Contents

List of Figures

1 序論

1.1 研究背景及び目的

ある目的に対してより完璧に (accuracy が高くなるように) 命令を実行をする Artificial Intelligence が求められている昨今の AI 競争の時代に対し、自然言語処理やゲーム AI のようなタスクは極めて複雑な課題を抱えている。例えばそれは “言葉” という問題である。これは人間がコンピュータに正解となるものを提供することが極めて難しく、“なんとなく良い感じに” 目的を達成してくれることを期待することが多い。この問題に対処するための手段として、入手でき得る限りの大規模なデータを用意して中心極限定理的に尤もらしい中心部を得る方法や、とにかく何らかの単一のモデルに押し込めて問題を解くという方法 [1] がある。それに対して、データそのものを一旦精査・前処理すること、問題を整理・分解しそれぞれを解くことも研究 [2] として存在している。

自然言語処理の、特に対話システムについて考えたとき、小問題に分割した上で対話システムを達成した例として、例として Amazon Alexa Prize [3] というコンテストや Microsoft 社が研究・開発している “りんな” [4] を挙げることができる。これらは対話を行うという問題に対して小さな部分問題を解くタスクを設定し、それぞれを組み合わせることで元の問題を解くというスタイルを取っている。

本研究ではこれらを参考に、日本語の対話システムを作成するという問題に対して小問題を設定しそれを解くための手法を提案・実験する。またその前準備としてデータ収集に絡めて日本語データとその前処理について考察する。尚本研究が最終的に望むものは、キャラクタ性を持った対話可能なエージェントを作ることであることを強調する。

もう少し言及すると、本研究ではデータからモデルにかけて 5 つの少テーマについて研究を行った。概要をそれぞれ説明すると以下のようになる。

1. 日本語データの取り扱いについて
   我々は一般に日本語を話しており、それを用いた対話システムの構築が本研究の主目的である。しかし機械学習等のデータセットや実験で多く使われているのは、日本語とは使っている文字や文型で大きく異なっている、英語のものが多い。その前提のもとで日本語のデータ、特にセンテンスに対して、どのような性質があるのかを調査し、また提案する漢字→ひらがな変換という前処理とそれによって得られる性質についても議論を行う。

2. 文抽出を念頭においた不均衡分散・サイズの分類問題
   テキストのカテゴリ分類を考えたとき、一般には n 個のカテゴリの中から任意の文が入力されることを想定している。今回はそれとはやや問題設定が異なり、いくつかの文をカテゴリ $1 \cdots n-1$、それ以外をカテゴリ n として扱うことを主問題した。しかしデータを設定・収集することが困難であったため、一部画像認識の問題として議論を行う。

3. 機械翻訳システムを用いた対話モデル
   一対一対話を行う際に、機械翻訳システムを用いることがある。今回はそれを、問題を文脈に依存しない発話に対する反応を学習することに再設定し、Transformer という先日 [5] State-Of-The-Art を獲得した機械翻訳の手法を用いて実験、有名な手法である Sequence to Sequence Attention を用いた一対一対話モデルと比較を行う。

4. 文のスタイル変換
   文のスタイルとは、例えば口調や訛り、書き言葉や話し言葉といったものを指す。これは日本語で特に顕著に見られるもので、テキスト上でもこれを確認することで相手のペルソナをある程度想定することができる。本研究が日本語を対象としていること、キャラクタ性を持たせたいというモチベーションがあることから文に特定のスタイルを持たせることを問題として取り上げる。

5. CoLA タスクを応用した対話システムのエラー検知
   対話システムを小問題に分割して解く弊害として、それぞれの問題でエラー (不適切な出力) が出てしまうというものがある。これに対処するため、特に何らかのモデルから生成された文に対しそれが自然であるかどうかを評価するモデルを作成し実験する。

[1] HRED (Sordoni et al. 2015) や VHRED (Serban et al. 2016) があるが、発話の多様性を得ること (一般的な受け答えを学んでしまい、同じような文ばかり生成してしまう) やデータを十分に集めることが難しいなど課題がある。

[2] 日本で人気を得ている “マルチモーダルエージェント AI” とは、複数のソースから問題を見直すという特徴があるが、これは複数のモデルを使っているという意味で同じではあるが、問題を分割しようとしているわけではないという点でこの研究と大きく違うと言えるだろう。

