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、一般にはおおよそ同程度のサンプル数が期待できる 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 $1 \cdots n-1$ 、それ以外のすべてををカテゴリ n として扱うことについて考える。そうすると、カテゴリ n のみ異様に得られるサンプル数、分散が大きくなってしまう。本テーマではこの問題について議論を行う。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に置き換え実験を行った。

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 2017 年 12 月時点で SOTA (State-Of-The-Art 最高水準) 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

[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異なる。

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

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

しまうというものがある。これに対処するため、特に何らかの機械学習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。

1.2 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。

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 [5] 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 [6] システムも構築している [Figure 1] 。

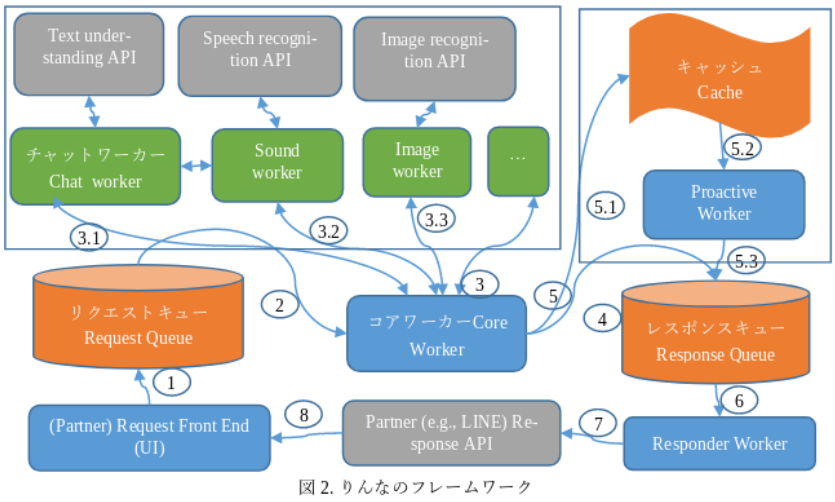

図 2. りんなのフレームワーク

Figure 1:　りんなのフレームワーク

[5] 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。

[6] 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。

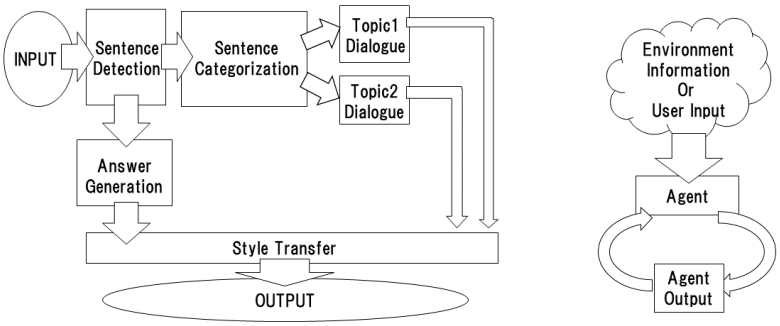

Figure 2: 本研究のシステム全体像

- 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・文抽出を念頭においた不均衡分散・サイズの分類問題]
  文を大雑把にカテゴリ分類する。例えばそれは livedoor news corpus [7] 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。

[7] 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[8]。

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や 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。

尚本研究では、形態素解析には MeCab 0.996、単語辞書として mecab-ipadic-neologd 20181112-01 を用いた。特に Twitter のようなデータは流行語や新語に対応するため、単語辞書を定期的に更新する必要がある。

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[9]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その一部を短くする前処理が施されることが一般的であること、を挙げる。

Table 1: 発話データに対して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

フィルタによって抽出された 54,368 発話を調査した。

まず発話データとして問題があると考えられる発話について報告する[10]。

[8]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。

[9]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。

[10]すべての報告における例は、個人情報を含んだ部分を含まないように編集されている。

Table 2: 発話データの調査結果 1

概要	詳細	例
他国語を用いた発話	中国語・英語等を用いた (含まれる) ツイートが 0.5 % 程度見られた	Very nice Merry Christmas! 謝謝 Guten Morgen!
テキストのみでは 理解できない発話	画像などのコンテンツに 対する発話が微量見られた ハイコンテクスト過ぎて 理解できないものが見られた	これ最高 れ!!!
(意図的・意図的でない) 誤字		オフトゥン イケメソ
顔文字や絵文字の多用	Twitter で許可されている絵文字や、 顔文字が含まれる発話が 8% 程見られた	(⋆´ω ｀⋆) お疲れ様です [(:3[■■]] (´∀ ｀)>
単語の一部や 語尾の繰り返し	特に感情的なつぶやきでは、 強調などの目的から 語の一部を繰り返す傾向が見られた	全全全休 ほにゃほにゃほにゃほにゃする やだぁあぁぁぁぁぁぁぁ!
略語の多用	長い単語、文は相互に理解できるような 形に省略されることが多かった	メリクリ! なるはや
別の表現	同じ意味を示すが 別の表記法があるものは 共通化されているわけではなかった	!/!!!/！/ !! /!!!!!!!! … /… こんど/今度 彼氏/カレ氏/カレシ デス/です
伏せ字	隠語など伏せ字を用いている場合があった	○ね
語尾の特徴付け等		ねれないぽよ ...と思うニョロ むいねー

次に主に情報の価値として問題があると考えられる発話について報告する。

Table 3: 発話データの調査結果 2

概要	詳細	例
個人情報の入ったもの	電話番号や SNS の ID などを 含まれるものが、 一万件に対して 5,6 件あった 個人名・アカウント名が含まれるものを 含めると 5%程になってしまった	
時刻など		2018.12.26 06:00
頻度が高すぎるもの	挨拶等	メリクリ！ おはよう
センシティブなもの		
Twitter 特有のもの		凍結された フォローありがとうございます
数値データ	英語での NLP の一部では積極的に削除されている 漢数字 ギリシャ数字	2018 200 円 一 V

最後にこの後実験として取り上げる極性判定のデータとして問題があると考えられる発話について報告する。

Table 4: 発話データの調査結果 3

概要	詳細	例
予定などのメモ書き	個人の予定や イベントの告知	
企業などの広告		
取引などのツイート		買) 鳥獣戯画のペンダント
豆知識や引用	特に深夜～早朝にかけては 自動ツイートのような形式の 豆知識や引用の頻度が高くなっていた 最大では 3 %程がこれに含まれていた	丁字染ちょうじぞめ オロバス ￥n ソロモン 72 柱の… [飲み会で使える？ダジャレ]… サーッ!(迫真)
感情が含まれているか 疑問のあるデータ		なぜ僕らは生きるのか

4.1.3 考察

データを収集した時間も相まって広告や豆知識・引用といった発話が多く観測された。これらのデータは極性判定やカテゴリ分類、ユーザクラスタリングなどに悪影響を与えることが論理的に考えられる。予定や広告、時刻などに関係したデータは、ほとんどの場合で一過性のものであるため長期的なシステムのためのデータとして見たときには適切であるか疑問が残る。

数値データや個人名のようなデータに関しては、英語での NLP、特に良い精度を持ったいくつかのタスクに対しては何らかの記号に置換されることが多い。しかし日本語でこれを適用しようとしたとき、1. 様々な表記方法があること、2. スペースで分割されていないため、形態素解析などの技術や NER(Named Entity Recognition 固有表現抽出) の技術を組み合わせなければ抽出できないこと、が問題として挙げられる。特に形態素解析に関しては Twitter のデータのような正規化されていないテキストに行った場合、精度が比較的に落ちるため、何らかの精度向上手法または別手法を提案する必要がある。

