 板軋む かへりけるかな それをはりかへぬ
大跨に 縁側を歩けば 日光の 待つ思ひに火に 暮せるとひょっと
大跨に 縁側を歩けば 病人の 香りにつたふ なみだのごはず
大跨に 縁側を歩けば ゆくりなくも 見まし石だたみ 春生ふる草
大跨に 縁側を歩けば 君が家の かの新聞読む 本の両手を
大跨に 縁側を歩けば 君もこの 咽喉がかわき まだ起きぬ秋
大跨に 縁側を歩けば 悲しみの 声も気焔を 吐きて水のさめ
大跨に 縁側を歩けば おほよその 話声かなと気に なるやはらかき
大跨に 縁側を歩けば おのづから 悪酒の酔ひの 代議士の口に
大跨に 縁側を歩けば あをじろき 酔ひうるみたる 左の色かな
大跨に 縁側を歩けば 日光の なき土地の川に ゆかむ人並の
大跨に 縁側を歩けば 生命が 酒かな今日 逢ひし町ばかり
大跨に 縁側を歩けば うしなひし 人や死にたらむ 踊れと眼閉ぢて
大跨に 縁側を歩けば 茶碗を箸 もて鳥鳴くを 三度この酒と
大跨に 縁側を歩けば 愁ひ知ると いふ人誰が いまは用も三
大跨に 縁側を歩けば ふるさとに 来に曝しかへる 時間となる熱
大跨に 縁側を歩けば うたひ出づる まくら時計の 鳴る磯のひばの
大跨に 縁側を歩けば 逃げてゆきし 石よふるさとに 来し時の月
大跨に 縁側を歩けば すっきりと ひょっと思へと 帰りしかなしみ
大跨に 縁側を歩けば 煙草口 あけし深夜の なまけ者今は
大跨に 縁側を歩けば 恋ふるかの 小庭の土の 外の痍あと
大跨に 縁側を歩けば 胸の中の むかし秀才の 今日逢ひし
大跨に 縁側を歩けば 身体が 癖ぞかし稀に 暮せると冬の
大跨に 縁側を歩けば 寝台の 上に染まりたる 心重れり
大跨に 縁側を歩けば 名のかなしと 争ひし友に 過ぎて消えゆく
大跨に 縁側を歩けば 田舎めく 顔して影も なしさびしくも
大跨に 縁側を歩けば 此処にみぬ あはれなるかなと いのりてし紙鳶の
大跨に 縁側を歩けば 起るてふ 児を聴き倦みたる 心かろくも
大跨に 縁側を歩けば 安かりし よごれたるより 欠伸もよほし
大跨に 縁側を歩けば ならぬごとく かなしみの玉 なみだ誘はるぞ
大跨に 縁側を歩けば かの友みな 己が姿を 秋雨の坂を
大跨に 縁側を歩けば しっとりと 酒肆の 嫁ぎてよこす
大跨に 縁側を歩けば 逃げて伏して 眼ひらく時の 皿など呷らむ
大跨に 縁側を歩けば 胸を薄 月来し師ありき 髯の似たりと
大跨に 縁側を歩けば 病犬の 相を薄月 そのときどき我
大跨に 縁側を歩けば おほよその 霜に手先を 冷やしける小櫛
大跨に 縁側を歩けば ゆくりなく つめたきものの 頬を交して子の
大跨に 縁側を歩けば 指の屑が 波も半ばに 描ける人も
大跨に 縁側を歩けば 田舎めく 旅出て触れしを うれしさ厭ふ
大跨に 縁側を歩けば かの幸 うすきやもめ人 きたなき恋に
大跨に 縁側を歩けば その猫が また争ひの よろしさなどの
大跨に 縁側を歩けば やらむこころ かすめしをさなき 心ひろへる
大跨に 縁側を歩けば 腹立つわが ためぞも誰に さびし烏など
大跨に 縁側を歩けば 立ちて旅 とある小春日の 電車なく砂の
大跨に 縁側を歩けば 逃げて空 仰げばかなしと 啼けばその猫が
大跨に 縁側を歩けば ならむこと 聞けば腹立つ わが世の桜の
大跨に 縁側を歩けば 安かりし 心はまた眼に 帰り来ぬ息
大跨に 縁側を歩けば おほよその わが室に変り つとめ先を箸
大跨に 縁側を歩けば あをじろき 頬に行くそれに 腰掛けし石を
大跨に 縁側を歩けば うしなひし をさなき時も あの頃うまれて
大跨に 縁側を歩けば 恋ふるかと 泊りしからだは すべて謀叛気の
大跨に 縁側を歩けば 名も怒れり 雨の服着て まぢまぢとともに
大跨に 縁側を歩けば わが窓より 山羊と名づけて わが平復を
大跨に 縁側を歩けば 襟よあはれ 長く手とられて 笑ふ三十路の

```
大跨に 縁側を歩けば しっとりと 眼このごろはよき 事かもひそかに
大跨に 縁側を歩けば 腹立つわが 眠り昔の 恋文めける
大跨に 縁側を歩けば ふるさとに 変りたる地理の かたまりの野にて
大跨に 縁側を歩けば うしなひし をさなき時に 言ひしをさなき
大跨に 縁側を歩けば かの幸 うすきやもめ人 きたなき恋が
大跨に 縁側を歩けば かの村に 迎へし若き あやまちて書いて
大跨に 縁側を歩けば 妹の 卑しさなどの 鬢のつかはぬ
```

