提出日１月２０日

ダジャレチャットボットで
考える自然言語処理

プロジェクト(赤石)-2022 秋レポート
担当教員：赤石美奈
法政大学情報科学部コンピュータ科学科２年
学籍番号：21K0134
氏名：日置遼平
Email：ryohei.hioki.5n@stu.hosei.ac.jp

§１．はじめに

今日、チャットボットやおしゃべりロボット、お絵かきAIなど、ユーザからの何らかの言葉の入力に対して、ユーザが期待する応答ができる人工知能が求められている。それらには、どうすれば応答の精度を高くすることができるのかという問題がある。このレポートでは、実際に自然言語処理を行うプログラムを作り、その問題点について考える。

§２．研究目的

今回の研究目的は、ユーザが入力した単語を使ってダジャレを返す、ダジャレチャットボットの作成を行いながら、ダジャレ（言葉）をコンピュータに理解させるにはどうすればいいか、また自然言語処理を行うプログラムにはどのような問題点があるのかについて考え、その解決方法を探すことである。すでに自然言語処理を行うプログラムを使った高度なものや、それに関する研究発表が世の中にたくさん存在するのを分かった上で、解決方法を探す手段としてダジャレチャットボットの作成を選んだ理由は、実際に作ってみたほうがたくさんの問題を発見することができ、さらにその問題について深く考えることもできると思ったのと、ダジャレが個人的に好きだからである。

§３．研究方法

ここで、ダジャレチャットボットを作るために必要な機能について述べる。応答のダジャレに関しては、ボットが自分の能力でダジャレを生成するという高度なものではなく、ファイルから読み込んだダジャレを使って応答するという仕組みにした。また、ユーザが入力した単語を使ってダジャレを返すだけでなく、ユーザがダジャレを入力してボットにダジャレを覚えさせる機能も追加することにした。この応答と記憶の２つの機能を実現するためには、英語をカタカナに変換、漢字や平仮名カタカナの変換、ダジャレに単語が含まれているかの確認、入力された単語をキーとして使ったダジャレになっているかの確認というような機能が必要だと考えた。

変換の機能が必要な理由は、ダジャレに単語が含まれているかの確認やダジャレになっているかの確認をする際に、文字が一致するかどうかで判定する部分があるので、文字の形式を１つの形（今回はカタカナ）に統一してプログラムを動かすことが必要になる場合があるからである。また、ダジャレに単語が含まれているかの確認が必要になる理由は、ダジャレのリストから応答として使うものを抽出しなければならないからである。この機能だけだとユーザの入力した単語をキーとしたダジャレではなく、ただダジャレの文の中に偶然そのキーが含まれているだけの、別の単語をキーとしたダジャレを応答してしまう可能性があるので、ユーザが入力した単語をキーとして使ったダジャレになっているかの確認が必要になる。この機能は、ダジャレをボットに覚えさせるときに入力した文章がダジャレに

なっているかどうかの判定にも用いる。変換の機能は主にライブラリを用いて行ったため説明を省略し、その他の重要な機能の詳細についてのみ後述する。

§４．結果(成果)

まず実行結果を示す前に、最終的にどのような機能の実装になったのかについて詳細に説明する。

４－１．ダジャレに単語が含まれているかを確かめる機能について

①まずファイルから読み込んだダジャレをリストに保存しておき、入力された単語がリストのダジャレに含まれているかを確認する。含まれていれば候補に挙げて、その中からランダムに１つ出力する。

②候補が出なければ、入力された単語を平仮名からカタカナ、もしくはカタカナから平仮名に変換して同様の処理を行う。

③それでも候補が出なければ、漢字も含めてリストのダジャレをすべてカタカナに変換し、入力された単語はカタカナに変換する。ただし、入力の漢字は変換しない。

②は、平仮名かカタカナかの違いだけで、リストのダジャレとユーザの入力した単語が一致している場合のための処理である。③は、リストのダジャレのキーが漢字であるという違いだけで、その意味とユーザの意図した単語が一致している場合のための処理である。注意点としては、入力された単語が平仮名やカタカナの場合は単語の意味が１つに定まらず、文字を読み取るだけでその意味を１つに絞ることはできないため、ユーザの意図した単語と一致したダジャレを出力できない可能性があるのを問題ないとしている。入力された単語の漢字を変換しないのは、ユーザが意図した単語の意味は１つに定まっており、その意味が変わってしまうのを防ぐためである。例えば、以下の図１の場合である。