また同じ意味を表す文でも様々なバリエーションがあることがわかった。例えば “おはよう” を例に取ってみると、“おはようございます”、“おはよー”、“おは”、“おはよおおお”、“おは (愛称等)” といったバリエーションが見られた。これらはキャラクタ性を持たせるためには必要な分散であるが、意味のみに注目した場合や、語彙数の問題を考慮した場合には極力減らされたほうが良いと考えられる[11]。更にバリエーションのある文は平均的に出現頻度が高い[12]ため、これを集めすぎるとデータに偏りが生まれてしまうことも考慮する必要があるだろう。

極性判定のみに絞った議論をするならば、例えば自動ツイートされた発話にはユーザの極性があるとは考えにくいため、これを省くのが適当であると考えられる。しかし以上のことを踏まえてデータの再抽出・編集をフィルタリング後のデータの中の、15,000 程度のデータに対して行ったところ、1,500 程度のデータしか得られなかった。尚特にこの結果を招いた要因を挙げるとすれば、個人情報を含んだデータを編集・削除したこと、極性を持たないと思われるデータ (中性という意味ではない) を省いたことだ。

更に極性判定のためのデータとしてこのデータを考えると、顔文字や絵文字等は極めて感情を含んでいると感じられた。例えば、“おはようございます。(ノД｀)” と “おはようございます。(*´ω｀*)” では極性判定上全く違う評価を下さざるを得ない。しかし顔文字や、特に絵文字については、そのバリエーションに際限がないことや機種依存文字などの入力可能性について議論しなければならない。これらを解消するためには、それらを例えば文字単位、或いはそれに準ずる単位で分割するなどしてある程度のカテゴライズを行えるようにする手法が要求される。

[11]英語の NLP (例えば機械翻訳) でも前処理として、“he’s” を “he is” にするなどの前処理が行われることがある。
[12]例えば 26 日午前 6 時ちょうど頃は 3 割程度が宛先や顔文字などの付加情報の差はあれど “おはよう” の意味の発話であった。

形態素解析で成功した例

りかちゃんありがとう

<形態素解析結果>
りか 名詞, 固有名詞, 人名, 名,*,*, りか, リカ, リカ
ちゃん 名詞, 接尾, 人名,*,*,*, ちゃん, チャン, チャン
ありがとう 感動詞,*,*,*,*,*, ありがとう, アリガトウ, アリガトー

形態素解析で失敗した例

山さんに …

<形態素解析結果>
山 名詞, 一般,*,*,*,*, 山, ヤマ, ヤマ
さん 名詞, 接尾, 人名,*,*,*, さん, サン, サン
に 助詞, 格助詞, 一般,*,*,*, に, ニ, ニ
…

※人名を指すが一般名詞として認識されてしまっている。
このよう場合には単語分割した後、NER を用いて検出することが望ましいと言える。

4.2 調査) 対話データ

対話データとして、2018 年 8 月から 12 月にかけて不定期に Twitter から収集した対話データ、一般公開されている書き起こしの対話コーパス、一般公開されているチャットの対話コーパスについてデータを観測した。

以下に調査結果として何らかの問題があると考えられる特徴について報告し、それに対する考察を述べる。

4.2.1 調査結果

1. Twitter から収集した対話データ

   収集方法は 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の中から、3 発話以上対話が続いているものを収集した。この処理によって生データが 10,767 の対話ペアが得られた。そして生データに対しては 4.1.1 と同様にハッシュタグと宛名、そして URL リンクを削除したが、文字制限は対話間の意味を観測するため行わなかった。

Table 5: 対話データの調査結果 1

概要	詳細	例
センシティブな内容	3 %程はセンシティブな内容の対話であった。	
ゲームに関する内容	5 %程はゲームに関する内容であった。 その中には一過性の内容 (情報共有や待ち合わせ等) が含まれていた	
顔文字や絵文字等が含まれるもの	15 %程は顔文字や絵文字を含んでいた そのうちの 2 割ほどは顔文字・絵文字のみが 発話になっているものがあった	おはよーございます! ((⋆゜Д゜)ノ ('Д`)
似たような内容	特に挨拶など同じような 内容の対話頻度が高かった 朝方には半数が “おはようございます” の内容であった	おはようございますよ
事前知識を必要とする内容	間柄や話題 (例えばゲーム) の内容に 関する事前知識がいるものが 多く感じられた。[13]	line カメラたのしい

次ページに続く

前ページからの続き

概要	詳細	例
固有表現が含まれるもの	名前等固有表現が含まれるものは 3 割程度であった。	

2. 名大会話コーパスから収集したデータ

名大会話コーパス (逸子, 美恵子, and ディヴィッド義和 2011) から入手できる 129 会話について観測した。名大会話コーパスとは日本語母語話者同士の雑談を文字化したコーパスで、129 会話を収録、その合計時間は 100 時間に及ぶ比較的大規模なものだ。ライセンスがクリエイティブ・コモンズ表示-非営利-改変禁止 4.0 国際ライセンスで公開されているため、研究目的で用いることは非常に容易なコーパスであると言える。

非常に大規模かつ考察で述べるように複雑な内容であるため、出現頻度については言及しない。

Table 6: 対話データの調査結果 2

概要	詳細	例
言外のコミュニケーション	言語化せずに伝える内容があった	＜笑い＞ (共感の意)
長文や複文	相手が内容を理解したものとして 文を継続させる場合があった。	すごい勢いで走って。 私、あ、あーさっきの犬だとか 私たちが言っとるじゃん。 犬も気がついたじゃん。 じゃははって走ってきちゃって、犬が。
書き言葉・話し言葉の変化	あの → あん といった変化が見られた。	ほいでさあ、ずっと歩いていたんだけど、 そうすと上から、なんか町の中が見れるじゃん。
固有表現	個人情報保護のため 名前などの 固有表現は置換されていた	＊＊＊の町というのはちいちゃくって ... ほいで、あの、F023 さんはあたしが前の日に... Cが、あのー、写真を見せてくれたんだけど...

3. 対話破綻チャレンジから収集したデータ

対話破綻チャレンジ (TODO: conv_challenge) とは人間と対話システムとの間で生じる「対話破綻」(ユーザが対話を継続できなくなる状態) を自動検出することを目的とした、評価型ワークショップである。

このデータは対話システムと人間間とのテキストを用いた対話データと、その対話が成立しているかどうかを判定した複数人によるアノテーションが含まれており、本研究の目指すエージェントと人の対話の形に最も近いデータセットであると言える。

本データセットは問題点が少なく、アノテーションに従って、比較的成立しているとみなされた対話を抽出することで対話データを生成することが出来た。

4.2.2 考察

Twitter から収集した対話データに関しては Twitter データとして非常に有効であると考えられる。しかし比較的に個人的・センシティブな内容が多く、これを対話データとして学習させてしまうことによる、対話システムの倫理的な問題を考慮しなければならないだろう。また顔文字や絵文字等は 4.1 で考察したように単位で分割することが難しい。同様に同じような意味を持った対話が多く存在していたことから、これにも対処する必要があるだろう。

名大会話コーパスから収集したデータに関しては日常会話を分析・理解するには抽出するには非常に価値のあるデータセットであるが、これをチャットのようなテキスト入力等を介した対話には不適切なデータであると考えられる。このコーパスを観測して考察できる内容としては、1. 書き言葉・話し言葉の変化は想像以上に大きなものであったと言えること、2. 決して発話一つに対して返答が一つという形式になっているわけではないこと、3. 固有表現の取扱についてより深く考察する必要があること、であった。

対話破綻チャレンジから収集したデータはほぼ申し分ない自然さを持ったデータを集めることができることがわかった。しかし対話システムと人との対話データであるため、“人対人のような日常会話” 対話は