[3] https://developer.amazon.com/alexaprize

[4] https://twitter.com/ms_rinna

[5] 2017 年 12 月時点

1.2 本論文の構成

第 1 章に本論文の概要とその構成について説明を行い、第 2 章で関連研究を紹介し、第 3 章で本研究で掲げるシステムの全体像を示す。そして第 4 章から第 9 章にかけては 1.1 で述べたテーマについての順に議論する。その後付録として補足をまとめたものを 第 10 章 として示す。最後に議論として本論文のまとめ、今後の展望について述べる。

2 対話システムの関連研究

対話システムの関連研究としては、1.1 で述べたように Amazon Alexa Prize というコンテストや、Microsoft 社のりんなを挙げることができるだろう。Microsoft 社のりんなは日本語雑談対話 (Wo et al. 2016) を実現しており、2018 年現在 Twitter などで活動をしている。Amazon Alexa Prize は Amazon Alexa という音声会話を行うことのできる端末に搭載する対話システムを競う大会である。評価対象はユーザの印象であり、別の指標として対話時間が公開される。2018 年度の Amazon Alexa Prize では平均 10 分程度の対話を行うことの出来たシステム [6] が優勝した。顔や体といったテキスト以外の情報を用いることの出来ない対話システムでこのような結果が得られたことは注目すべきことである。

いずれも複数のモデルを組み合わせて構成されており、例えば言語理解部と文生成部、そして本研究で取り扱わないものとしては、音声理解部と音声生成部を挙げることができる。またりんなに関してはそれに加えて画像認識部などの対話以外の [7] システムも構築している [Figure 1] 。

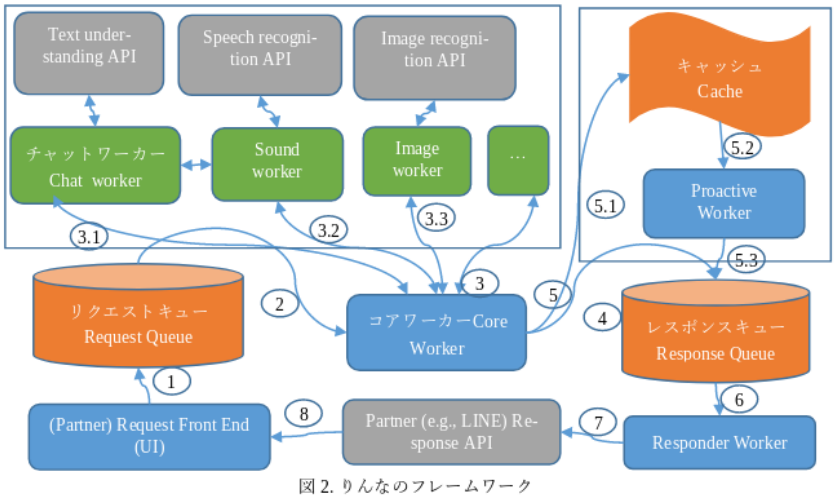

図 2. りんなのフレームワーク

Figure 1: りんなのフレームワーク

[6] 2018 年度優勝は カルフォルニア大学デイビス校のチームが開発したの Gunrock というシステムであり、また 2017 年度優勝はワシントン大学のチームが開発した Sounding Board というシステムである。この 2 つについての詳細は 9.1 で紹介する。なぜこれらを追実装しなかったのかという疑問もあるかもしれないが、いずれも大規模なデータを必要とする (例えば 10M を超える会話データ) ため、個人でそれを実装することは不可能である。

[7] 対話をテキストやそれを示す音声のみのコミュニケーションと定義した場合。実際には対話には身振り手振り、表情といった要素が複雑に絡んでいる。そのため 2017 年頃からは、表情を考慮した対話システムが提案され (Chu, Li, and Fidler 2018) 研究されている。

3 想定する対話システムの全体像

以下に本研究で想定する対話システムの全体像を示す [Figure 2]。

このシステムでは入力としてテキストと、環境情報を得る。このシステムにおける環境情報とはこのシステムが組み込まれているエージェントが居る場所の環境 (天候や気温・湿度)、エージェントの内部状態 (メモリ使用率等) を指す。これはテキストを用いた人対人の対話をイメージしたもので、つまり相手の居る環境、相手の体調をそれぞれ置き換えたものになる。また Answer Generation に用いる所謂個人データのようなものもエージェントの内部に持っているものとする。本論文で扱うものは、この内の Sentence Detection / Sentence Categorization / Topic Dialogue / Style Transfer である。また Topic Dialogue から Style Transfer への矢印・Answer Generation から Style Transfer への矢印・Style Transfer から Output への矢印におけるエラー検知についても議論する。