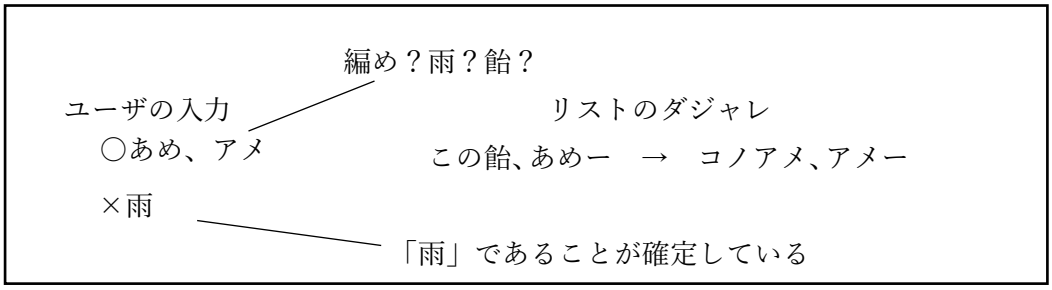

図１：変換の説明図

また、文章ではなく単語を入力としているので、形態素解析を用いて単語を抽出するようなことは行わなかった。修飾語がついている単語でもその修飾語付きの単語を使ったダジャレをユーザが求めていると考える。

４－２．入力をキーとして使ったダジャレになっているかを確かめる機能について

①キーとダジャレをカタカナに変換する。
②ダジャレの１文字目からキー＋キーの半分の文字数だけ読んで、キーの各文字と一致するか確かめる。ただし、文字が一致する順番は適切でなければならない。
③キーの半分より多くの文字数が一致したらカウントを１プラスする。ただし、１文字しか一致していない場合はカウントをプラスしない。
④カウントをプラスしなかった場合は、最初に読んだ文字の次の文字から②を行う。カウントをプラスした場合は、読み始めたところから、(読んだ文字数)－(最後に連続して一致しなかった文字数)だけ先の文字から②を行う。
⑤これをダジャレの最後の文字を読むまで繰り返して行い、終了時にカウントが２以上だったら入力をキーとして使ったダジャレになっていると判定する。
　ダジャレかどうかを判定するには、文の読みを調べないといけないので①の処理が必要になる。②のキー＋キーの半分という文字数にした理由は、キーが少し違った形でダジャレに登場しても高々キーの２倍の文字数になり、２倍の文字数を読むと２回目に現れるキーと被ってしまうことがあるからである。例えば、以下の図２の場合である。

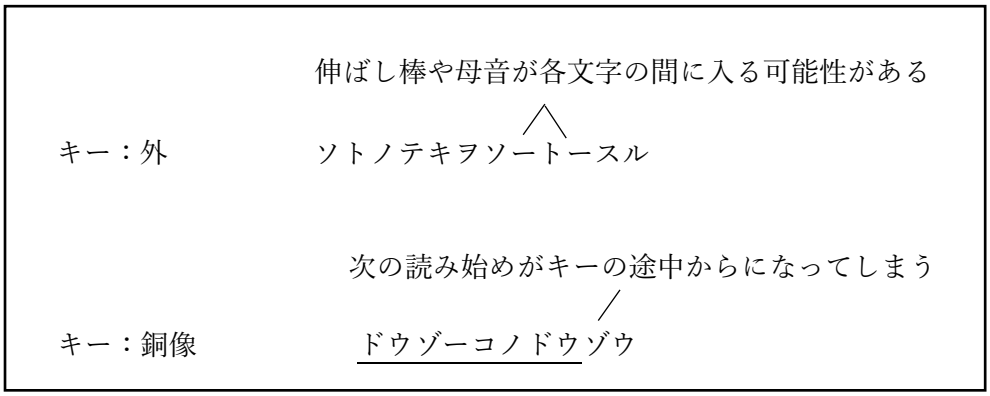

図２：１回あたりの読む文字数の説明図

　また、ダジャレの捉え方は人それぞれ異なるので、③の一致度に関しては緩めに設定した。例えば「５文字のキーに対して３文字が同じような読み方だったらダジャレである」と判断する人もいるということである。④に関して、ほぼ同じ文字を読んでキーを重複して数えないようにするために、カウントをプラスした場合には読んだ文字数だけスキップする。ただし、②で少しのマージをとって読んでいるせいで、最後の方に全くキーと関係ない文字を読んでいたり、２回目に登場するキーを読んでしまったりする可能性があるので、最後に一致しなかった文字数分だけスキップする回数を減らしている。⑤で行っていることを言い換えると、キーと同じような読み方をする文字列が文中に２回以上登場する文のことをダジャレとしている、ということである。

４－３．ダジャレを覚えさせる機能について

ダジャレの入力をリストに追加してファイルに書き込めば、ダジャレを覚えさせることはできるのだが、その際に入力がダジャレかどうかを判定するようにした。方法としては、４－２の機能を応用して使うことで実現する。キーとなるのは名詞か動詞なので、形態素解析を用いて名詞と動詞を抽出し、それぞれをキーとして４－２の処理を行えばダジャレであるかどうかを判定できる。しかし、実際にユーザがダジャレを入力するときには、高度なダジャレや複雑なダジャレが入力されて形態素解析が完璧に機能しない可能性がある。そこで今回は、入力された文の中にある２文字以上の文字列全部をキーとしてダジャレ判定処理を行うことにした。この文字数は漢字を読みに変換した上での文字数である。この場合、同じ２文字が２回以上登場さえすればダジャレだと判定されてしまうのだが、前述した通り、ダジャレの捉え方は人それぞれであって、ダジャレの判定は緩く行うべきであるため、問題ないとした。

４－４．実行結果

ダジャレチャットボットのプログラムは、インターフェースとして tkinter を用いて作成し、五十音順のダジャレの一覧とダジャレかどうかの判定をした時のキーが IDLE 上で表示されるようになっている。その実行画面が以下の図である。

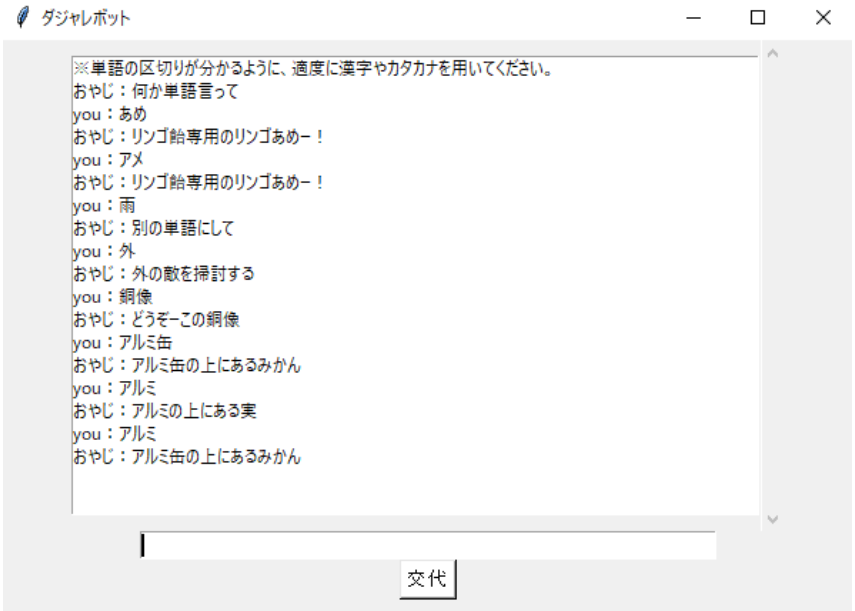

図３：ダジャレ応答の実行画面

ユーザが入力した単語に応じたダジャレを出力していることが分かる。また、図１や図２の説明にあったようなダジャレもきちんと出力されている。複数候補が出ている場合もランダムに両方出力されている。

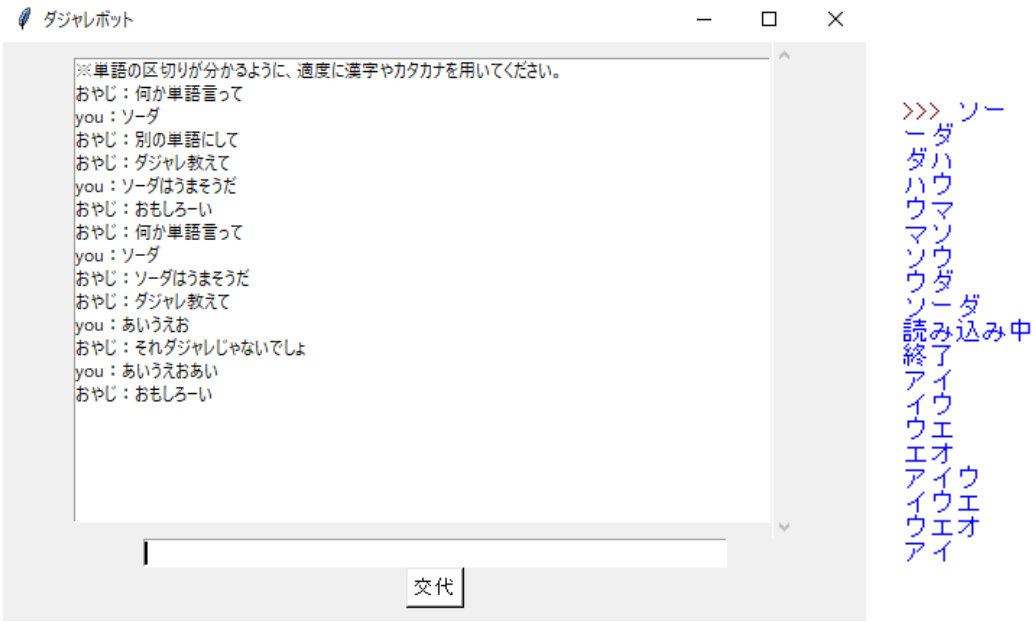

図４：ダジャレ入力の実行画面と IDLE 画面

画面下の交代ボタンを押すと、ユーザがダジャレを入力するモードに移行することができる。図５では「ソーダ」をキーとしてダジャレが記憶されていることが IDLE 画面から分かり、追加する前は「ソーダ」と入力してもダジャレが出力されなかったのが、出力されるようになっていることが分かる。また、適当な文字列を入力してもダジャレとして記憶されないようにはなっているものの、同じ文字列が並んでいる文を入力すると、意味が通らないものでもダジャレと判定されてしまっている。

§５．考察

まず、今回作成したダジャレチャットボットの改善点について考察する。ダジャレに単語が含まれているかを確かめる機能について、入力が平仮名やカタカナだと単語の意味が１つに定まらないという問題点があったが、もし音声入力でイントネーションまで読み取ることができたら、ある程度ユーザの意図した単語を１つに定めやすくなると考えられる。入力をキーとして使ったダジャレになっているかを確かめる機能については、キーと同じような読み方をする文字列が文中に２回以上登場する文のことをダジャレと定義して、少しのマージをとって文字を読んでいき、文字の関係性ではなく、文字の一致や順序、余分な文字がどれくらい含まれているかに注目してダジャレを判定するようになっていた。しかし、このダジャレの定義は曖昧であるため、しっかりダジャレになっているのにダジャレと判定されないという可能性があると考えられる。そこで、文字の一致や順序に関してマージをとるのではなく「句読点や“っ”が途中であってもいい」「こういう文字の後にはこの母音や伸ばし棒が続いても問題がない」など、ダジャレの定義をもっと厳密に定義して判定する機能を作る必要があるだろう。さらに、実行結果でもあったように、同じ文字列が並んでいる

文を入力すると、意味が通らないものでもダジャレと判定されてしまっていたので、しっかり文として成立しているのか、または名詞と名詞が合わさった単語（例：栗のクリーム）になっているかを構文解析する必要もあるだろう。

次に、自然言語処理を行うプログラムについての考察をする。先ほど、意味が通らない文についても考えなければならないことについて言及したが、ユーザが勝手に言葉を増やしたり変に言葉を学習させたりするのを防ぐために、現実に存在する単語なのかどうかについても考えなければならないだろう。これは、入力があるたびに検索をかけたり、あらかじめ言葉の辞書のようなものを作っておいたりすることで解決できると考えられる。しかし、実行時間が増加したり、とても膨大なメモリが必要になったりするというデメリットもあるだろう。また、言葉の捉え方についても考える必要があるだろう。日本人と英語圏の人では文の捉え方が全く違うし、ダジャレの定義の話でもあったように、言葉や概念の定義が人それぞれで異なる場合があるからだ。

よって、自然言語処理を行うプログラムについて、はっきりとした正しいプログラムを作ることは難しいと考えられる。しかし、存在する言葉としない言葉の区別や、言葉や文の捉え方、概念の定義についての研究が進めば進むほど、より良い自然言語処理を行うプログラムが作れるようになるだろう。もちろんダジャレチャットボットにおいても、より正確な判定ができるようになるだろうし、ボットが自分の能力でダジャレを生成して出力するものも作れるかもしれない。

§６．おわりに

今回、ダジャレチャットボットを作成しその実装、改善点などについて深く考え、さらに自然言語処理を行うプログラムについての考察を行った。そして多くの問題点を発見することができた。次は、自然言語処理に関する文献を読んで知識を身に付け、より良いダジャレチャットボットの作成に挑戦したい。

参考文献

・CIS Moodle：プロジェクト(赤石)-2022 秋、UI サンプル、CIS Moodle(オンライン)、入手先　〈https://cms.cis.k.hosei.ac.jp/pluginfile.php/42511/mod_resource/content/2/チャットボット UIsample.pdf〉

・Qlita：python 文字列変換　[カタカナ⇔ひらがな]、Qlita(オンライン)、入手先〈https://qiita.com/mocha_xx/items/07465240d4212d946148〉

・YUMARU BLOG：[Python]　日本語のテキストをひらがな、カナ、ローマ字に変換する、YUMARU BLOG(オンライン)、入手先　〈https://yumarublog.com/python/japanese-text-hira-kana-romaji/〉