[13]アノテータが一人のため境界を判定することは難しいため、割合を明言することは出来ない。

比較的少なく、“人のような” 対話エージェントを作成するならば、不足している対話を外部から付け加える必要があると考えられる。

4.3 問題設定

NLP の研究分野の一つについて単語分散を用いた言語モデル生成がある。単語といったある単位ごとの意味をベクトルなどの数値にする手法であり、この利点としては、単位ごとの距離を考えたとき、意味的に近い要素は近く、遠い要素は遠くなることで様々な NLP のタスクで自然言語を数値化する際に、自然言語の特徴を強く表すことができるようになるというものがある。

本研究ではこの単語分散を得るという問題に対してデータの前処理がどのように影響するのかを理解する目的で、2 つの実験を行う。

一つは、1. 漢字・かな入り混じり文、2. かな飲みに変換した文、によって得られる単語分散の性質の違いを確認する実験、もう一つは得られた単語分散を用いて極性判定を行う実験である。

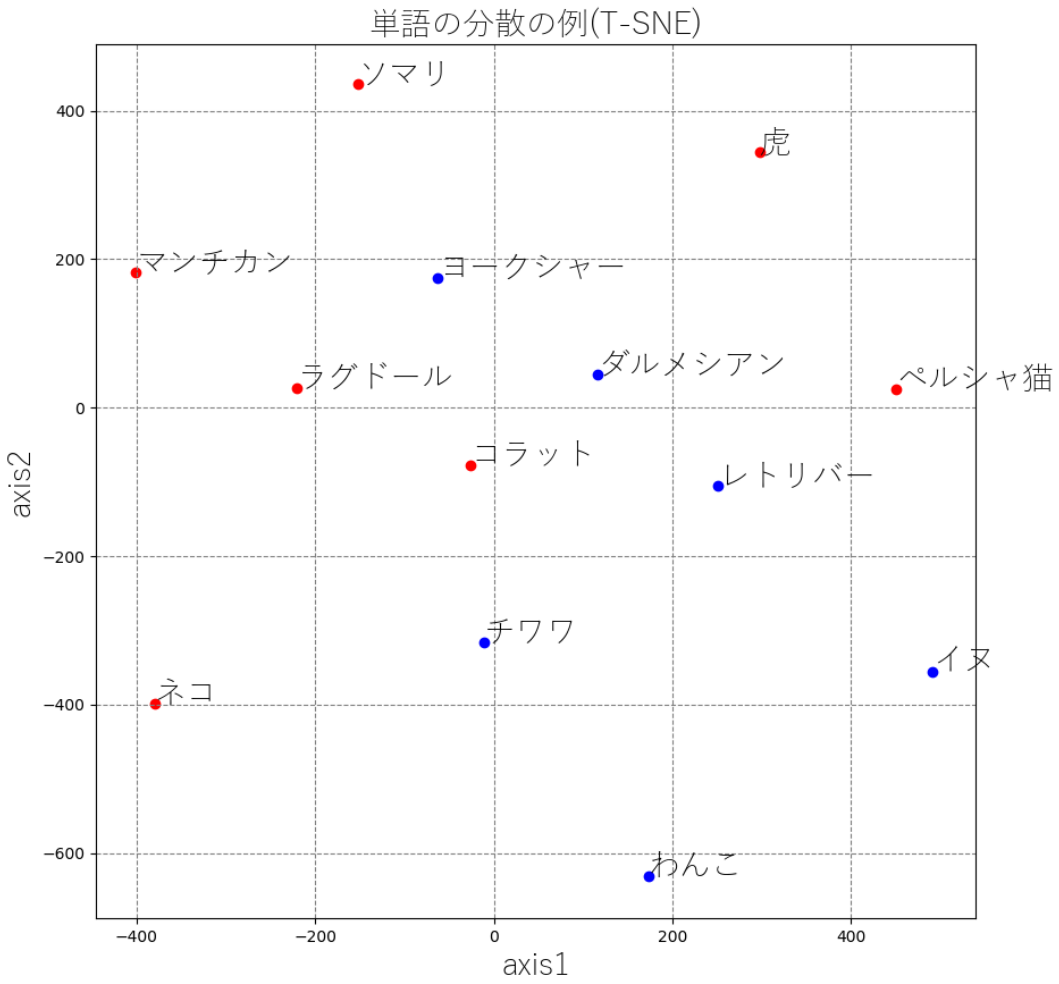

Figure 3: 単語分散の例 (T-SNE を用いて二次元平面に描画)

4.3.1 関連研究

単語分散を得るための手法としては、SVD(特異値分解)(TODO:svd) や Word2Vec(TODO:word2vec) や glove(TODO:glove)、fasttext(TODO::fasttext) といった手法が有名だ。また昨今、NLP では文単位での解析が多いこと、文全体の意味も考慮したほうが良いというモチベーションから、単語分散のみならず、文ごとの関係も考慮してベクトルを生成する手法が提案されている。その代表例が、ELMo(TODO:ELMo)、BERT(TODO:BERT) と言った深層学習のモデルであり、昨今の様々な NLP のタスクで SOTA を達成している。

4.4 実験) 漢字かな問題に対する単語分散獲得

この実験では、日本語特有に存在する “漢字とかなによる同意表現の複数表記” を解消するための漢字 → かな変換を行い、それによって得られる性質の変化を調査する。

上記の調査で明らかになったように、日本語には同意でありながら様々な表現が存在している。その中でも比較的簡単に差がわかる・前処理が簡単であるものとして、“漢字とかな” について挙げることができる。例えば “寒い” という単語は “さむい”、“寒い” といった場合があるが、これらは単語的にはほとんど同じ意味を示す。また漢字とかなが入り交じることによって文字の種類が増加し、英語に比べて解析時の次元数が増大してしまう可能性が直ちにわかる。更に日本語のみならず英語を代表とした他国語をそれらの文字のまま併用し、それを当然のように会話に組み込んでいるという特徴から、日本語の文字種数を削減することは重要であると考えられる。しかしこの前処理を行う弊害として、例えば “すなわち”、“即ち”、“則ち”、“乃ち” といった微妙にニュアンスの異なる同音の単語がまとめられてしまうことによる影響をについて憂慮する必要があり、考察しなければならない。

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 から入手したコーパスを用いた。Wikipedia コーパスを選択した理由として、プライバシーや料金といったデータの入手難易度が低いこと、言語モデルを作成することを視野にいれた際に、百科事典的な特徴から大まかに日本語の語彙を網羅することが期待でき魅力的であることを挙げられる。

実験に用いるモデルは、fasttext の subword を用いた、Skip-gram(TODO:Distributed Representations of Words and phrases and their compositionality) である (TODO:enriching word vectors with subword information)。subword とは活用や語幹といった単位で単語を分割することで、例え単語が文字上一致しなくともその単語間の距離が近くなることを保証できるという利点が得られる。これは特に英語が、単語が小さな意味を持つ文字群に分割できることに大きく影響する。この利点は日本語にも応用可能であるという理屈としては、任意の国語辞典を開けばわかることだ。

Skip-gram はターゲットとなる単語からその周囲単語を予測する単語分散の獲得手法である。Skip-gram の詳細は 9.2.6 で説明する。

subword の例

・英語の場合
inspire → in・spire (中に+吹き込む)

・日本語の場合
鶏肉 → 鶏・肉 (鶏 (の)+肉)

議論の対象は以下の 3 点についてだ。

- 語彙数の変化
  漢字 → かな変換によりどれだけ語彙を縮小させることが出来たのかを調査する。ここでいう語彙数とは subword ではなく形態素解析で得られる本来の単語の数である。
- それぞれの、単語埋め込みベクトルの次元数と損失の変化
  それぞれの場合で、単語埋め込みベクトルの次元数に対して、訓練後の損失がどの程度変化するのかを調査する。
- それぞれで得られた最良のモデルに対する、類似語の変化
  それぞれの場合で、“日本 (ニホン)” という単語に対してどのような類似単語が得られるのかを調査する。