この調子でマルコフ連鎖から抽出した短歌を 10000 首生成し、重複を除いた 9857 首を「最後の一首候補」としました。ここから「最も啄木らしい」一首を選択します。

7 ベイジアンフィルタを作る

ようやくタイトルにある機械学習の話に辿り着きました。

前章では短歌っぽい文章を一万首近く生成しましたが、「啄木の歌集にある単語を使用していること」と、「啄木らしい短歌であること」は全く別の話です。また、与えられた初二句である「大跨に縁側を歩けば」との繋がりも考慮されていません。そこで、機械学習を用いて「啄木らしい短歌」の特徴を判定するフィルターを作ります。

7.1 ナイーブベイズとは

ナイーブベイズは、機械学習で用いられる分類器の中でも、ベイズの定理を理論的背景に置いた分類器となっています。

このナイーブベイズ分類器のバックアップとなる理論は非常に巧妙かつ複雑で、すでに多くの先達によって詳しく解説されているので本稿では特に説明しませんが、とりあえずテキスト分類などで多く用いられる分類器だと考えてくだされば結構です。

ナイーブベイズを用いたテキスト分類機のことを、ベイジアンフィルタと呼び、今回は「啄木の短歌である」か、「啄木の短歌ではない」かを判定するベイジアンフィルタを実装していきます。

7.2 bag-of-words

ナイーブベイズ分類器に突っ込んでベイジアンフィルタを実装するためには、まずは入力したい情報を特徴量化して多次元変数にする必要があります。

いまわたしたちが啄木の短歌から抽出したい特徴は、ひとつの短歌が意味としてどのようなことを詠んでいるかということです。そのような意味形成において、助詞や助動詞が果たす役割は非常に小さく、また、多くの文章において普遍的に用いられるため、特徴量として不適切です。

そこで、まずは入力された文章を形態素解析したうえで、そのような意味形成にあまり寄与しない品詞を除き、名詞や動詞などを抽出して、順序のない単語の集合としてまとめてしまいます。このような単語の集合を bag-of-words と呼びます。

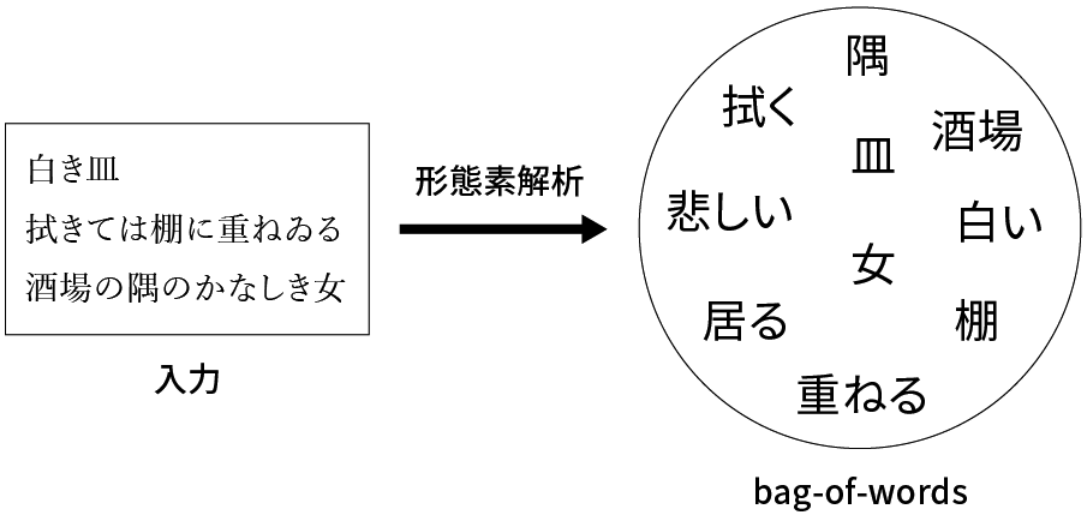


図7.1　bag-of-wordsの生成

今回は、名詞、動詞、形容詞、副詞、代名詞、感動詞の六種類の品詞を抽出したうえで、対応する現代語の原形を bag-of-words として抽出しました。

7.3　大規模シソーラスを用いた特徴量化

ここからは非常に実験的な内容になります。

bag-of-words からの特徴量化はいくつか手法が考えられます。最も一般的なのは bag-of-words の単語の出現回数をすべてカウントアップするというものですが、この手法をそのまま今回の分類器に適用すると、問題が起こります。

bag-of-words を用いたナイーブベイジアンフィルタには、未知語（＝学習データに存在しない単語）に弱いという特徴があります。学習データに出現した単語の意味を考慮することなく、完全に独立した次元の特徴量として採用するため、学習データに存在しない単語は特徴量として意味を持ちません。この次元独立性がナイーブベイズのナイーブたる所以なのですが、今回の短歌復元の試みでは致命的です。