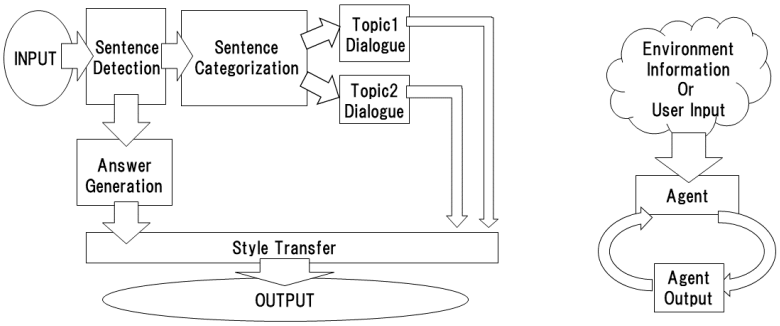

Figure 2: 本研究のシステム全体像

- Sentence Detection [該当部:文抽出を念頭においた不均衡分散・サイズの分類問題]
  ある特定の文を取り出す。取り出された場合はどの意味として取り出されたのかという情報とともに、Answer Generation へ向かい、取り出されなかった場合には付加情報なしで Sentence Categorization へ入力を受け流す。 最終的にはほとんどの文をここで抽出し、それに対する返答を Answer Generation でエージェントの内部状態ないし外部知識ベースを参照しながら生成する。

- Sentence Categorization [該当部:日本語データの取り扱いについて]
  文を大雑把にカテゴリ分類する。例えばそれは livedoor news corpus [8] で議論されるような スポーツ/IT/家電 といったようなカテゴリである。ここでカテゴリ分類された文はそれぞれ対応する Topic Dialogue に流される。

- Topic Dialogue [該当部:機械翻訳システムを用いた対話モデル]
  与えられたカテゴリに対する一対一応答を行う。例えばゲームについての話題を受け持つ Topic Dialogue はゲームに関する入力文を期待しており、それに対する出力を学習しているものとする。そのモデルはエージェントのペルソナに応じて置換することが可能であり、例えば好きなゲームカテゴリについての好意的なデータを多分に含んだデータセットで訓練した Topic Dialogue はそのゲームカテゴリが好きな (好きになった) エージェントが持つことになる。

- Style Transfer [該当部:文のスタイル変換]
  文のスタイルを変換する。ここで言う文のスタイルとは例えば書き言葉や話し言葉、各ペルソナに基づいた語尾変化を示す。

- エラー検知についての議論 [該当部:CoLA タスクを応用した対話システムのエラー検知]
  上記のシステムで発生するエラーデータと正常なデータを分類する。

[8] https://www.roundhuit.com/download.html#ldcc

4　日本語データの取り扱いについて

日本語データは英語データに比べていくつかの問題を抱えている。問題の例としては、文字の数が多すぎること、スペースといった意味ごとの分割がないこと、容易にペルソナを特定できるような多彩な語尾変化があること、多国語も日本語であるかのように用いること、同意同音の語でも様々な表記方法があることが挙げられる[9]。

また一般に公開されている対話データセットを対話テキストのみで学習させると想定したとき、背景知識の欠如を指摘せざるを得ない。更に言えば日本人の特徴として “言外にわかり合う” というコミュニケーションスタイルも問題を難しくしていると言えるだろう。

この章では上記の問題があることを公開されているデータセット、Twitter から収集したデータセットを用いて調査するとともに、“漢字をかなに変換する” という前処理を用いることでどのようにデータの性質が変化するのかを、単語分散を得るというタスクについて実験する。

4.1　調査) 発話データ

発話データとして、2018 年 12 月 25 日 23:00 頃 から翌 26 日 10:00 頃 までに収集した 7 万件の Twitter データを収集し、その性質を観測した。

データの収集手法としては Twitter 社が公開している API を用い、日本のユーザから呟かれている内容を集めるものとした。この処理によって生データが 77,285 発話得られた。

4.1.1　フィルタ

データを収集するにあたり、タグや宛名、URL リンクと言った Twitter に特有な部分を省いた。その上で、4 文字以上、60 文字以下のデータをすべて抽出し、データを 54,368 発話にした。

Twitter に特有な部分を省いた理由として、全体の目的から考えて Twitter データに特化させる必要がなかったこと、タグは時系列で発生・消滅すること、宛名に関してはそのユーザの背景情報が必要になることが容易に想像できること、URL リンクを発話として認めるべきではないと考えたこと[10]を挙げる。