実験上の固定されたパラメータを以下に示す。パラメータの詳細な意味は 9.2.6 で説明する。

Table 7: fasttext を用いた単語分散獲得学習の共通パラメータ

パラメータ名	値
許容最低語彙頻度	5
学習係数	0.1
学習係数向上率	100
epoch 数	5
ネガティブサンプル数	5
ウィンドウサイズ	5

4.4.2 実験結果

実験結果を示す。

ここでいう次元数とは単語埋め込みベクトルの次元数 dim であり、default とは漢字かな入り混じり文、yomi とは漢字 → かな変換を行ったものを示す。

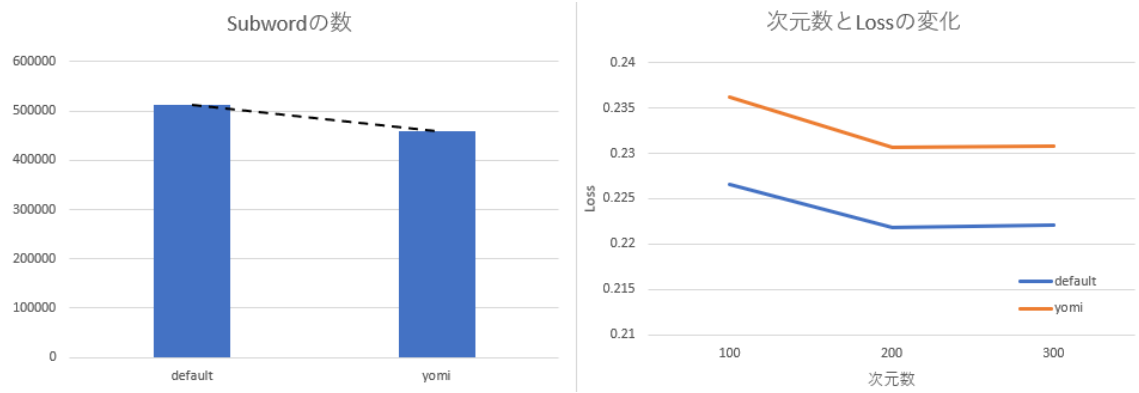

Figure 4: 漢字かな問題に対する単語分散獲得

1. 漢字かな入り混じり文 の類似単語

   用いた単語埋め込みの次元数は 200 である。

Table 8: 漢字かな入り混じり文 の類似単語

ターゲット	日本
類似単語	韓国
	米国
	台湾
	にっぽん
	中国
	日本さくらの会
	海外
	実業
	国内
	日本税理士会連合会

2. かなのみの文 の類似単語

   用いた単語埋め込みの次元数は 200 である。

Table 9: かなのみの文 の類似単語

ターゲット	ニホン
類似単語	ニホンヤモリ
	ニホンバレ
	ニホンシカ
	ニホンウンソウ
	ニッポンザル
	ニホンズイセン
	ヒトツオボエ
	ゴジセイ
	ニホンカジョシュッパン
	ニホンドケン

4.4.3 考察

漢字 → かな変換によって語彙が 10%程度減少したことは確認できたが、損失は増加してしまったことがわかる。しかしいずれの場合でも次元数と損失の変化の外形は似ていることから Skip-gram の損失のみを見るならば変換前のテキストの方が良い単語埋め込みを獲得できていると考えられる。

また類似単語であるが、漢字かな入り混じり文は国として類似する単語を取り出していることがわかるのに対して、かなのみの文では 生物名 や、“日本晴れ” といった 慣用的な表現 を多く抽出している。このことから変換を行ったほうが、subword を活かすことが出来ていると考えられる。

これらのいずれが良いのかについては議論の余地があるだろうが、少なくとも汎用的な言語モデルを作成するならば前者の Skip-gram としての損失が小さい方を選択する方が良いと考えられる。

4.5 実験) 得られた単語分散を用いた極性判定

この実験では、4.4 で得られた単語分散を用いて極性判定を行うことで 2 つの単語分散の極性判定における性能を調査する。

一般に単語分散を獲得することで得られる言語モデルは極性判定やカテゴリ分類等に活用されることが多いが、今回は特に極性判定のうちの、陽性・中性・陰性の 3 値分類について挑戦する。3 値分類を選んだ理由は、データとして Twitter のデータを収集した際に、4.1 にあるように必ずしも陽性・陰性の 2 値を取らなかったこと、5 値のようなより複雑な分類にすると、データのラベリングコストが高くなってしまうことを挙げる。

4.5.1 実験概要

用いた単語分散は 4.4 で得られた中で損失が最小であった 200 次元のものを用いた。極性判定のデータセットは 4.1 で抽出・編集したデータだ。抽出条件として、4.1.2 で得られた結果を用い、今回はこのいずれかに該当するものすべてを削除・編集した。

データ数は総データ数 1270 発話、この内ランダムに抽出した 10 %を学習に用いない検証データとした。

用いたモデルは CNN(Convolutional neural network) と 双方向 LSTM(Bidirectional long short term memory) を合わせたものであり、構成を以下に示す。

構成しているレイヤーの説明は 9.2.7 で行う。

Table 10: 実験に用いた CNN の概要

パラメータ (レイヤー)	値	補足
1 層	1 次元畳み込み	フィルターサイズ 64 / カーネルサイズ 3 / 活性化関数 elu
2 層	1 次元畳み込み	フィルターサイズ 64 / カーネルサイズ 3 / 活性化関数 elu
3 層	1 次元畳み込み	フィルターサイズ 64 / カーネルサイズ 3 / 活性化関数 relu
4 層	最大プーリング	プーリング幅 3
5 層	双方向 LSTM	隠れ層サイズ 256 / ドロップアウト率 0.2 / 再帰中のドロップアウト率 0.3
6 層	全結合層	ユニット数 256 / 活性化関数 sigmoid
7 層	ドロップアウト層	ドロップアウト数 0.25

次ページに続く

前ページからの続き

パラメータ (レイヤー)	値	補足
8 層	全結合層	ユニット数 256 / 活性化関数 sigmoid
9 層	ドロップアウト層	ドロップアウト数 0.25
10 層	全結合層	ユニット数 256 / 活性化関数 sigmoid
1 層	ドロップアウト層	ドロップアウト数 0.25
12 層	全結合層	ユニット数 3 / 活性化関数 softmax
epoch	十分に学習できるまで	過学習が起きる直前の値を訓練後の精度とした
最適化関数	Adam	適当に調整した
損失関数	クロスエントロピー	

4.5.2 実験結果

以下のようにいずれの場合でも accuracy という面では若干の精度向上が見られた。しかし検証データの損失に関しては増大してしまっている。

Table 11: 得られた単語分散を用いた極性判定

	漢字かな	かなのみ
訓練データの損失	0.9523	0.7016
訓練データの accuracy	95.2%	98.2%
検証データの損失	1.204	2.096
検証データの accuracy	61.5%	64.8%

4.5.3 考察

ひらがなにすることでやや精度が向上したようにも見えるが、複数回実験をしたものの大きな違いが得られるような結果とはならなかった。この原因として、Wikipedia コーパスと収集したデータの距離が離れていることを考えることが出来る。

本実験でははは以下の表に示すように学習した語彙以外の単語が、いずれの場合でも 30%ほど、学習データに含まれてしまった。これは subword を用いての結果であるため、単語区切りやそれ以上の区切りのもので単語分散を学習した場合には、より語彙外の単語が増えてしまうことが想定できる。これに対処する方法として、学習に用いるデータも合わせて fasttext で単語分散を得ることが提案できるが、Wikipedia コーパスに比べ学習データは極端に少ないため、2 つのデータを合わせてもそれらは語彙外の単語として切り捨てられてしまった。

以上のことから、前処理もさることながらより目的にあった密な (語彙数の増加よりもデータ数の増加が大きくなるような) データを効率よく大量に収集する必要があると考えられる。