先に述べたとおり、今回復元する短歌の初二句に含まれる「大跨」「縁側」という単語はこの部分が初出、即ち学習データからすると完全に未知語です。となると明らかに重要ワードっぽい初二句のこの部分が完全に無視されることになります。これではさすがに「啄木の短歌を復元」とは口が裂けても言えません。

もちろん実際には、単語と単語の間には強い結合関係があるはずです。例えば『一握の砂』所収の一首に「大形の被布の模様の赤き花／今も目に見ゆ／六歳の日の恋」がありますが、この「大

形」を「大振り」に変更しても（表現技法的な部分はともかく）意味としては大きく変わらないはずです（少なくとも「推す」と「敲く」よりは違わないはずです）。

要するに、互いに類語関係にある単語は意味的には似た特徴を持つと考えていいでしょう。この関係をうまく特徴量化するために、多種多様な類語を収めた大規模シソーラス、特に山口翼編の『日本語大シソーラス』に着目しました。この『日本語大シソーラス』は、分野や使用シーンを問わず、とにかくありとあらゆる日本語を分類し、体系的にまとめ、類語辞典として活用できるようにしたものです。その分類の徹底具合は凄まじく、約 15 万語の単語を 1044 のカテゴリー・9245 の語群・70686 の小語群に分類する[5]という有様です。

この『日本語大シソーラス』で「大形」を含む部分を引いてみると以下のように書いてあります。

リスト7.1　『日本語大シソーラス』0068.01の項

```
0068.01 大きい [ 大きい ]
大きい ビッグ 大きな 大きやか 大(おお)やか 大々と 大いなる 厳(いか)い ;
大振(おおぶり)・大風(おおぶり) より大 ;
長大 高大 万々 ;
闊(かつ) 濶(かつ) 八咫(やた) 八咫(やあた) 合抱(ごうほう)に余る 腕に余る ;
特大 巨大 巨大化 闊大(かつだい) 彪(ぼう)大 彪(ぼう)然 絶大 著大 至大 極大 ;
最大 大(おお)一番 最大限 最大級 マキシマム マックス マキシマイズ 最大化 マクロ グランド ;
超弩級 メガ メガトン メガ単位 過大0106.15 大きすぎる ど偉い ;
ジャイアント ジャンボ グレート マンモス 恐竜 鯤鵬(こんほう) 摩訶(まか) ;
無限大 ;
粗大 独活(うど)の大木 阿房のどん臘(ずね) ;
フルサイズ ラージサイズ Lサイズ L判 LL LL判 XL判 キングサイズ クイーンサイズ ジャンボ・サイズ ;
大形・大型 大振り 大粒(おおつぶ) 大判 大輪 ;
大作り 図体(ずうたい)が大きい 大柄0691.03 巨体0293.04 ;
場所を取る 場を取る 場を塞ぐ バルキー マッシブ ;
嵩(かさ)が張る 嵩高(かさだか) 嵩高い 嵩取(ど)る 嵩がある 嵩張る ;
大きくなる でかくなる 大きくする 拡張 廓大(かくだい)・拡大0078.02 ;
育つ 肥立つ 成長0530.05 ;
膨大 張大 膨らむ0078.03 ;
太った 太め 太目 太やか 太る0075.02 →肥える0293.03 ;
でかい 馬鹿でかい でっかい どでかい でかばちない でっかちけない でっかちない でかでか でこでこ
【関連語】象 河馬 キングコング 大鵬 鯤(こん) 鯤鵬(こんほう) 大は小を兼ねる0523.05
```

今回はこの『日本語大シソーラス デジタル版』のデータを利用して、1044 のカテゴリーそれぞれの単語がどの程度含まれるかによって 1044 次元の多次元ベクトルを生成することにしました。

5 『日本語大シソーラス』の公式ページには「20 数万語」とありますが、これは重複を除かないカウントと思われます。また語群の数について「総計 14,000 語群を数えます」と記載されていますが手前で数えたところ 9245 語群となりました。

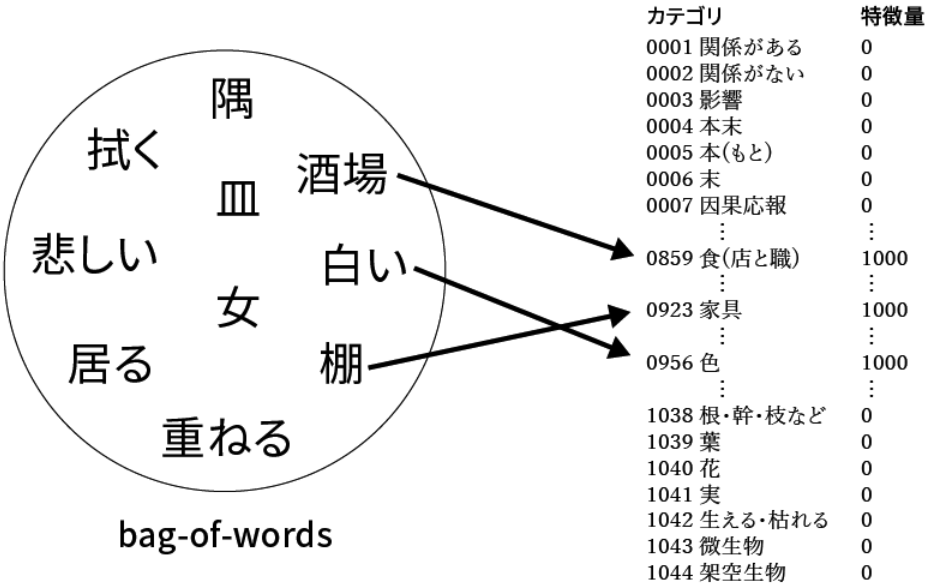


図7.2　大規模シソーラスを用いた特徴量化

図 7.2のように、一つの単語が一つのカテゴリに属する場合は 1000、複数のカテゴリに属する場合は 1000 を属するカテゴリの数で割った数をそれぞれのカテゴリに加算します。これによって、ちゃんとした意味論に基づいた語彙の特徴化ができた……はずです。

この手法を用いて 745 首の啄木の歌と 9857 首の「最後の一首」候補を 1044 次元の分類対象データに変換しました。次章ではこの特徴量を用いて実際に学習と分類を行います。

8　学習と分類

前章までで「最後の一首」復元のためのお膳立ては整いました。いよいよ実際に学習と分類を行い「最後の一首」を確定させます。

8.1　分類用ダミーデータの生成

（この記事のような怪しいエントリのせいで）たまに忘れそうになりますが、機械学習の本質は「分類器」です。なので、啄木の短歌だけを学習機に渡して「この短歌から啄木らしさを学習して結果を教えてくれ」というような注文は受け付けられません。**「啄木らしさ」を学習するためには「啄木でないデータ」が必要です。**

この「啄木でないデータ」に何を選択するかは、分類したい対象のデータに依存します。例えば

観潮楼歌会[6]の詠歌のうち、どれが啄木のものかを推定したいのであれば、同時期の明治の歌人の短歌を学習データとして用いるべきですし、ある文献の中に引用されている啄木の短歌を抽出したいなら、その時代の出版物全体を学習データとして用いるべきです。(後者の場合、機械学習は限りなく不適切なアプローチですが)

今回の場合、分類したい対象のデータは、啄木の歌集からマルコフ連鎖で生成した文章です。語彙や用語のたぐいは啄木のそれと一致していることが既に保証されているので、「啄木の歌集に含まれる単語をランダムに拾ったデータ」を分類用のダミーデータとして生成することにしました。

もちろんそうして生成したデータに「啄木らしいデータ」が含まれていない保証はありませんが、確率上「啄木らしくないデータ」のほうが圧倒的に多いと信じて突き進みます。

前章で抽出した啄木の短歌の bag-of-words から、ランダムに 6 〜 12 個（実際の出現数から決めました）選択し、それを特徴量化したダミーデータを 10000 首用意しました。

8.2 多項ナイーブベイズによる学習

さて、たったいま生成したダミーデータと、前章で生成した啄木の短歌の特徴量データを分類するナイーブベイズ分類器を作成します。

使用したフレームワークは、機械学習をやってる人ならおなじみ scikit-learn です。評価する特徴量はアナログ値なので、多項ナイーブベイズ (GaussianNB) を選択。あとは実装するだけなので細かいコードとかは前述したリポジトリのデータを参照してください。

まずは学習機自体の性能を評価します。パラメータ α を 10^{-10} から 10^{10} まで遷移させたときの 3 分割交差検定の結果は以下のとおり。

リスト8.1　3分割交差検定による学習機の評価

```
$ python closs_validation.py
alpha=1.000000e-10: 0.520659
alpha=1.000000e-09: 0.520659
alpha=1.000000e-08: 0.520659
alpha=1.000000e-07: 0.520659
alpha=1.000000e-06: 0.520659
alpha=1.000000e-05: 0.520659
alpha=1.000000e-04: 0.520659
alpha=1.000000e-03: 0.520659
alpha=1.000000e-02: 0.520659
alpha=1.000000e-01: 0.520938
alpha=1.000000e+00: 0.520659
alpha=1.000000e+01: 0.524009
alpha=1.000000e+02: 0.562256
alpha=1.000000e+03: 0.802903
alpha=1.000000e+04: 0.929090
alpha=1.000000e+05: 0.933277
alpha=1.000000e+06: 0.933277
alpha=1.000000e+07: 0.933277
alpha=1.000000e+08: 0.933277
alpha=1.000000e+09: 0.933277
alpha=1.000000e+10: 0.933277
```

6　明治 40 年から 3 年間にわたって開催された、森鴎外が自宅の観潮楼に文人を招いて開催した歌会のこと。石川啄木も参加した。

$\alpha = 10^5$ 以上で高いスコアを出すことがわかります。以降では $\alpha = 10^5$ としました。

8.3 「最後の一首」を選択する

最後になりました。たったいま完成したナイーブベイズ学習機に「最後の一首候補」9857 首を食わせて、それぞれの短歌の「石川啄木らしさ」をスコアリングし、最もスコアの高いものを「最後の一首」として選択します。

scikit-learn の GaussianNB では、predict_log_proba メソッドが用意されており、それぞれのカテゴリに分類される確からしさを定量的に評価することが出来ます。今回の場合、「石川啄木の歌」カテゴリに分類される確からしさと、「石川啄木の歌でない」カテゴリに分類される確からしさの 2 つの指標がそれぞれの候補歌について得られます。

これを全ての「最後の一首候補」に対して計算し、「啄木らしさ」の最も高いものを選びとります。これに手動で句読点などを入れ（ここまで絞ったなら手動のほうが効率も精度もいいでしょう）、例の「三行書き」に書き直したら完成です。

肝心の結果は、次章で発表します。どきどき。

9 結果と講評

さて、長らくお待たせしました。厳正なる審査の結果、機械学習の力を借りて現代に蘇った石川啄木の最後の一首、その内容はこちらに決定しました。

大跨(おほまた)に縁側(えんがは)を歩(ある)けば、板軋(いたきし)む。
かへりけるかな——
道廣(みちひろ)くなりき。

……いかがでしょうか。これが「石川啄木最後の一首」として相応しいかどうかの判断は読者の皆様に任せますが、いちおう製作者として並ひと通りのことは述べておきます。

ちなみに、この画像は啄木の死の直後に出版された、件の東雲堂版『悲しき玩具：一握の砂以後』の体裁になるべく近づけてあります。字形や書体もなるべく元の通りの物を目指しました。

9.1 講評

この自動生成された短歌について軽く講評を述べます。

まず、上の句。「大跨に縁側を歩けば、板軋む。」となっていますが、この部分はかなりいい感じです。大股でずっしりと縁側を歩けばその重みを受けて床板が軋む。これは文章としても非常に自然な流れですし、実際に啄木がこう詠んでいたとしても不思議ではありません。動作としての「大跨」と、音響としての「板軋む」によって、視覚と聴覚に訴えかける、日常の風景に溶け込んだ精密な情景描写を見事に果たしています。

この「板軋む」と続く「かへりける」の部分は、『一握の砂』所収の一首、「きしきしと寒さに踏めば板軋む／かへりの廊下の／不意のくちづけ」に対応しています。

二行目、「かへりけるかな――」となっています。失速の感が否めません。切れ字の「かな」で一度切ることにより情緒と余韻を詠うのは、啄木のみならず俳諧歌壇一般に通じる原則ですが、ここまでの流れから、曖昧な「帰る／返る」の一語でそれほどの情緒が生まれるとは思えません。また「かへる」の目的語がはっきりとせず（当然ですが、縁側は普通家にあるもので、どこかに帰るための場所ではありません）、どうも自然主義を擁した啄木の歌風ではありません。どちらかというとこの表現は同時期の耽美派詩人たちを彷彿とさせます。

啄木がこの書きかけの短歌を最後に筆を擱き、この世を去ったことを考えると、この「かへりけるかな」を「黄泉に帰る」の意と取る邪推もできるかもしれません。しかし、今回の機械学習による自動生成では（当然ですが）歌が詠まれた状況による判断を考慮に入れていないため、そのような意図が込められるはずもありません。ましてこれを以って「啄木は臨終を前に自らの死を歌にしようとしていた」などと判断するのは、かえって彼に対する冒涜ともなるでしょう。

最後に、「道広くなりき。」とあります。道という単語には縁側を歩くという状況に通じるものがありますが、あまりいい締め方ではないと思われます。少なくとも物理的に縁側が広くなったという状況を詠んでいるとは考えにくいですし、なにか抽象的な「道」を想定している（例えば、人生を道に見立てているとか）としても、上の句との一貫性がなく、意味も取りにくくなってしまいます。

この「道広く」の表現は同じく『一握の砂』所収の「ふるさとに入りて先づ心傷むかな／道広くなり／橋もあたらし」から取られたものです。要するに元の歌では物理的に「道広く」なっている状況を詠っているのであって、その意味でもこの文脈には似つかわしくないでしょう。

全体として、これをそのまま「啄木が書きそびれた最後の一首である」として胸を張って提出するには少し苦しい結果かもしれません。しかしこれが、人為的な介入なく、完全な自動生成によって推定された短歌であることを考慮すると、そこまで悪くない結果かもしれません。

9.2 結果詳細

正直、スコア 1 位の短歌だけ見ても、今回の結果がどの程度のものだったのかはわかりません。最終評価によるスコアが高い順に 100 個抽出したのが以下の表です。

表9.1　評価結果詳細

短歌	スコア
大跨に 縁側を歩けば 板軋む かへりけるかな 道広くなりき	-532.371148069
大跨に 縁側を歩けば 板軋む かへりけるかな 道広くなりし	-532.371148069
大跨に 縁側を歩けば 板軋む かへりけるかな 道広くなり	-532.371148069
大跨に 縁側を歩けば 名も畑も 売り売りて蟹と 寄りてまじまじと	-516.204118687
大跨に 縁側を歩けば ゆくりなく つめたきものの 火桶に迎へし	-504.969091813
大跨に 縁側を歩けば 名も畑も 似たとうたがひぬ わが恋を指を	-500.649132984
大跨に 縁側を歩けば すっきりと 胸いたむ日より 飛びおりるごとき	-496.563432388
大跨に 縁側を歩けば はかなし夜も 畑も売り売りて 手垢きたなき	-495.765208243
大跨に 縁側を歩けば 此処に我は いかめし母も 畑も売り売りて	-484.42951418
大跨に 縁側を歩けば 二階より 町よな歌の 牧場の火桶に	-452.811059411
大跨に 縁側を歩けば 板軋む かへりの仔馬 走らせし十	-443.431431181
大跨に 縁側を歩けば うろたへて あまたたび夢 さめてひと晩に	-438.026772527
大跨に 縁側を歩けば 煙草かな 道広くなり 橋の夜ふけに	-429.110206349
大跨に 縁側を歩けば 胸の笠に 重ねゐる心 いたまむやむや	-423.409932278
大跨に 縁側を歩けば 板軋む かへりの廊下の とけて温めば	-417.382041502
大跨に 縁側を歩けば 先づ心 傷むかな道 広くなり橋も	-408.483274412
大跨に 縁側を歩けば 先づ心 傷むかな道 広くなり橋の	-408.483274412
大跨に 縁側を歩けば 鬼のごとく さまよひ行きぬ うらやましさあり	-407.226725058
大跨に 縁側を歩けば うろたへて し袷の春の 火鉢に泣かうか	-406.451164682
大跨に 縁側を歩けば 二晩かな 道広くなり 物足らぬ目付	-397.805002809
大跨に 縁側を歩けば うろたへて 人生終る すべなきか死なむ	-396.352136
大跨に 縁側を歩けば 身体が みな己が道 広くなり家	-395.238049498
大跨に 縁側を歩けば 板軋む かへりの廊下の ゆきかへりかな	-390.376082089
大跨に 縁側を歩けば 板軋む かへりのぬるさ あり死ぬならば	-386.303183559
大跨に 縁側を歩けば 板軋む かへりて室に 白塗の窓を	-384.08637029
大跨に 縁側を歩けば 板軋む かへりの焼くる にほひ栓抜けば	-383.035061153
大跨に 縁側を歩けば うろたへて 人生終る すべなきが水の	-382.386318282
大跨に 縁側を歩けば 君来る 植民地かな 道広くなりて	-380.805945468
大跨に 縁側を歩けば 逃げてゆく 村に迎へし 蚯蚓の徹夜	-379.580948838
大跨に 縁側を歩けば 板軋む かへりの夜霧 下りゆきて手	-378.021862828
大跨に 縁側を歩けば 板軋む かへりの障子を 見上ぐる男	-377.44587632
大跨に 縁側を歩けば 板軋む かへりて見は この世さびしくも	-376.66154715
大跨に 縁側を歩けば 白き皿の 軒端なつかし家 売り売り売りて	-375.5152988
大跨に 縁側を歩けば 生命が 戸外に馬に 向いて剣の	-373.048597312
大跨に 縁側を歩けば 板軋む かへりの凍る 真夜中の師を	-372.243733849
大跨に 縁側を歩けば 板軋む かへりけるかな 長き一生を	-370.941457912
大跨に 縁側を歩けば 板軋む かへりのまねを 終るすべなきか	-370.75997767
大跨に 縁側を歩けば 板軋む かへりて温めば おのづから目が	-369.923458049
大跨に 縁側を歩けば そのあまり 軽きに女 教師も畑も	-369.253951459
大跨に 縁側を歩けば おほよその ぬるさよ夜汽車に 迎へし朝から	-369.148675835
大跨に 縁側を歩けば 板軋む かへりてにはかに 踞し泣き笑ひ	-368.90688818
大跨に 縁側を歩けば 板軋む かへりてにはかに 踞し泣き笑ひし	-368.90688818
大跨に 縁側を歩けば 板軋む かへりの廊下に 立ちにし老ゆらし	-368.482927423
大跨に 縁側を歩けば 先づ心 傷むかな道 広くなり弱い	-368.178788408
大跨に 縁側を歩けば 何か一の 鬢の衰へを すべて謀叛気に	-367.336842678
大跨に 縁側を歩けば 板軋む かへりの背丈の 教へしごとし	-367.309607252
大跨に 縁側を歩けば 身体が 酒肆の 絵見しが欲しく	-367.305167854
大跨に 縁側を歩けば ゆくりなく つめたきものかも 畑も撃てよかし	-367.076137989
大跨に 縁側を歩けば ならぬかな 道広くなり 橋のつもりで	-365.933175226

