

Contents

1	序論	1
1.1	研究背景及び目的	1
1.2	本論文の構成	2
2	対話システムの関連研究	2
3	想定する対話システムの全体像	2
4	日本語データの取り扱いについて	2
4.1	調査) 発話データ	2
4.1.1	フィルタ	2
4.1.2	調査結果	2
4.2	調査) 対話データ	2
4.3	問題設定	3
4.4	実験) 漢字かな問題に対する単語分散取得	3
4.4.1	実験概要	3
4.4.2	実験結果	3
4.4.3	考察	3
4.5	実験) 得られた単語分散を用いた極性判定	3
4.5.1	実験概要	3
4.5.2	実験結果	3
4.5.3	考察	3
5	文抽出を念頭においた不均衡分散・サイズの分類問題	3
5.1	問題設定	3
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	4
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	4
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	4
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	4
5.6	考察	4
6	機械翻訳システムを用いた対話	4
6.1	問題設定	4
6.2	実験) Seq2Seq Attention と Transformer の精度比較	4
6.2.1	実験概要	4
6.2.2	実験結果	4
6.2.3	考察	5
7	文章のスタイル変換	5
7.1	関連研究	5
7.2	問題設定	5
7.3	実験) 書き言葉→話し言葉のスタイル変換	5
7.3.1	実験概要	5
7.3.2	実験結果	5
7.3.3	考察	5
8	CoLA タスクを応用した対話システムのエラー検知	5
8.1	問題設定	5
8.2	実験) 対話システムのエラー検知	5
8.2.1	実験概要	5
8.2.2	実験結果	5
8.2.3	考察	6

9	付録	6
9.1	対話システムの関連研究	6
9.1.1	Sounding Board	6
9.1.2	2018 年の Amazon Alexa Prize	6
9.2	日本語データの取り扱いについて	6
9.2.1	単語分割	6
9.2.2	Word Piece	6
9.2.3	Sentence Pieces	6
9.2.4	Skipgram	6
9.2.5	CNN-RNN	6
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	6
9.3.1	画像データ	6
9.3.2	文データ	7
9.4	機械翻訳システムを用いた対話	7
9.4.1	Seq2Seq Attention	7
9.4.2	Transformer	7
9.5	文章のスタイル変換	7
9.5.1	Sequence to Better Sequence	7
9.5.2	CopyNet	7
9.5.3	Denoising Auto Encoder	7
9.6	CoLA タスクを応用した対話システムのエラー検知	7
9.6.1	BERT	7
10	結論	7
10.1	今後の課題	7

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、何気なしに“なんとなく良い感じに”目的を達成してくれることを祈るという有様である。この問題に対処するための手段として、入手できるだけ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スタイル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 Prize³ 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。

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

我々は一般に英語のみを話しているわけではない。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英語のものである。しかし日本語と英語には、例えば使っている文字や文型で大きく異なっていると言える。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する。

2. 文抽出を念頭においた不均衡分散・サイズの分類問題 テスト

[abstract]

まず受付やオンラインチャットなどにおいて対話システムの需要が増えていること、Amazon Alexa や Siri などを例に出して説明する。次にキャラクターを持ったマルチモーダル対話システムとして、りんなを例に上げる。

その実装例として、Amazon Alexa Prize のコンテストを例に出す。

本研究の目的として、日本語環境下で、りんなのような機能を持つシステムを構築すること、ゲーム AI への転用などを視野にいれていることを説明する。そしてシステムの概要として、対話システムという大問題に対して、いくつかの小問題に分割し、それらを組み合わせるモジュール分割という手法を用いることにしたことを説明する。

¹HRED (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こと) 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>

⁴https://twitter.com/ms_rinna

1.2 本論文の構成

1 章として論文の導入をしていること。2 章で対話システムの関連研究として 1 章の研究背景で紹介した (Amazon Alexa Prize やりん) の概要を説明していること。3 章として対話システム全体の実装として目標としている構成図を示していること。4 章については日本語のデータをどのように扱うべきか考察したこと。5 章から 8 章にかけてはそれぞれのモジュールに対する研究について説明をしていること。9 章に対してはそれぞれの章で説明が不足しているであろう内容を補足するための付録であること。を説明する。

2 対話システムの関連研究

関連研究として、Amazon Alexa Prize の問題設定の説明、2018 年、2017 年の最優秀賞団体がモジュール分割して問題を解決したことを説明する。りんなの概要についても紹介する。

3 想定する対話システムの全体像

全体像図のグラフを示す。
それぞれの問題をリストとして示す。

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があること。その例として。語尾の多彩な変化や漢字かな問題があることを紹介する。また一般に公開されている対話データセットをどのように用いるべきなのかについての考察を行ったこと、漢字かな問題に対して単語分散を得るための手法を二種類想定し、それぞれの性質を比較する。

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。

4.1.1 フィルタ

フィルタとして、タグや宛名、リンクを省いた後、4 字以上、60 字以下のデータを対象とした。その理由として、長文のツイートは説明の内容が含まれること、このデータの取扱先として深層学習を中心にした機械学習 (要約タスクを除く) を想定しているため、あまり長すぎるテキストは切り落とす可能性があること、短すぎるつぶやきはリンクやタグのみのツイートが多かったことを挙げる。

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。

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章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。
(特定の文章で分岐を行い、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、文章類似度を測り、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

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 文章のスタイル変換

7.1 関連研究

この分野の関連研究として sequence 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章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。

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

9.1.1 Sounding Board

Sounding Board

9.1.2 2018 年の Amazon Alexa Prize

Amazon Alexa Prize

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