また検証データに対する精度が向上しながらも損失が不安定になってしまうという問題が多く発生した。これは損失がクロスエントロピーを用いていることで、以下のような現象が起きていると考えられる。

クロスエントロピーを用いて損失が増大しまうシナリオ

真のラベルを $[1.0, 0.0]$ とする。出力をそれぞれ $[0.8, 0.2]$ 、 $[0.6, 0.4]$ とする
勿論いずれの場合においても正しく識別できている。
しかし真の分布 p(x) と 推定された分布 q(x) を用いてクロスエントロピーは以下のように定義されるものであるから、前者 (0.223) よりも後者 (0.510) の方が損失の値が大きくなってしまう。

$$cross\_entropy = -\Sigma_x p(x) \log q(x)$$

Table 12: 学習データ中の語彙外の単語数

	漢字かな	かなのみ
全語彙数	19265	20975
語彙外の単語数	6512	6453
割合	33.8%	30.1%

Table 13: 得られた単語分散を用いた極性判定 (Wikipedia + 学習データ)

	漢字かな	かなのみ
訓練データの損失	0.1161	0.1010
訓練データの accuracy	96.8%	96.9%
検証データの損失	1.7960	1.8166
検証データの accuracy	64.0%	64.7%

5 文抽出を念頭においた不均衡分散・サイズの分類問題

任意の文の入力を受け付ける際に、いくつかのある特定の内容の文が入力された場合のみ、何らかのイベントを発したいという状況について考える。このとき “任意の文” と “ある特定の内容” という領域の比を考えるといくつかのパターンが考えられる。例えば、“任意の文” が極性判定のようなネガティブ・ポジティブな文の集合であり、“ある特定の内容” がポジティブな文であったとき、これはネガティブな文とポジティブな文を区別するシンプルな 2 クラス分類問題と考えることができる。ここで用いる、シンプルな、という意味は、おおよそ 2 つのデータの自然言語空間上の分散、領域の大きさが一致していると考えられ、おおよそ同じくらいのデータサイズのサンプルを確保できるということだ。ところが、“任意の文” が例えば病院の診察記録であり、“ある特定の内容” が 1,000 万人に一人の発症率の難病、しかもそれを複数取り扱いたいと考えたとき、この問題は極めて難しいものとなる。これは n クラス分類問題でありながら、1 つのクラスが異様に全体データの領域を占め、そして残りの $n-1$ クラスが得られるデータのサンプル数が極端に少ない。こうなると通常のクラス分類ではうまく行くとは考えにくい。

本テーマでは、うまく行かないということを確かめるため、まずデータが充実している画像処理についてこの問題を考え、次にデータが不揃いであるものの自然言語処理でも同様にすることでクラス分類がうまく行かないことを確かめ、提案する手法である、点類似度を用いたクラス分類を実験し、その効果を確認する。

5.1 問題設定

3 つの問題設定で実験を行う。

一つは ImageNet という 2 万種類以上のラベルを持つ画像認識のデータセットを用いた 2 クラス分類で、猫の画像と犬の画像を分類する場合と、猫の画像ととランダムな画像を分類する場合、そしてそれぞれでデータ数に偏りをもたせた場合の精度比較する。

もう一つは news20 という英語の 20 種類のラベルを持つ英語の自然言語処理のデータセットを用いた 2 クラス分類で、やや問題が元の問題設定よりもずれているものの、自然言語処理の領域で問題を解く必要があると考えたため、これを用いて精度を比較する。

最後に提案する点類似度を用いたクラス分類を行う。この提案手法は、与えられた文と判定したいクラスのテキストのサンプルデータすべてに対するの類似度をとり、その値群を考えることでその文がクラスに含まれているかを考えようというもので、値群を合計するのか、最大値を取るのかという 2 つの指標の下実験する。

5.2 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 のデータを用いた画像タスクで、猫・犬分類と猫・ランダム画像でのクラス分類を行い、その精度の変化を実験する。

5.2.1 実験概要

ImageNet (Deng et al. 2009) とは 2 万件のラベルを持つ画像を合計で 1,500 万枚有しているデータベースである。

つまりここから得られる画像データセットを利用すれば、19,999:1 の比率のクラス分類を実験することができる。また深層学習の分野では積極的に画像認識で使われている技術が自然言語処理でも使われている [14] ことから、こちらで精度が出ていればそれを自然言語処理に転用することも容易であることが伺える。

[14] 例えば最近では RNN(reccurent neural network) で文章のベクトルを生成していたものと、画像認識分野で広く使われている CNN(convolutional network) を用いて同様のことを行う研究 (TODO:Pervasive Attention:2D Convolutional Networks for Sequence-to-Sequence Prediction) が流行している。

以上のことからこれは元問題の設定にそれなりに近い設定であると言えるだろう。
その上でデータの分散が異なると見られる犬とランダムな画像を相手として、猫の画像と分類する 2 クラス分類問題を実験する。
尚今回は比較のため、用いるモデルは統一している。そのモデルは AlexNet(TODO:AlexNet) を参考にした CNN (Convolutional Neural Network) であり、概要は以下の通りであり、詳細は 9.3.1 で述べる。
データは 28×28 の 3 チャンネル (rgb) の画像、データ数は猫・犬 (ランダム画像) で、その比率は 200:1000 / 400:800 / 600:600 / 800:400 である。検証データについてはいずれの場合でも 30:30 に統一した。

Table 14: 実験に用いた CNN の概要

パラメータ (レイヤー)	値	補足
1 層	2 次元畳み込み	フィルターサイズ 32 / カーネルサイズ 3×3 / 活性化関数 relu
2 層	2 次元畳み込み	フィルターサイズ 64 / カーネルサイズ 3×3 / 活性化関数 relu
3 層	最大プーリング	プーリング幅 2×2 / プーリング間のストライド 2
4 層	ドロップアウト層	ドロップアウト率 0.25
5 層	2 次元畳み込み	フィルターサイズ 128 / カーネルサイズ 2×2 / 活性化関数 relu
6 層	最大プーリング	プーリング幅 2×2 / プーリング間のストライド 2
7 層	2 次元畳み込み	フィルターサイズ 128 / カーネルサイズ 2×2 / 活性化関数 relu
8 層	最大プーリング	プーリング幅 2×2 / プーリング間のストライド 2
9 層	ドロップアウト層	ドロップアウト率 0.25
10 層	全結合層	ユニット数 1500 / 活性化関数 relu
11 層	ドロップアウト層	ドロップアウト数 0.5
12 層	全結合層	ユニット数 2 / 活性化関数 softmax
epoch	十分に学習できるまで	過学習が起きる直前の値を訓練後の精度とした
最適化関数	Adam	
損失関数	クロスエントロピーに重みを付けたもの	重みはデータ数 x:y に対して y:x の比率

5.2.2 実験結果

図中の Train_acc は訓練データに対する accuracy、Val_acc は検証データに対する accuracy、Train_loss は訓練データに対する損失、Val_loss は検証データに対する loss だ。尚 accuracy が 0、或いは損失が 1 となっているのは学習率などを変更しても収束しなかったことを示している。