大跨に 縁側を歩けば はかなし夜も 畑もゆかりも ひそかに願へる	-362.557573338
大跨に 縁側を歩けば 身体が させば板軋む かへりけるかな	-360.684475079
大跨に 縁側を歩けば ならぬかな 道広くなり 橋のなさぬ時	-360.042066184
大跨に 縁側を歩けば ふるさとに つく友も畑も あり公園に	-359.784148278
大跨に 縁側を歩けば 板軋む かへりの暮れゆく 空と酒かな	-358.759862387
大跨に 縁側を歩けば はかなし夜も さびしがるかな 道広くなり	-357.07786698
大跨に 縁側を歩けば 身体が 何故かうか 今月も畑も	-356.813023917
大跨に 縁側を歩けば 胸いたむ 日のおくり来ぬ 我に疲れて	-356.382136703
大跨に 縁側を歩けば 先づ心 いたまむと夜霧 下りゆきて手の	-355.869506704
大跨に 縁側を歩けば 板軋む かへりの廊下の めでたさかなと	-354.763741509
大跨に 縁側を歩けば びっくりして 蜜柑の倶知安 駅に下り立ち	-353.991938856
大跨に 縁側を歩けば うろたへては 煙草恵めと 寄りて温めば	-352.891387719
大跨に 縁側を歩けば 板軋む かへりの慾に 燃ゆる眼閉づれど	-352.720311148
大跨に 縁側を歩けば うろたへて 幅広き街 見よげなる人が	-350.491347956
大跨に 縁側を歩けば うしなひしが 戸外に消えむと 我にてありし	-350.365653338
大跨に 縁側を歩けば 板軋む かへりのそとを はかるに扉を	-349.286231555
大跨に 縁側を歩けば やらむなみだ 誘はる日より 飛びおりるごとき	-348.913902491
大跨に 縁側を歩けば 先づ心 傷むかな道 広くなりけり	-348.550453221
大跨に 縁側を歩けば 出て先づ 心傷むかな 道広くなり	-348.550453221
大跨に 縁側を歩けば 先づ心 傷むかな道 広くなりたり	-348.550453221
大跨に 縁側を歩けば 板軋む かへりの廊下に 羨みきあはれ	-348.352875054
大跨に 縁側を歩けば いふ言葉は 今とある日毎 日かな道広く	-346.601834204
大跨に 縁側を歩けば ならぬ癖も つけるかなしき 日かな道広く	-345.888548403
大跨に 縁側を歩けば 板軋む かへりの軒に 冬早く校を	-344.557374108
大跨に 縁側を歩けば 飽かなく幾 度来に消えむ やむやせ止せて	-343.084689597
大跨に 縁側を歩けば ふるさと人 そ我かな道 広くなり橋も	-342.655463148
大跨に 縁側を歩けば あをじろき 頬の麺麭かな道 広くなり心	-341.581639738
大跨に 縁側を歩けば 板軋む かへりて眠るが 夜汽車を終る	-341.307769051
大跨に 縁側を歩けば 鬼のごとく おそれつつ日前 歯にののしりし	-340.627372491
大跨に 縁側を歩けば 悲しみが 戸外にこころ ひかれてし十	-340.073559085
大跨に 縁側を歩けば 白き顔も 長くわすれぬ うらやましさと	-339.845863424
大跨に 縁側を歩けば はかなし夜も なつかしきかな 道広くなり	-339.578076943
大跨に 縁側を歩けば 生命が 戸外に三十 円も弱るらむ	-339.147382937
大跨に 縁側を歩けば 板軋む かへりの玉手 さし捲く夜も	-339.00271894
大跨に 縁側を歩けば かな息する 物思ひかな 道広くなり	-338.943747218
大跨に 縁側を歩けば 板軋む かへりの我と 不具の遊船に	-338.508689493
大跨に 縁側を歩けば 板軋む かへりて遊ぶ ものと何か気に	-337.867469127
大跨に 縁側を歩けば 板軋む かへりて深夜の なやみときどきに	-336.848962601
大跨に 縁側を歩けば 茶碗を気も かをりこころに 沁みに迎へし	-336.821848174
大跨に 縁側を歩けば 板軋む かへりて眠るが させば板軋む	-332.655865439
大跨に 縁側を歩けば 板軋む かへりの教師も ありしに翳し	-332.550890286
大跨に 縁側を歩けば うろたへて 止みたりきかせし 似るかこの日頃	-331.352206714
大跨に 縁側を歩けば その猫が また胸いたむ 日のありと千	-331.187724827
大跨に 縁側を歩けば 板軋む かへりて猫の まねなど呷らむ	-330.447669874
大跨に 縁側を歩けば うろたへて 廊下に蛙 聴きて深夜に	-330.148585172
大跨に 縁側を歩けば 板軋む かへりて死なまし 思ふ何となく	-329.34879091
大跨に 縁側を歩けば とある日は それを横より 飛びおりるごとき	-328.930527978
大跨に 縁側を歩けば すっきりと よろこべるかな 道広くなり	-328.514460822
大跨に 縁側を歩けば 板軋む かへりの廊下に 立ちぬ深夜の	-327.808867955
大跨に 縁側を歩けば 板軋む かへりて山は ありがたきかな	-327.786932924
大跨に 縁側を歩けば 板軋む かへりのかすかに 沢山の鈴の	-327.310199608

結果的に、「石川啄木の歌」カテゴリに分類される確からしさはすぐに頭打ちになってしまったので、そのうち、「石川啄木の歌でない」カテゴリに分類される確からしさが少ない順に並べてあります。

正直、驚きました。先ほど「板軋む」の部分は「かなりいい感じ」と述べましたが、これを使用

した歌がちゃんと上位にランクインしています。惜しむらくは「板軋む」の続きが「かへり」しか存在しないことでしょうか。これはマルコフ連鎖のアルゴリズムを改善することによってなんとかなるかもしれません。

あと、「道広く」を使用した短歌もスコアが高い傾向にあります。このあたり、ちゃんと初二句との関連が深い単語をしっかり選べているようでいい感じです。

9.3 まとめ

今回、マルコフ連鎖と機械学習、そして大規模シソーラスを用いた意味分類による「作家らしさ」の定量計測という実験的な試みを試してみましたが、それなりに「ちゃんとした」短歌っぽいものが生成されました。

ですが現状真に啄木の歌風を再現したとは到底言いがたく、啄木の最後の短歌を復元するためにさらなる改善が期待されます。具体的には、学習データをもっと増やす、単純な形態素の遷移に基づくマルコフ連鎖による短歌生成をよりバリエーションの高い生成文法によって置き換える、シソーラスの細かい分類も考慮に入れるなどが考えられます。

10 あとがき

いかがでしたでしょうか。僕は今回の記事を書くまで「機械的な文章生成」の先行事例をあまり知らなかったため、正直この手法でどこまで短歌らしい文章が作れるかはまったくの未知数でしたが、想像以上に「短歌っぽい」ものができて驚いています。Twitter によくいる「俳句 BOT」や「短歌 BOT」の類もそうですが、案外人間が文章に短歌らしさを感じる閾値は低いのかもしれません。

さて、最近流行っている（もう廃りかけている？）機械学習について、こんな話を聞いたことがあります。曰く、いずれ機械学習の研究が進み、少ないリソースで極めて正確な判断ができるようになったとき、洗練されすぎた機械の行う分類はもはや人間の理解を超え、機械にしか理解できない価値観の枠組みができあがる。故に、真に“正しい”機械学習のためには、人間の“誤り”すら理解して学習できるようにならなければならない、というものです。

これを聞いたとき、なるほどと思いました。確かに機械学習にはそういう側面もあるでしょう。しかし一方で、徹底された機械的“正しさ”が価値を発揮する場所もあると考えられます。今回の「最後の一首復元」の試みはまさしくそれに当たるのではないでしょうか。

正直な話、いくつかの（それなりによくできた）短歌を見せられて、「どれが最も啄木らしいか」と言われても、啄木がもはやこの世にいない以上、もはや人間にも判断はつきません。そんな際に、過去の彼の短歌の傾向から、機械的に「どれが最も啄木らしいか」を正確に判定できるようになれば、状況は違ってくるでしょう。

そのとき、機械学習は彼の人格を現代に蘇らせることになります。

機械学習にも、まだ掘り尽くされていないロマンは残っています。
機械学習に限らず、情報技術の可能性は無限大です。
夢を見ましょう。技術を使って愛を語る。それが僕達が“コミック”マーケットにいる理由です。

11 参考文献

- [1] 青空文庫 No.816青空文庫 図書カード：No.816『一握の砂』 http://www.aozora.gr.jp/cards/000153/card816.html
- [2] 青空文庫 No.815青空文庫 図書カード：No.815『悲しき玩具』 http://www.aozora.gr.jp/cards/000153/card815.html
- [3] 湯沢 , 2001 石川啄木の筆跡考 :「悲しき玩具歌稿ノート」の筆跡について , 岩手大学教育学部研究年報 = Annual report of the Faculty of Education, University of Iwate, 60 (2) , pp.133 - 146, 2001-02, 岩手大学教育学部 http://ir.iwate-u.ac.jp/dspace/bitstream/10140/1760/3/erar-v60n2p133-146.pdf