また文字数でフィルタを行った理由として、1. 4 文字未満のデータは少なく、この後議論する単語分割が出来ないようなデータ、それのみでは意味が通じないデータが多く含まれていたこと、2. 60 字超過のデータは何らかの内容に対する説明と言った発話データとはややベクトルの異なるデータが多かったこと、深層学習を中心とした機械学習を用いた自然言語処理 (要約タスクを除く) に用いるデータであると考えたとき、長すぎるテキストは短くされる前処理を施すことが一般的であること、を挙げる。

Table 1: 適用したフィルタとその理由

フィルタの概要	詳細	理由
Twitter 特有の内容		
	タグ	時系列で発生・消滅するため
	宛名	宛名のユーザに対する情報が必要であるため
	URL リンク	リンクを発話として認めるべきか議論の余地があるため
文字数		
	4 文字未満	データ数が少なかったため
		単語分割が出来ないため (極端な略語など)
	60 文字超過	発話データというよりは説明のようなデータが多かったため
		適用する予定の手法では情報の一部が切り落とされてしまうため

4.1.2　調査結果

調査結果を表を用いて示す。そして後述の実験である極性判定実験のために抽出できたデータが 10% 程度であったことを説明する。

[9]前 2 つに関しては、中国語も共通して抱えている問題と言える。

[10]勿論タグに意味が込められている例 (“#○○を許すな” など) も多く見られたが、タグを認めるとタグのあるすべてのデータを手動で確認する必要があったため今回はすべて省いた。

4.2 調査) 対話データ

対話データとして、Twitter のデータ、一般公開されている書き起こしの対話コーパスの内容について言及し、前者に比べ後者は文字だけでは学習することが難しい (背景知識が必要である) ことを説明する。

4.3 問題設定

英語では単語分散を得るために space で区切られた単語ごとに id を振る手法が有名であったが、最近では単語の一部 subword を用いる手法が出てきている。その例として google の出した wordpiece があることを紹介する。 (単語分散を得る際に、日本語は英語と違って、単語ごとに分割されていないことを上げ、WordPirce SentencePiece 単語分割を用いる手法があることを紹介し、最近では単語分散を得ることのできる有力な手法として ELMo、 BERT が台頭してきたことを紹介し、そこでは SentencePiece が有力であるという実験結果が出ていることを示す。) 今回は単語分割+subword を用いることを想定し、1. fasttext の skipgram を用いて漢字かな入り混じり、かなのみのテキストに対して語彙数、損失、ある単語の類似語について実験をすること 2. 得られた単語分散を用いて極性判定の実験をすることを説明する。

4.4 実験) 漢字かな問題に対する単語分散取得

4.4.1 実験概要

単語分散を得るためのコーパスとして Wikipedia コーパスを用いたことなど、実験の概要を示す。

4.4.2 実験結果

実験結果を示す。

4.4.3 考察

考察を示す。

4.5 実験) 得られた単語分散を用いた極性判定

4.5.1 実験概要

4.4 で得た単語分散を用いて極性判定を行ったこと、極性判定のデータセットは 4.1 で抽出・編集したデータであることを示す。(抽出・編集条件 を再度示す) また実験に用いたネットワークについて説明する (CNN-RNN)

4.5.2 実験結果

実験結果を示す。

4.5.3 考察

考察を示す。

5 文抽出を念頭においた不均衡分散・サイズの分類問題

5.1 問題設定

入力された文が特定の意味を持った文であるかどうかを抽出する問題において、どのように分類するべきなのかを検討する。一般的なクラス分類との比較として、この問題は特定の意味を持った文の集合であるクラスと、それ以外のクラスとでデータの分散やデータの数に大きな差があること、画像認識と違ってアップサンプリング (水増し) が難しいことを問題点としてあげ、まず一般的に用いられている分類問題として解き、次に提案する手法である点類似度を用いたクラス分類を説明する。(特定の文で分岐を行い、その組み合わせを用いてユーザとの対話を試みる、シナリオ型対話システムがあることにも触れる。) 考察は比較のためにすべての実験のあとにまとめることを説明する。

5.2 実験) 自然言語処理の場合における一般的なクラス分類

news20 というデータセットを用いて CNN を用いた 1 クラス分類 (1 カテゴリ : 19 カテゴリ) を行う。相手のクラスの分散が想定よりも小さいことを注記する。

5.3 実験) 画像タスクに置換した場合における一般的なクラス分類

imagenet の画像タスクで、猫・犬分類と猫・ランダム画像でのクラス分類を行う。

5.4 実験) 自然言語処理の場合における点類似度を用いたクラス分類

BERT モデルを用いて、文類似度を測り、それを用いてクラス分類を行う。

5.5 実験) 画像タスクに置換した場合における点類似度を用いたクラス分類

画像の類似度を測り、それを用いてクラス分類を行う (実験が間に合えば)