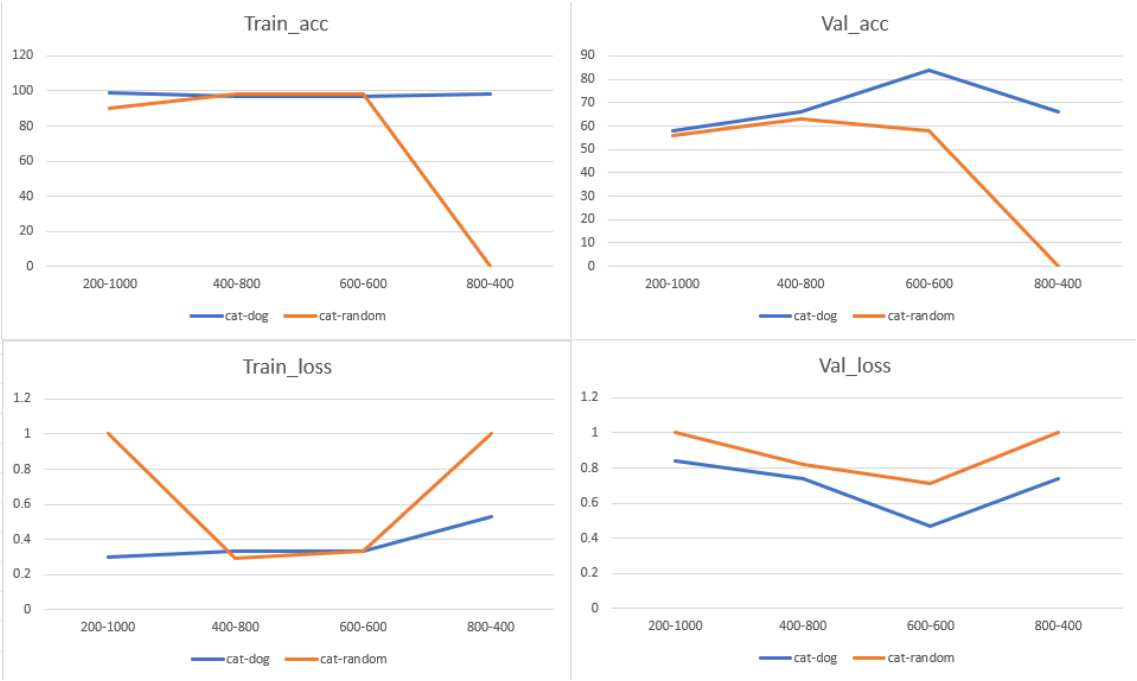

Figure 5: 画像タスクに置換した場合における一般的なクラス分類

5.2.3 考察

全体的にランダム画像とのクラス分類の方が精度が悪いとわかる。このことから、通常のクラス分類を転用してクラス分類を行うよりはそれにふさわしいモデルを作成した方が良いとわかる。

またランダム画像とのクラス分類に関しては、ランダム画像が多いほうが検証データに対する accuracy が向上するという予想があったが、ほとんど向上しないことがわかった。しかし犬画像との検証データに対する accuracy を比較すると、犬画像がデータ数が等しい場合を頂点として対称に精度が落ちているのに対して、ランダム画像に関しては 400:800 の時が最も精度が高くなっていることが興味深い。しかしいずれの場合でもデータの偏りが生じると損失は増加してしまう傾向にあるため、これが健全な学習結果であるすることは難しいだろう。

またより損失の重み付けを大きくした場合についても実験を行ったが、この場合には学習が荒れてしまい結果を得ることが出来なかった。

5.3 TODO: 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。

5.4 TODO: 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 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

反射応答を行うシステムを作成するという問題について、機械翻訳の手法を用いることを提案、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。

3 においてはカテゴリごとに別のモデルを作成することを提案しているが、本実験では十分なデータを入手できなかったため、利用可能なデータを集めたもので実験を行った。

6.2 実験) Seq2Seq Attention と Transformer の精度比較

6.2.1 実験概要

応答の中でも前後の文脈がなくともある程度意味が通じる反射応答に近い部分に適用されることを想定し、“入力文対出力文が一対一” であるようなシステムを考える。このときこの問題は、一般の機械翻訳の形に落ち着けることが出来る。この仮定のもとで、古くから有力な機械翻訳手法であるとされる Sequence to Sequence モデル (TODO: seq2seq) [15] に補正をかけた Sequence to Sequence Attention (TODO: seq2seq_att) と、最近提案され 8 で取り上げる BERT という言語モデルの主軸となるアイデアにもなっている Transformer (TODO:att_need) を用い、それらから得られる出力の性質について考察する。

データとして用いたものは、自作の 1k 程度の対話と 4.2 において対話破綻チャレンジ (TODO: conv_challenge) から得ることが出来た 15k 程度の対話を組み合わせたものだ。自作のデータを用いた理由としては、対話破綻チャレンジで得られたデータセットを確認したところ、期待する反射応答である、挨拶といった対話が不足していると考えられたことを挙げる。

6.2.2 実験結果

以下に Sequence to Sequence Attention によって得られた出力と、Transformer から得られた出力、そしてそれぞれの BLEU スコア (Bilingual Evaluation Understudy Score) を示す。BLEU スコアについては 9.4.3 にて紹介する。

Table 15: データセットの入力と出力の例

入力	出力
チーズ ケーキ は 好き です ね 。	チーズ ケーキ の 美味しい 店 を 知っ て い ます 。
お は よー 。 熱中 症 に 気 を つけ て 。	おはよう ござい ます 。
テレビ は 民主党 支持 が 多い です ね 。	そんな 事 は ない でしょ ?
プレイ は 普通 です 。	どんな スポーツ を さ れ ます か ?
晴れ の 日 は 嫌い ? 暑い ?	晴れ の 日 が いい です ね 。
です よ ね 。	はい 。
すごい . 子供 とか 泣く もん ね 。	打ち上げ花火 だっ たら 連発 で ガンガン 上がる の が 楽しい 。
お は よー ありがとう 。	最近 映画 見 まし た ?
晴れ です か 。	気分 が いい です ね 。
ポテトチップス は うす 塩 が 好き です ね 。	うす 塩 は 定番 です ね 。
こんにちは ー っ 。	元気 が いい です ね 。

Table 16: Sequence to Sequence Attention と Transformer によって得られた出力

Seq2Seq Attn	Transformer
チーズ ケーキ の 美味しい 店 を 知っ て い ます 。	チーズ ケーキ の 美味しい 店 を 知っ て い ます 。
漫画 は 読み ます か 。	おはよう ござい ます 。
そんな 事 は ない でしょ ?	そんな 事 は ない でしょ ?

次ページに続く

[15] 一対一な機械翻訳手法を一般に “Sequence to Sequence (Seq2Seq) モデル” と呼ぶこともあるが、今回指すそれはそのうちの手法の一つであるモデル (TODO:Seq2Seq) を指す。

前ページからの続き

Seq2Seq Attn	Transformer
どんな スポーツ を さ れ ます か？	どんな スポーツ を さ れ ます か？
晴れ の 日 が いい です ね 。	晴れ の 日 が いい です ね 。
夏 って 感じ が し ます 。	私 も スポーツ が 好き です 。
なるほど 。	気温 は いい です ね 。
最近 映画 見 まし た？	最近 映画 見 まし た？
気分 が いい です ね 。	気分 が いい です ね 。
うす 塩 は 定番 です ね 。	うす 塩 は 定番 です ね 。
元気 が いい です ね 。	元気 が いい です ね 。

Table 17: 学習に用いたデータの BLEU スコア

	BLEU スコア
Seq2Seq Attn	66.92
Transformer	77.11

Table 18: 学習外のデータについての BLEU スコア

	BLEU スコア
Seq2Seq Attn	61.80
Transformer	64.33

6.2.3 考察

それぞれのモデルからの出力文そのものを眺めると、いずれも文法的に不自然でないテキストを出力していることがわかる。しかしおおよそ短文としては成立している一方で文脈の考慮という点では今ひとつという出力が見られることがわかる。

しかし教師データである入力と出力がそもそも文脈上でのみ成り立っているものも含まれていることがわかるため、この点を考慮すればおおよそ期待通りの学習が出来たと考えている。

また本テーマでは翻訳とは違って単語対単語の直接的なつながりが比較的薄く、RNN のような文全体を読む機能が必要であるように思われたが、Transformer に組み込まれている単語間の関係を示す Positional Encoding が効いているおかげか単語対単語の対応ではない学習が出来ていると考えられる。(実際に Positional Encoding を削除した場合で実験を行った際には Sequence to Sequence Attention よりも精度が悪くなってしまった。)

訓練時間については Sequence to Sequence Attention よりも Transformer のほうが圧倒的に早かった。これはまず Transformer が RNN を用いていないという影響が大きいと考えられる。しかしそれであったとしても、epoch 数が前者は 700 程度必要であったのに対して、後者は 60 程度で収まっているという点が興味深く感じられた。こちらの理由に関しては、Attention 機構と RNN の機構を組み合わせることでモデルが比較的に大きくなってしまったというということが考えられるが、確証のある説を提示することは出来なかった。尚具体的には以下の自宅環境で実験を行った際に、前者は 60 分、後者は 25 分ほどで学習することが出来た。[16]

Table 19: 実験環境

OS	Windows 10 Education
CPU	Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz
RAM	16.0GB
GPU	NVIDIA GeForce GTX 1080
Python	3.6.5

[16]自宅環境で実験を行った理由は、研究室の計算資源よりもこちらのほうが計算速度が早かったためだ。

7　文のスタイル変換

日本語は英語と比較してペルソナに伴う語尾などの言葉遣いの変化が顕著である。これは日本語を対象とした統計・機械学習を行う際に、英語で用いられる手法を直ちに用いることができるか、という点で議論が生じる。その意味で日本語の文に対して何らかのスタイルを付与するという手法について既存の英語で用いられている手法と、2 つの提案手法を用いて実験する。

7.1　関連研究

スタイルを変換するという問題に対して、画像認識では VAE(Variational autoencoder)(TODO:VAE) や GAN (Generative Adversarial Network)(TODO:GAN) が提案されており、 いずれも様々な派生が研究されている。NLP の舞台でも同様の試みが行われているが、現在文を変換するというタスクに対しては特に、VAE を用いた研究が盛んである。例としては、Toward Controlled Generation of Text(TODO:tcg-text) や、Sequence to Better Sequence(TODO:seq2bseq) や Style Transfer from Non-Parallel Text by Cross-Alignment (TODO:cross-align) があり、これらは非並行、つまり必ずしも元の文とスタイルが付与された文が対になっている必要がないという点で優れている。

しかしこれらが議論しているスタイルとは日本語で用いられるようなペルソナを象るようなものではなく、むしろ極性や単語並び替えといった議論に集中している。唯一 Sequence to Better Sequence に関してはシェークスピアの作品と現在の言葉との変換を行っているため、本実験ではこれを用いて実験を行う。

7.2　問題設定

日本語での書き言葉 → 話し言葉変換を行うことを問題として取り上げた。これは、元の問題である対話エージェントが何らかの人型、ないし何らかのキャラクタを持つことを想定したこと、学習させるための研究資材として個人で収集できる範囲の有効なデータが、Wikipedia や青空文庫 (TODO:青空文庫) の書籍といった言った比較的書き言葉に近いテキストを用いることになるだろうと考えていること、文章生成の段階では書き言葉の方がテキストの情報を正規化して持っているのではないかという予想があったことのためだ。

この実験における書き言葉と話し言葉の例を以下に引用する。これらのデータはデータセットを入手することが出来なかったため、自作のものを用いた。データの総数は約 300 と小さめのデータセットだ。

ここで並行なデータを用意していることについて触れる。まず第一にデータを作成する際に片方のデータセットを作成した後、も片方のデータセットを作成した方が効率が良かったことを挙げることができる。また今回議論するスタイル変換は上記の例のように極めて変化が乏しいものである。そのためこの変換を “特定の条件で特定の単語を置換する” という問題として見直し、これを解く手法を提案したことを挙げることができる。以上のことから今回は並行なデータを用いて実験を行う。

Table 20: 文スタイル変換に用いる学習データ例

書き言葉	話し言葉
おはようございます。	おはよう。
明日も会社です。	明日も会社だ。
明日はゆっくりできそうです。	明日はゆっくりできそう。
きっとうまく行きますよ。	きっとうまく行くよ。

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

Sequence to Better Sequence と Sequence to Better Sequence に Denoising autoencoder を加えたもの、CopyNet を用いたものの 3 つについて同じデータセットで実験を行った。それぞれのモデルの説明は、9.5 で行う。

7.3.2　実験結果

以下に得られた結果を示す。学習精度については学習に用いたデータが少ないため議論できない。尚、S2BS は Sequence to Better Sequence、S2BS with DAE は Sequence to Better Sequence に Denoising autoencoder を加えたものを示す。

Table 21: 文スタイル変換の実験結果

実装	入力	出力
S2BS	おはようございます。	おはよう。
	応援する。	応援してる。
	今日は寒かった。	今日は寒かった。
	夕飯は？	夕飯はどうしようか？
	早く寝たい。	お風呂に入ろう。
	何か不安だなぁ。	何か口に入れてはどうでしようか
S2BS with DAE	S2BS と同じ結果が得られた	
CopyNet	おはようございます。	おはよう。
	今日は良い天気ですね。	今日は良い天気。
	こんにちは。	こんにちは。
	頑張るぞい！	頑張るぞい！
	進捗どうですか？	進捗どう？

7.3.3 考察

本実験結果は学習データが極めて少ないものの、データが極めてノイズが少ないこともあってか、ある程度求めていた出力を得ることが出来たと考えている。しかし Sequence to Better Sequence に Denoising autoencoder がどのような影響を示すのかを確認することは出来なかった。ただ学習を行って見た感想としては、 Denoising autoencoder を加えた方が学習が難しくなっているように感じたが、これは入力の一部をマスクしている性質上当然とも言えるだろう。

Sequence to Better Sequence の出力例の後者 2 つについては非常に興味深い出力と言える。勿論学習データにはこのような変換を指定していないが、このように入力文に対して飛躍した文が生成されている。しかしこの入出力には全く相関がないとは言い切れないところが面白い。例えば、“早く寝たい” から “お風呂に入ろう” という変換は、“寝る前に風呂に入る” という学習内容に含まれない “生活習慣” を学習しているとも取れるもので、つまりは 4 で話題としたような言外の知識が学習されている可能性を示唆しているとも考えられる。

CopyNet に関しては、“単語の置き換えをする” という目的を達成しているということが、適当な入力をしてもそれが変換されずに飛ばされているという点から推測できる。このことから、任意の別のペルソナを持つ発言や文章を収集し、それぞれを学習データとしたとき、ペルソナを象りやすい単語を抽出することができるのではないかという可能性を想像することができる。またデータ数が少ないという問題を考慮したとき、CopyNet はその構造上未知語への対応が比較的容易であるため、データを集めることが困難な個人の研究者にとっては有効な手段であると考えられる。

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

機械学習を用いて文章を作成する手続きの中では、ほぼ間違いなく “不自然な文” が生成されてしまう。ここで定義する不自然な文とは、語順や文法、そして意味を挙げることが出来る。

英語を用いた研究では、不自然な文と自然な文とを識別するためのタスクとして CoLA (The Corpus of Linguistic Acceptabillity) (TODO: CoLA) と呼ばれるものが存在している。

本テーマでは、機械学習モデルから生成された日本語の文を、不自然な文と自然な文とに識別するという問題設定を行い、実際に 6 から生成されたテキストに対してラベリングを行い、その識別を行う。

8.2 実験) 対話システムのエラー検知

8.2.1 実験概要

言語モデルである BERT (TODO:BERT) を用いて自作のデータセットを用いて自然な文と自然でない文を判定するファインチューニングを行った。

データは 844 の文とそのラベルであり、ラベルは 0(不自然な文), 1(自然な文) の 2 値である。6 で期待できるデータに対して少ないように考えられるが、これは 6 の出力が想定以上に良いものであり、不自然な文を十分に用意できなかったためだ。またそのうちの 10% を検証データとして用いた。　以下に実験に用いたデータの例を示す。

Table 22: 対話システムのエラー検知のデータ例

ラベル	文
0	塩は強めです。
1	コーヒーとか?
0	の袋にてます。
1	このあたりの好みは似ていますね。
1	うす。

8.2.2 実験結果

ここでの loss と accuracy は検証データに対するものである。

Table 23: 対話システムのエラー検知の実験結果

	epoch	accuracy	loss
最も accuray が高いもの	30	0.702381	2.375742
最も loss が低いもの	3	0.619048	0.712082

以下に epoch と accuracy, loss についてのグラフを示す。

8.2.3 考察

元の CoLA タスクでの精度が 60% 前後であったのに対してより良い結果が得られたため、この結果が満足できるものであると考えられる。しかし loss と accuracy、精度の関係について疑問が今後の課題として残った。実験結果で示したように、accuracy と loss の最良値をとる epoch 数は一致しておらず、グラフとして見ても理想的な外形を得ることが出来なかった。検証データを取り替えてもこれ以上の結果を得ることが出来なかったため、データを増やすか、いずれかの値を “精度” の判断基準として採用する必要があると考えられる。“一般的には”[17] loss を判断基準として用いることが多いため、こちらを採用するべきだと予測できる。

[17]厳密にどちらかと明言された文書を見つけることが出来なかった他、ここ (https://stats.stackexchange.com/questions/258166) に興味深い議論があるように、頭ごなしに loss のみを観測して過剰適合かどうかを判定するのは早計であると考えたため、“一般的” という表現を用いた。

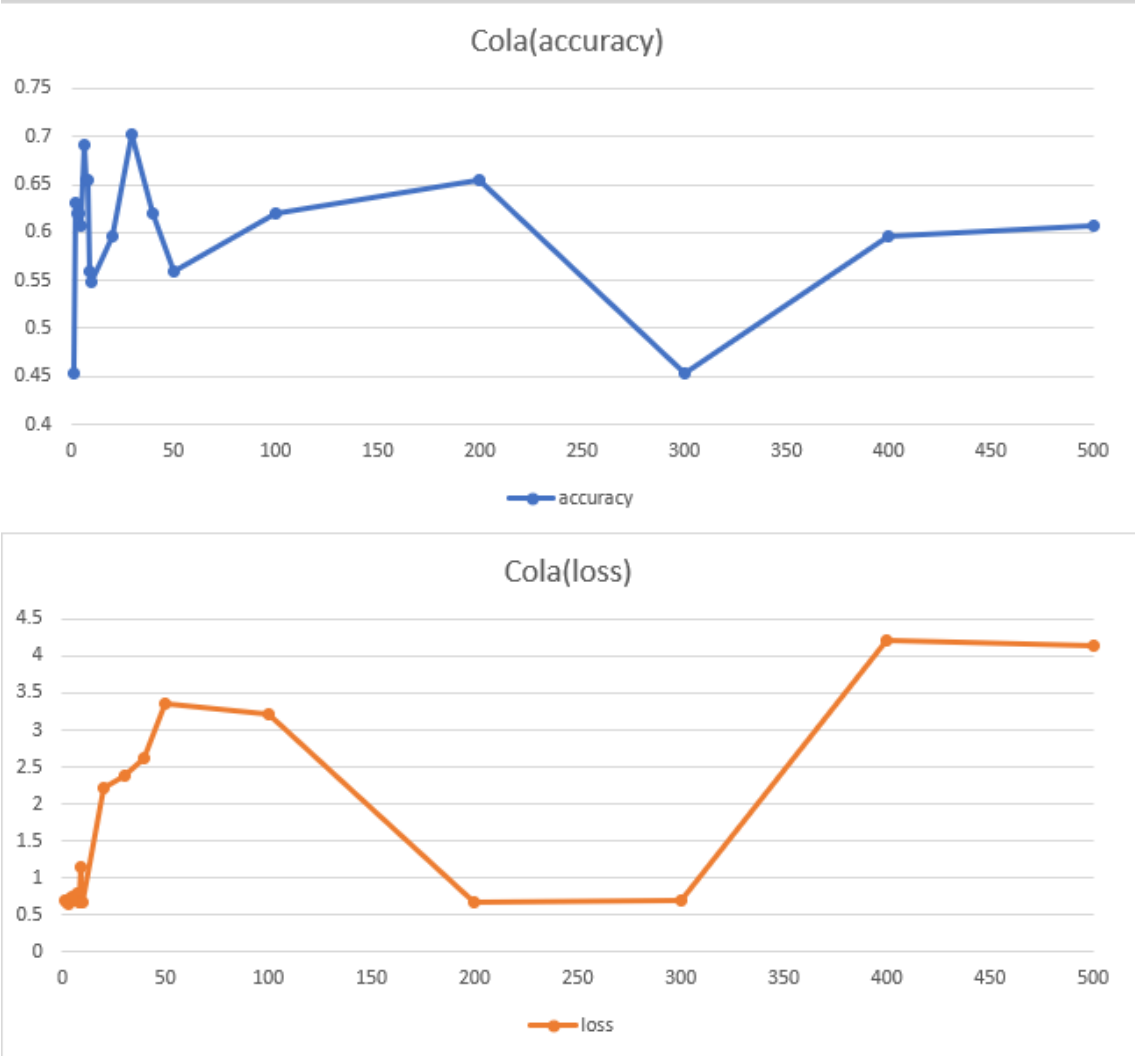

Figure 6: 対話システムのエラー検知の実験結果 における epoch と 精度の変化

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 形態素解析

9.2.3 NER

9.2.4 Word Piece

Word Piece

9.2.5 Sentence Pieces

Sentence Pieces

9.2.6 Skip-gram

Skip-gram のアルゴリズムは以下 (9.2.6) のとおりである。[18]

Skip-gram のアルゴリズム

1. 正のサンプルとして、ターゲットの単語とその周辺の単語を取り出す。
2. 負のサンプルとして、単語辞書の中からランダムにサンプルされた単語を取り出す。
3. ロジスティックス回帰を用いてこの 2 つのサンプルを区別できるようにネットワークを訓練する。
4. ネットワークの重みを単語埋め込みとみなす。

(TODO:画像)

Table 24: fasttext を用いた単語分散獲得学習のパラメータ

パラメータ名	説明
許容最低語彙頻度	語彙として認める単語の頻度。 これを下回る単語は頻度の少ない単語として学習の対象としない。
学習係数	目的関数 Adagrad の学習係数。
学習係数向上率	学習率の更新率、単語がこの数だけ訓練されると更新される。
epoch 数	語彙の数 に対して何倍訓練を行うかを決定する。

次ページに続く

[18]計算の都合上、辞書全体の単語を取り上げることが不可能なため、ネガティブサンプリングを行っている。またこのサンプリングは均一ではなく、高頻度な単語は程よく省かれるようになっている。(TODO:Distributed Representations of Words and phrases and their compositionality)

前ページからの続き

パラメータ名	説明
ネガティブサンプリング数	学習ごとに負のサンプルをどのくらい抽出するか。
ウィンドウサイズ	アルゴリズムで説明した m の値
損失関数	損失関数
dim	埋め込みベクトルの次元数

9.2.7 CNN-LSTM

CNN-LSTM

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

9.4.3 BLEU スコア

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.

Deng, J. et al. (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*.

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. 言語処理学会 第 22 回年次大会 発表論文集. Microsoft Japan Inc.

逸子, 藤村, 大曽 美恵子, and 大島 ディヴィッド義和 (2011). 言語研究の技法：データの収集と分析. Ed. by 藤村逸子、滝沢直宏. ひつじ書房.