5.6 考察

後者のほうが拡張性があること、前者の場合に猫・犬よりも猫・ランダムのほうが精度が悪くなる傾向があることを指摘する。

6 機械翻訳システムを用いた対話モデル

6.1 問題設定

反射応答のような問題について、機械翻訳を用いて発話を行わせることを提案、その手法として昨今機械翻訳の分野で SOTA を取っていた Transformer を用いることを実験し、その性能を考察する。

6.2 実験) Seq2Seq Attention と Transformer の精度比較

6.2.1 実験概要

データセットなどの実験概要を示す

6.2.2 実験結果

実験結果を示す。

6.2.3 考察

考察を示す。

7 文のスタイル変換

7.1 関連研究

この分野の関連研究として seqence to better sequence(本実験) や、(夏季レポートに記載したもの)　を例に挙げる。(画像認識の分野におけるスタイル変換についても触れておく必要があれば触れておく)

7.2 問題設定

書き言葉→話し言葉変換を行うことなどを説明する。またこの実験における話し言葉、書き言葉の定義についても言及しておく。

7.3 実験) 書き言葉→話し言葉のスタイル変換

7.3.1 実験概要

データセット、モデルの説明を行う。

7.3.2 実験結果

実験結果を示す。

7.3.3 考察

考察を示す。

8 CoLA タスクを応用した対話システムのエラー検知

8.1 問題設定

深層学習を用いた対話モデルや、文生成のモデルを用いる際に出てしまう可能性のある不自然な文を検出するという問題設定について説明する。

8.2 実験) 対話システムのエラー検知

8.2.1 実験概要

BERT を用いて実験したことを示す。 (このモデルを作成するにあたり文の自然さを評価するための CoLA タスクというものに注目し、これを解いている BERT と呼ばれるモデルを用いる。)

8.2.2 実験結果

実験結果を示す。

8.2.3 考察

考察を示す。

9 付録

この付録の存在意義について説明する。(論文の補足であることを説明する)

9.1 対話システムの関連研究

この章では 2 で引用した対話システムのうち、 Sounding Board と Gunrock について詳細な説明を行う。りんなに関しては非公開情報が多いため説明を省略する。

9.1.1 Sounding Board

Sounding Board Fang et al. 2018

9.1.2 Gunrock

Gunrock Chen et al. 2018

9.2 日本語データの取り扱いについて

9.2.1 単語分割

単語分割

9.2.2 Word Piece

Word Piece

9.2.3 Sentence Pieces

Sentence Pieces

9.2.4 Skipgram

Skipgram

9.2.5 CNN-RNN

CNN-RNN

9.3 質問文抽出を念頭においた不均衡分散・サイズの分類問題

9.3.1 画像データ

画像データ

9.3.2 文データ

文データ

9.4 機械翻訳システムを用いた対話

9.4.1 Seq2Seq Attention

Seq2Seq Attention

9.4.2 Transformer

Transformer

9.5 文のスタイル変換

9.5.1 Sequence to Better Sequence

Sequence to Better Sequence

9.5.2 CopyNet

CopyNet

9.5.3 Denoising Auto Encoder

Denoising Auto Encoder

9.6 CoLA タスクを応用した対話システムのエラー検知

9.6.1 BERT

BERT

10 結論

10.1 今後の課題

今回できなかった文生成の問題・論文に載せることのできなかった推論の内部状態の更新等について言及する。また精度向上や今後取り組みたい問題設定 (Unity などで仮想世界を作り、その中で対話を行えるようにするエージェント作成したい旨) について話す。

References

Chen, Chun-Yen et al. (2018). *Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data*.

Chu, Hang, Daiqing Li, and Sanja Fidler (2018). "A Face-to-Face Neural Conversation Model". In: eprint: arXiv:1812.01525.

Fang, Hao et al. (2018). *Sounding Board: A User-Centric and Content-Driven Social Chatbot*. eprint: arXiv:1804.10202.

Serban, Iulian Vlad et al. (2016). *A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues*. eprint: arXiv:1605.06069.

Sordoni, Alessandro et al. (2015). *A Hierarchical Recurrent Encoder-Decoder For Generative Context-Aware Query Suggestion*. eprint: arXiv:1507.02221.

Wo, Xianchao et al. (Mar. 2016). "りんな：女子高生人工知能". In: