

2023年度(令和5年度) 修士論文

GPTを用いた対話破綻検出

令和6年2月

鳥取大学大学院 持続性社会創生科学研究科
工学専攻 情報エレクトロニクスコース

自然言語処理研究室

M22J4011Y 尾崎 虎之介

概要

近年、Apple 社の「Siri」や Microsoft 社の「りんな」などの対話システムや、対話を機能の1つとして取り入れているサービスが増えている。このような対話システムが出力する文そのものは、人間が話すように自然で流暢なものとなってきている。こういった自然な発話を行う対話システムは大規模言語モデル(以下、LLM)を基に作成されており、大量のウェブ上のテキストデータを学習に用いている。

しかし、対話システムと人間の対話において、人間同士での対話と同じように対話が行えるかという点必ずしもそうとはいかない。その理由として、対話システム側が論理的に破綻している内容を出力する可能性があるからである。ここでの対話破綻は文脈との整合性が保てないことであったり、そもそも知識ベースで間違った情報を出力することで引き起こされる。先行研究の対話破綻検出チャレンジ [1, 2, 3, 4] では、対話システムとユーザの対話履歴に対し、ラベルを付与することをタスクとしていたが、人手でのラベル付与といった観点から十分なデータ数で開発・評価を行えていなかった。

そこで本研究では、LLM の1つである GPT[5] の汎化性能に着目した。この言語モデルは膨大な事前学習を行っているだけでなく、推論時に1つもしくは少量のタスクのデモンストレーションを提示する One-shot learning と Few-shot learning[6] を適用することによって性能向上が期待できる。このことから、対話履歴とラベル付与例を One-shot と Few-shot の提示例として GPT にプロンプトする手法を提案する。

実験では、先行研究のデータを対象に先行研究の結果と提案手法の比較を行った。評価指標はラベル一致システムと分布距離システムの2種類があり、ラベル一致システムは先行研究と比べ提案手法の正解率が 0.5841 から 0.6562、F 値が 0.5387 から 0.5485 と性能向上が確認できた。しかし、分布距離システムの評価指標では、先行研究と比べて性能が劣るということが分かった。

目次

第1章	はじめに	1
第2章	先行研究	3
2.1	対話破綻検出チャレンジ (DBDC)	3
2.2	GPT(Generative Pre-Training)	4
2.2.1	Improving Language Understanding by Generative Pre-Training	4
2.2.2	Language Models are Few-Shot Learners	4
第3章	提案手法	5
3.1	提案手法の概要	5
3.2	命令とルール定義	7
3.3	対話履歴とラベル付与例	8
3.4	ラベルを付与する対話の提示	12
第4章	実験	14
4.1	使用する GPT モデル	14
4.2	実験データ	14
4.3	データセット内のラベル付けされた対話履歴	16
4.4	ラベル付与例の選定	18
4.5	評価方法	19
4.5.1	ラベル一致システム	20
4.5.2	分布距離システム	20
4.6	ベースライン	21
4.7	実験結果	21
4.7.1	データセット全体の結果	22
4.7.2	各データセットの結果	23
4.7.3	データセット全体をまとめて集計した結果	31

4.7.4	評価指標ごとの上限値	32
第5章	考察	33
5.1	データセットごとのF値性能の調査	33
5.2	データセットごとのラベルの偏り	33
5.2.1	ラベル付与が適切な対話履歴	34
5.2.2	ラベル付与が不適切な対話履歴	37
第6章	おわりに	40

目次

3.1 提案手法概要	5
3.2 One-shot, Few-shot learning 箇所	6

表目次

4.1	モデルごとのトークン制限	14
4.2	データセットの統計 (JCDC-1146, DBDC1, DBDC2)	15
4.3	データセットの統計 (DBDC3, DBDC4)	15
4.4	データセットの統計 (Live Competition)	15
4.5	モデルごとの提示例の数	18
4.6	モデルごとの提示ラベルの割合 (GPT-3)	18
4.7	モデルごとの提示ラベルの割合 (GPT-3.5)	19
4.8	モデルごとの提示ラベルの割合 (GPT-4)	19
4.9	正解率, F 値 (マクロ平均法)	22
4.10	JS divergence(マクロ平均法)	22
4.11	Mean squared error(マクロ平均法)	22
4.12	正解率, F 値 (DCM)	23
4.13	JS divergence(DCM)	23
4.14	Mean squared error(DCM)	23
4.15	正解率, F 値 (DIT)	24
4.16	JS divergence(DIT)	24
4.17	Mean squared error(DIT)	24
4.18	正解率, F 値 (IRS)	25
4.19	JS divergence(IRS)	25
4.20	Mean squared error(IRS)	25
4.21	正解率, F 値 (Live Competition IRS)	26
4.22	JS divergence(Live Competition IRS)	26
4.23	Mean squared error(Live Competition IRS)	26
4.24	正解率, F 値 (Live Competition MMK)	27
4.25	JS divergence(Live Competition MMK)	27
4.26	Mean squared error(Live Competition MMK)	27

4.27	正解率, F 値 (Live Competition MRK)	28
4.28	JS divergence(Live Competition MRK)	28
4.29	Mean squared error(Live Competition MRK)	28
4.30	正解率, F 値 (Live Competition TRF)	29
4.31	JS divergence(Live Competition TRF)	29
4.32	Mean squared error(Live Competition TRF)	29
4.33	正解率, F 値 (Live Competition ZNK)	30
4.34	JS divergence(Live Competition ZNK)	30
4.35	Mean squared error(Live Competition ZNK)	30
4.36	正解率, F 値 (マイクロ平均法)	31
4.37	JS divergence(マイクロ平均法)	31
4.38	Mean squared error(マイクロ平均法)	31
4.39	正解率, 適合率 (X), 再現率 (X), F 値 (X) の上限値	32
4.40	適合率 (T+X), 再現率 (T+X), F 値 (T+X) の上限値	32
4.41	JS divergence の上限値	32
4.42	Mean squared error の上限値	32

第1章 はじめに

近年, Apple 社の「Siri」や Microsoft 社の「りんな」などの対話システムや, 対話を機能の1つとして取り入れているサービスが増えている. このような対話システムが出力する文そのものは, 人間が話すように自然で流暢なものとなってきている. こういった自然な発話を行う対話システムは大規模言語モデル(以下, LLM)を基に作成されており, 大量のウェブ上のテキストデータを学習に用いて学習を行っている.

しかし, 対話システムと人間の対話において, 人間同士での対話と同じように対話が行えるかという点必ずしもそうとはいかない. その理由として, 対話システム側が論理的に破綻している内容を入力する場合があるからである. ここでの対話破綻は文脈との整合性が保てないことであったり, そもそも知識ベースで間違った情報を入力することで引き起こされる. 先行研究の対話破綻検出チャレンジでは, 対話システムとユーザの対話履歴に対しラベルを付与することをタスクとしていたが, 人手でのラベル付与といった観点から十分なデータ数で開発・評価を行えていなかった.

そこで本研究では, LLMの1つである GPT[5]の汎化性能に着目した. この言語モデルは膨大な事前学習を行っているだけでなく, Fine-tuningの代わりとして推論時に少量のタスクのデモンストレーションを提示する方法の One-shot learning, Few-shot learning[6]によって性能向上が期待できる. このことから, 対話履歴とラベル付与例をデモンストレーションの提示例として GPT にプロンプトする手法を提案する.

本研究の主張点を以下に示す.

- GPT に対して, One-shot learning と Few-shot learning を適用することにより, GPT の持つ汎化性能を対話破綻検出に適用し, 学習データの不足を解消する.
- 対話履歴に対して, GPT がラベルを付与するようにプロンプトを調整した.
- 先行研究では Fine-tuning を適用した BERT で, 正解率が 0.584 と F 値が 0.539 であったが, 提案手法により, 正解率が 0.656 と F 値が 0.549 と性能向上が確認できた.
- しかし, ラベルの分布距離を評価する評価指標では先行研究に劣る性能となった.

本論文の構成は以下の通りである.

第 2 章 本研究に関連する研究としてどのような研究が行われてきたかを記述し, その研究と本研究との関連を説明する.

第 3 章 提案手法について説明する.

第 4 章 本研究が行った実験についての説明と, その結果について記述する.

第 5 章 考察を行う.

第 6 章 まとめを行う.

第2章 先行研究

本章では、先行研究について記述する。

2.1節では、東中ら [1, 2, 3, 4] が行った評価型ワークショップについて記述する。2.2節では、本研究で用いた深層学習モデルである GPT(Generative Pre-Training) モデルの紹介として Radford ら [5], Brown ら [6] が行った研究について記述する。

2.1 対話破綻検出チャレンジ (DBDC)

評価型ワークショップである対話破綻検出チャレンジ (Dialogue Breakdown Detection Challenge; DBDC) にて東中ら [1, 2, 3, 4] は、開発・評価用に人間と対話システムとの対話データを収集した。その後、複数名のアノテータが対話システム側の発話に対し、破綻しているかどうかのラベルを付与した。ラベルは NB(not breakdown, ‘O’), PB(possible breakdown, ‘T’), B(breakdown, ‘X’) の3つに分けられており、それぞれの定義は以下の通りである。

NB(O)：破綻ではない 当該システム発話のあと対話を問題無く継続できる。

PB(T)：破綻とは言い切れないが、違和感を感じる発話 当該システム発話のあと対話をスムーズに継続することが困難。

B(X)：あきらかにおかしいと思う発話，破綻 当該システム発話のあと対話を継続することが困難。

以上のラベル付けされた対話データを用いて、参加者らは深層学習モデルやルールベースによる手法によって対話破綻検出を行った。第4回となる DBDC4 では、日本語のデータセットに対し BERT や LSTM といった方法が用いられた。評価尺度は2種類あり、ラベルを付けた箇所を検出器が正しく検出できるか、実際のラベルの分布を推定できるかで評価実験を行った。詳しくは4.5節で説明を行う。

2.2 GPT(Generative Pre-Training)

本研究で用いた深層学習モデルである GPT について記述する。

2.2.1 節では, GPT の概要について記述する. 2.2.2 節では, GPT の汎化性能について記述する.

2.2.1 Improving Language Understanding by Generative Pre-Training

自然言語処理 (NLP) において, 推論やテキスト分類といったタスクを行うには, 手作業によってラベル付けされた大量のデータが必要となる. しかし, ラベル付けされたデータは限られていることから, 言語情報を活用できるドメインのタスクが限られてしまう.

Radford ら [5] は, ラベル付けの行われていないテキストによって事前学習を行った後に, 推論やテキスト分類といった複数のタスクが含まれたデータを用いて, モデルに対し教師ありの Fine-tuning を行った. これにより, ラベル付きデータの必要数を削減することができ, 幅広いタスクに対応することが可能となった.

2.2.2 Language Models are Few-Shot Learners

GPT モデルを用いることで, 必要となるラベル付きデータの数は削減することが可能となった. しかし, Fine-tuning には何千ものデータが必要である. また, Fine-tuning の際に, タスクによっては汎化性能が出ない場合がある.

そこで, 人間が多くの教師ありデータを必要とせず, 簡潔な指示や少ない数のデモンストレーションのみで言語タスクを解くという観点から, GPT モデルに対し One-shot learning や Few-shot learning が有用であるとされている. この手法は, 推論時にタスクのデモンストレーションを 1 つ, もしくは複数の比較的少量の例示的な例を提示した後に, モデルに対しタスクを行わせる手法である. Brown ら [6] は, メタ学習や言語モデルの大きさを大きくすることにより, Zero-shot learning, One-shot learning, Few-shot learning において性能の向上が可能であるとした. この研究により, ユーザ側が少量の具体的な例を提示するだけで言語モデルが新たなタスクに対応することができ, GPT モデルの柔軟性を実証することができた.

第3章 提案手法

本章では、本研究の提案手法について記述する。

3.1 節では、本研究の提案手法の概要についての説明を記述している。3.2 節と 3.3 節と 3.4 節では、GPT に対しての具体的なプロンプトについて記述している。

3.1 提案手法の概要

本研究では、GPT モデルの持つ幅広いタスクに対応する認識性能を活かし、ユーザと対話システムの対話における破綻検出の精度向上を目的とする。そこで提案手法では、GPT モデルに対し One-shot learning と Few-shot learning を適用することで、タスク内容と対話履歴に対する付与例を1つもしくは複数提示したのち、最終的にラベルが付与されていない対話履歴を提示することでラベルを付与させる (図 3.1)。

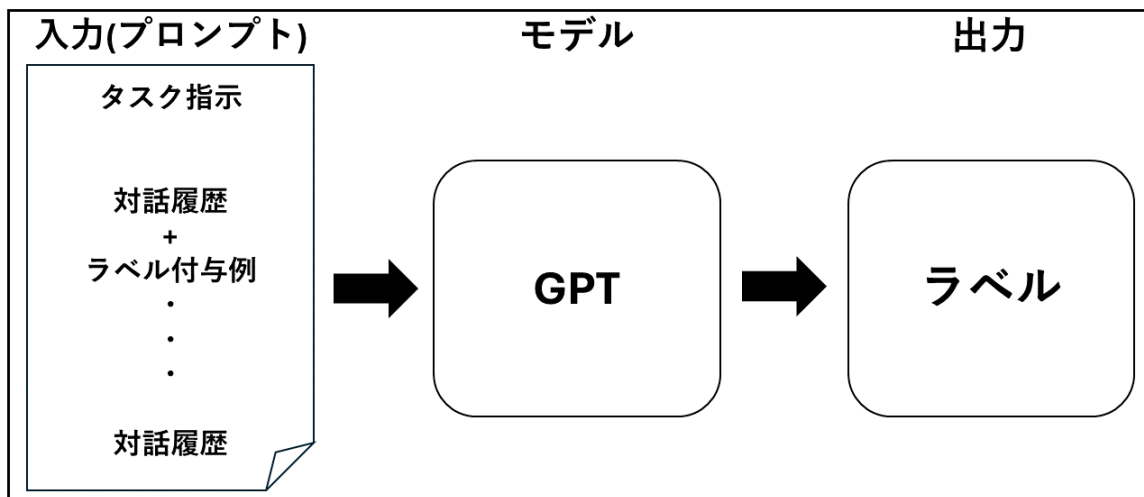


図 3.1: 提案手法概要

ラベル付与のプロンプトは3つに分割できる。まず最初に、命令とルール定義である。ここでは行うタスクの内容とラベル付与の基準を提示する。続いて、対話履歴の提示と

ラベル付与例を複数提示することにより、どの箇所にもいったラベルを付与するのかのデモンストレーションを提示する(図3.1)。最後に、実際にラベルを付与する対話履歴を提示し、GPT にシステムの発話箇所が破綻しているかどうかのラベルを付与させる。これら3つを順に1つにまとめてプロンプトとする。また、具体的なプロンプトの詳細については3.2節と3.3節と3.4節に示す。

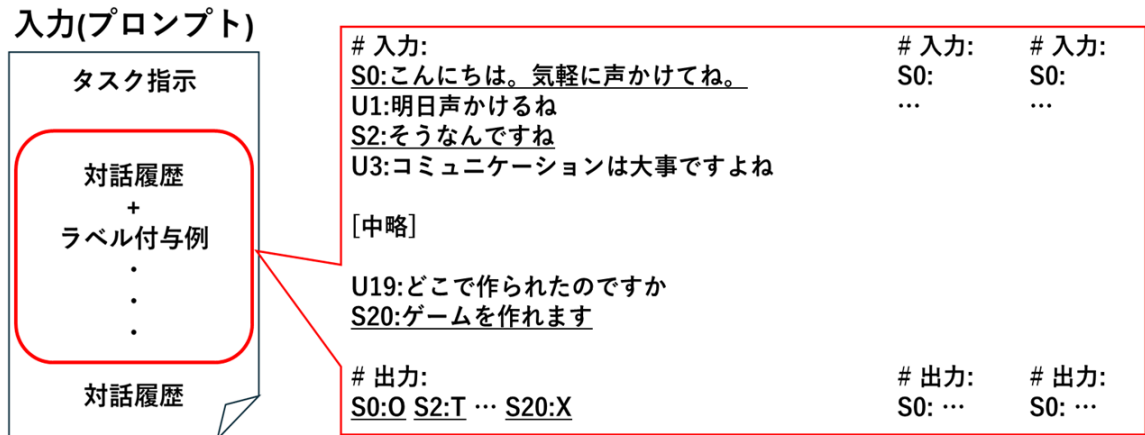


図 3.2: One-shot, Few-shot learning 箇所

3.2 命令とルール定義

まず GPT に対し、行うタスクの内容を提示する。ここでは、システムの発話が S であり、ユーザの発話が U であることを記述する。そして、システムの発話箇所のみを対象に、破綻しているかどうか検証することを記述する。続いて、GPT 側がどのように出力すればいいのかを [# ルール:] 部に提示する。具体的な内容として、破綻の基準、最終的な出力の仕方をここに記述する。

以下は、GPT に対して提示するプロンプトの命令とルール定義である。

以下はシステム (S) とユーザ (U) の対話です。システムの発話である S0 から S20 が破綻しているか検証してください。ただし、以下のルールに従って出力をしてください。

ルール:

破綻ではない場合、システムの発話のあと対話を問題なく継続できる場合はラベル”O”を出力すること。

破綻とは言い切れないが違和感がある場合、システムの発話のあと対話をスムーズに継続することが困難な場合はラベル”T”を出力すること。

破綻している場合、システムの発話のあと対話を継続することが困難な場合はラベル”X”を出力すること。

システムの発話番号とラベルのみを出力すること。

出力は 1 行でまとめること。

3.3 対話履歴とラベル付与例

次に、ユーザと対話システムの対話履歴と、対話システムの発話箇所に対するラベル付与例のセットを1つもしくは複数提示する。この例では、対話システム(S)とユーザ(U)の対話が20ターン続いている。ラベルはシステムの発話箇所のみが付与されており、システムの発話箇所はS0, S2, S4のように偶数で振り分けられている。そのため、[#出力:]部にはS0, S2, S4のように、[#入力:]部と対応付けされた発話箇所にラベルが付与される。2.1節でも説明したように、システムは発話箇所が破綻しているかどうかを‘O’, ‘T’, ‘X’の3つのラベルを付与することで判定している。

以下の例では、この[#入力:]部と[#出力:]部を1セットとし、5つの対話とラベルのセットを提示している。こうすることにより、GPTが‘S’かつ偶数が並ぶ文字列と同じ行が対話システムの発話箇所であると判断することができ、実際にラベルを付与する対話に対して、どのようにラベルを出力すればいいのかを判断させることができる。

入力:

S0: こんばんは。声かけてくれるのを待ってたんだ。

U1: こんばんは。

S2: こん

U3: 今日は暑かったですね。

S4: そういえば思い出したんですけど、遊園地にある大きな池は、防災も兼ねてるらしいですよ。

U5: そうなんですか。それは知りませんでした。

S6: 外で見ようですよ

U7: 遊園地は好きですか

S8: そうですね、子供の頃行きましたね。

U9: ジェットコースターは苦手です

S10: ジェットコースターに乗った直後のカタカタって昇っていく音がたまらなく好きなんだ。

U11: ジェットコースターが落ちる時が怖いです

S12: カタカタを使いますよねー

U13: どんな乗り物が好きですか

S14: 乗り物は自転車が好きですね

U15: 私もよく乗ります

S16: 乗り物は好きですね

U17: サイクリングは楽しいです

S18: サイクリングは楽しいんですか?? 乗り物は大好きですね

U19: ドライブも好きです

S20: ドライブは好きですか。行き先を決めてないドライブもたまにはいいよね。

出力:

S0:O S2:X S4:T S6:X S8:O S10:O S12:X S14:O S16:O S18:T S20:O

入力:

S0: こんにちは。夏といえばスイカだね。

U1: そうですね。でもスイカはあまり好きではないです。

S2: スイカで有名か??

U3: 違います。

[中略]

U17: それにしても、気持ちいいとはいえ、今日は暑いですね。

S18: 屋内から屋外へ出た時の、開放感ってすっきりするよね。

U19: そうですね。

S20: そうですよ

出力:

S0:O S2:X S4:X S6:X S8:X S10:X S12:O S14:X S16:O S18:X S20:O

入力:

S0: こんにちは。気軽に声かけてね。

U1: こんにちは

S2: おはようございます

U3: 好きなスポーツは何？

[中略]

U17: とんこつ味はうまいよね

S18: 相性は注意が必要です

U19: そうか

S20: そうさ

出力:

S0:O S2:T S4:X S6:T S8:X S10:O S12:O S14:O S16:X S18:X S20:O

入力:

S0: こんにちは。海へ行きたいね。

U1: 最近だといつ海に行った？

S2: 海に行かないんですね？

U3: いや、一週間前に言ったよ

[中略]

U17: それな

S18: だよな

U19: じゃあね。

S20: じゃあの

出力:

S0:O S2:X S4:O S6:T S8:T S10:T S12:X S14:X S16:O S18:O S20:O

入力:

S0: こんにちは。ちょうど退屈してたんだ。

U1: こんにちは。今日も暑いね。

S2: 健康に注意されていますか？

U3: 小まめに水分補給するようにしているよ。君は？

[中略]

U17: 私も夏休みにブラジルへ旅行に行くよ。

S18: 夏休みで福岡に行きます

U19: 福岡のラーメンは美味しいよね。

S20: ラーメン屋で食べますよねー

出力:

S0:O S2:O S4:X S6:X S8:T S10:X S12:T S14:O S16:X S18:X S20:T

3.4 ラベルを付与する対話の提示

最終的にラベルを付与する対話の [# 入力:] 部のみを提示することで、システムの発話箇所に対応するラベルを GPT に分類させる。

入力:

S0: こんにちは。気軽に声かけてね。

U1: 明日声かけるね

S2: そうなんですね

U3: コミュニケーションは大事ですよ

S4: コミュニケーションは大事かね？

U5: 信頼関係が築くためですよ

S6: 信頼は長い時間をかけて築くけど、崩すのは一瞬だね。

U7: お互いに尊敬し合うことになりますね

S8: 尊敬出来る人が誰かいるだけでも人生は変わるでしょう。

U9: 誰を尊敬していますか？

S10: 人生は長いようで短いから好きなことをやった方が良いでしょう。

U11: 趣味は大事ですね

S12: 趣味はゲームかな。

U13: どのようなゲームですか

S14: 趣味がある人って毎日が楽しそうで羨ましいよね。

U15: スポーツの趣味は良いですね

S16: 趣味はゲームかな。

U17: ゲームも面白いですよ

S18: 世界で最初に作られたコンピューターゲームは、チェスのゲームなんだよ。

U19: どこで作られたのですか

S20: ゲームを作れます

GPTは3.3節において、対話履歴とそれに対応するラベル付与の Few-shot learning を行っている。これにより、GPTは3.4節の対話履歴に対して、3.3節を参考にラベルの付与を行う。この例での GPT-4 の付与結果は以下のようになる。

出力:

S0:O S2:T S4:X S6:O S8:O S10:T S12:O S14:X S16:T S18:T S20:X

第4章 実験

本章では、本研究で行った実験について記述する。

4.1 節では、使用した GPT モデルについて記述している。4.2 節では、実験で用いたデータセットについて記述している。4.3 節では、データセットの具体的な形式について記述している。4.4 節では、GPT モデルごとのラベル付与数について記述している。4.5 節では、評価方法について記述している。4.6 節では、ベースラインについて記述している。4.7 節では、実験結果について記述している。

4.1 使用する GPT モデル

一般公開されている GPT モデルは複数存在するが、今回使用するモデルは GPT-3(text-davinci-003), GPT-3.5(gpt-3.5-turbo-16k-0613), GPT-4 の 3 つを使用する。GPT-3 と GPT-3.5 は API[7] で実験を行い、GPT-4 は 2023 年の 8 月にウェブ版 [8] で実験を行った。

各モデルの違いは、事前学習に用いられるデータ量やパラメータ数である。また、今回使用したウェブ版や API の GPT では一度に入力できるトークン制限にも違いがある(表 4.1)。

表 4.1: モデルごとのトークン制限

	GPT-3	GPT-3.5	GPT-4
トークン数	4,096	16,385	4,096

4.2 実験データ

以下に今回使用したデータセットの統計を示す。使用したデータセットは Project Next NLP 対話タスク [9] で収集された雑談対話コーパス、対話破綻検出チャレンジ [1, 2, 3, 4] で収集された DBDC コーパスである。このデータセットは対話システムとユーザの対話履歴を収集した後、アノテータによってシステム側の発話が破綻しているかどうかをラ

ベル‘O’, ‘T’, ‘X’のいずれかを付与するかで判断している。また、収集に用いられた対話システムは現在主流となっている深層学習を用いたものではない。今回評価する対話データは表4.3のDBDC4と表4.4に該当し、合計で223の対話セッション数が含まれている。その内訳としてDCM, DIT, IRSからそれぞれ50対話ずつと、対話システムライブコンペティションから73対話を評価する。

表 4.2: データセットの統計 (JCDC-1146, DBDC1, DBDC2)

	JCDC-1146		DBDC1	DBDC2		
	init100	rest1046	開発/評価	DCM(開発/評価)	DIT(開発/評価)	IRS(開発/評価)
対話数	100	1046	20/80	50/50	50/50	50/50
アノテータ数	24	2 or 3	30	30	30	30
O	59.2%	58.3%	37.1%	39.8%	33.0%	37.4%
T	22.2%	25.3%	32.2%	30.2%	27.4%	24.3%
X	18.6%	16.4%	30.6%	29.9%	39.5%	38.3%
Fleiss' Kappa(O, T, X)	0.28	0.28	0.20	0.31	0.24	0.36
Fleiss' Kappa(O, T+X)	0.40	0.40	0.27	0.44	0.38	0.48

表 4.3: データセットの統計 (DBDC3, DBDC4)

	DBDC3			DBDC4		
	DCM(評価)	DIT(評価)	IRS(評価)	DCM(評価)	DIT(評価)	IRS(評価)
対話数	50	50	50	50	50	50
アノテータ数	30	30	30	15	15	15
O	34.9%	25.3%	29.3%	47.7%	34.0%	38.8%
T	34.2%	28.3%	23.8%	31.6%	34.3%	29.2%
X	30.9%	46.4%	46.9%	20.7%	31.8%	32.0%
Fleiss' Kappa(O, T, X)	0.24	0.14	0.27	0.22	0.13	0.29
Fleiss' Kappa(O, T+X)	0.32	0.20	0.37	0.28	0.17	0.40

表 4.4: データセットの統計 (Live Competition)

	Live Competition				
	MMK(開発/評価)	MRK(開発/評価)	TRF(開発/評価)	ZNK(開発/評価)	IRS(開発/評価)
対話数	15/14	15/14	14/16	16/14	13/15
アノテータ数	30	30	30	30	30
O	61.0%	48.9%	65.7%	42.7%	34.6%
T	27.4%	34.0%	22.4%	31.4%	28.7%
X	11.6%	17.1%	12.0%	26.0%	36.7%
Fleiss' Kappa(O, T, X)	0.11	0.15	0.19	0.21	0.29
Fleiss' Kappa(O, T+X)	0.16	0.21	0.26	0.29	0.27

4.3 データセット内のラベル付けされた対話履歴

対話システムの発話 (S) は複数のアノテータによって ‘O’, ‘T’, ‘X’ のラベルが付与されている。データセットは json 形式で保存されており、以下に示すデータセットの例では、15 人のアノテータによってラベルが付与されており、冒頭のシステム発話箇所 S では、全アノテータが破綻でないと判断したため、ラベル ‘O’ が 15 個付与されている。

dialogue-id : 1550628075

speaker-id : DBDC4-01

group-id :

S: こんにちは。気軽に声かけてね。 O O O O O O O O O O O O O O O O O O

U: 明日声かけるね

S: そうなんですね T O T X O T X O X X X O T O T

U: コミュニケーションは大事ですよ

S: コミュニケーションは大事かね? T O X T O O T O X T O O T O O

U: 信頼関係が築くためですよ

S: 信頼は長い時間をかけて築くけど、崩すのは一瞬だね。 O O O O O O O T T
O O O O O O

U: お互いに尊敬し合うことになりますね

S: 尊敬出来る人が誰かいるだけでも人生は変わるでしょう。 T O O O O O O O O
X O O O O O O O

U: 誰を尊敬していますか?

S: 人生は長いようで短いから好きなことをやった方が良いよ。 X O T X X X X
X X X X X T O T

U: 趣味は大事ですね

S: 趣味はゲームかな。 O O O X O T T T T O O O T O O

U: どのようなゲームですか

S: 趣味がある人って毎日が楽しそうで羨ましいよね。 X O O X X T T X T X T X
T O T

U: スポーツの趣味は良いですね

S: 趣味はゲームかな。XOTXXXTTXXXTXTOX

U: ゲームも面白いですよ

S: 世界で最初に作られたコンピューターゲームは、チェスのゲームなんだよ。T
OOOOOOOOTOOOTO

U: どこで作られたのですか

S: ゲームを作れますXTOXXXXXXXXXXTOX

4.4 ラベル付与例の選定

提示するラベル付与例は、同じシステムで収集された学習データの中からランダムに選定した。例えば、DCMの評価データをGPTにラベル付与させる場合は、同じくDCMの開発データからランダムで選定をすることで、システムの発話の傾向が似ているものになるようにした。また、使用するGPTモデルによってプロンプトに用いることのできるトークン数に限りがある。そのため、モデルやデータセットによっては提示するラベル付与例の数が異なる場合がある(表4.5)。

表 4.5: モデルごとの提示例の数

	GPT-3	GPT-3.5	GPT-4
DCM	5	5	5
DIT	2	5	2
IRS	3	5	3
Live Competition IRS, MMK, MRK, TRF, ZNK	1	5	1

また、表4.2, 表4.3, 表4.4に記述しているように、データセットごとにラベルの分布が異なる。これにより、プロンプトのFew-shotの箇所で用いたラベルの付与の割合もモデルによって異なる。それぞれのラベルの割合を表4.6と表4.7と表4.8に示す。

表 4.6: モデルごとの提示ラベルの割合 (GPT-3)

	GPT-3		
	O	T	X
DCM	0.42 (23/55)	0.18 (10/55)	0.40 (22/55)
DIT	0.27 (6/22)	0.36 (8/22)	0.36 (8/22)
IRS	0.42 (23/55)	0.18 (10/55)	0.40 (22/55)
Live Competition IRS	0.30 (10/33)	0.15 (5/33)	0.55 (18/33)
Live Competition MMK	0.75 (12/16)	0.25 (4/16)	0.00 (0/16)
Live Competition MRK	0.25 (4/16)	0.63 (10/16)	0.13 (2/16)
Live Competition TRF	0.94 (15/16)	0.00 (0/16)	0.06 (1/16)
Live Competition ZNK	0.25 (4/16)	0.44 (7/16)	0.31 (5/16)

表 4.7: モデルごとの提示ラベルの割合 (GPT-3.5)

	GPT-3.5		
	O	T	X
DCM	0.42 (23/55)	0.18 (10/55)	0.40 (22/55)
DIT	0.29 (16/55)	0.33 (18/55)	0.38 (21/55)
IRS	0.31 (17/55)	0.13 (7/55)	0.56 (31/55)
Live Competition IRS	0.43 (34/80)	0.24 (19/80)	0.34 (27/80)
Live Competition MMK	0.75 (60/80)	0.23 (18/80)	0.03 (2/80)
Live Competition MRK	0.56 (45/80)	0.39 (31/80)	0.05 (4/80)
Live Competition TRF	0.84 (67/80)	0.08 (6/80)	0.09 (7/80)
Live Competition ZNK	0.54 (43/80)	0.21 (17/80)	0.25 (20/80)

表 4.8: モデルごとの提示ラベルの割合 (GPT-4)

	GPT-4		
	O	T	X
DCM	0.42 (23/55)	0.18 (10/55)	0.40 (22/55)
DIT	0.27 (6/22)	0.36 (8/22)	0.36 (8/22)
IRS	0.30 (10/33)	0.15 (5/33)	0.54 (18/33)
Live Competition IRS	0.31 (5/16)	0.25 (4/16)	0.44 (7/16)
Live Competition MMK	0.75 (12/16)	0.25 (4/16)	0.00 (0/16)
Live Competition MRK	0.25 (4/16)	0.63 (10/16)	0.13 (2/16)
Live Competition TRF	0.94 (15/16)	0.00 (0/16)	0.06 (1/16)
Live Competition ZNK	0.25 (4/16)	0.44 (7/16)	0.31 (5/16)

4.5 評価方法

対話破綻検出タスクにおいて、確立された評価尺度は存在しない。そのため、DBDCではラベル一致システムと分布距離システムの2つの評価尺度を用いている。本研究でも同様の評価方法を用いており、それぞれの説明については4.5.1節と4.5.2節に記述する。

4.5.1 ラベル一致系統

ラベル一致系統では、まず対話システムの発話箇所に対して、付与されたラベルの多数決を取り、該当の発話が破綻かどうかの正解ラベルを1つに決める。その後、評価するデータの正解ラベルと GPT が付与したラベルを比較することで評価を行う。評価尺度は以下となる。

- 正解率：O,T,X の全ラベル一致率
- F 値 (O,T,X)：ラベル X を正解としたときの再現率，適合率，F 値
- F 値 (O,T+X)：ラベル X とラベル T を正解としたときの再現率，適合率，F 値

4.5.2 分布距離系統

分布距離系統では、Jensen-Shannon Divergence による分布間距離 (JS divergence) と、分布間距離の平均二乗誤差 (Mean squared error) の 2 種を用いる。この 2 種の評価尺度はラベル分布をそのまま用いる場合 (O,T,X)，ラベル T とラベル X を同一ラベルとみなした場合 (O,T+X)，ラベル O とラベル T を同一ラベルとみなした場合 (O+T,X) の評価尺度があり、合計で 6 つの評価尺度としている。

- JS Divergence(O,T,X)：Jensen-Shannon Divergence による分布間の距離
- JS Divergence(O,T+X)：ラベル T と X を同一の破綻ラベルとみなした場合の Jensen-Shannon Divergence による分布間の距離
- JS Divergence(O+T,X)：ラベル O と T を同一の破綻ラベルとみなした場合の Jensen-Shannon Divergence による分布間の距離
- Mean Squared Error(O,T,X)：分布間の平均二乗誤差
- Mean Squared Error(O,T+X)：ラベル T と X を同一の破綻ラベルとみなした場合の分布間の平均二乗誤差
- Mean Squared Error(O+T,X)：ラベル O と T を同一の破綻ラベルとみなした場合の分布間の平均二乗誤差

4.6 ベースライン

ベースラインとして、DBDC4で用いられたCRFに対し、過去のワークショップで用いられた学習データと評価データを学習させたものをベースラインとした。ただし、DBDC1のデータセット(150対話)は論文サイトのリンクが切れていたため、本研究では学習に用いていない。そのため、実際に用いたデータは表4.2のJCDC-1146、DBDC2の開発データと評価データ、表4.3のDBDC3の開発データと評価データ、表4.4のLive Competitionの開発データである。

4.7 実験結果

実験結果について記述する。

4.7.1節では、マクロ平均法で集計した結果を記述している。4.7.2節では、各データセットの結果を記述している。4.7.3節では、マイクロ平均法で集計した結果を記述している。4.7.4節では、評価指標ごとの上限値を記述している。

また、先行研究の集計結果は全体の結果のみで、どのように集計したのか詳細な記述がなかった。そのため、4.7.1節のマクロ平均と4.7.3節マイクロ平均で記述している先行研究の結果はいずれも同じ値を記述している。

4.7.1 データセット全体の結果

データセットごとの結果を平均したラベル一致系統の評価結果を表 4.9 に示す.

表 4.9: 正解率, F 値 (マクロ平均法)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
先行研究 (best)	0.5841	0.5387	0.7254
ベースライン	0.5608	0.2984	0.5626
GPT-3	0.5518	0.3257	0.6503
GPT-3.5	0.5662	0.2561	0.5352
GPT-4	0.6562	0.5485	0.7086

データセットごとの結果を平均した分布距離系統の評価結果を表 4.10 と表 4.11 に示す.

表 4.10: JS divergence(マクロ平均法)

	O,T,X	O,T+X	O+T,X
先行研究 (best)	0.0947	0.0601	0.0612
ベースライン	0.3715	0.2631	0.2099
GPT-3	0.3613	0.2453	0.1687
GPT-3.5	0.3549	0.2637	0.1662
GPT-4	0.2992	0.1949	0.1536

表 4.11: Mean squared error(マクロ平均法)

	O,T,X	O,T+X	O+T,X
先行研究 (best)	0.0463	0.0635	0.0507
ベースライン	0.1906	0.2291	0.1836
GPT-3	0.1905	0.2087	0.1386
GPT-3.5	0.1840	0.2299	0.1356
GPT-4	0.1464	0.1516	0.1225

4.7.2 各データセットの結果

各データセットの結果を表 4.12 から表 4.35 に示す.

DCM

表 4.12: 正解率, F 値 (DCM)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
ベースライン	0.485 (267/550)	0.249	0.607
GPT-3	0.529 (291/550)	0.400	0.603
GPT-3.5	0.558 (307/550)	0.292	0.442
GPT-4	0.660 (363/550)	0.557	0.748

表 4.13: JS divergence(DCM)

	O,T,X	O,T+X	O+T,X
ベースライン	0.419	0.286	0.240
GPT-3	0.386	0.278	0.181
GPT-3.5	0.369	0.301	0.182
GPT-4	0.293	0.195	0.140

表 4.14: Mean squared error(DCM)

	O,T,X	O,T+X	O+T,X
ベースライン	0.225	0.254	0.221
GPT-3	0.206	0.246	0.155
GPT-3.5	0.194	0.273	0.157
GPT-4	0.145	0.153	0.111

DIT

表 4.15: 正解率, F 値 (DIT)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
ベースライン	0.449 (247/550)	0.466	0.793
GPT-3	0.496 (273/550)	0.389	0.804
GPT-3.5	0.445 (245/550)	0.331	0.793
GPT-4	0.505 (278/550)	0.567	0.841

表 4.16: JS divergence(DIT)

	O,T,X	O,T+X	O+T,X
ベースライン	0.419	0.228	0.311
GPT-3	0.383	0.222	0.192
GPT-3.5	0.402	0.229	0.200
GPT-4	0.384	0.207	0.264

表 4.17: Mean squared error(DIT)

	O,T,X	O,T+X	O+T,X
ベースライン	0.216	0.188	0.283
GPT-3	0.198	0.180	0.151
GPT-3.5	0.212	0.189	0.160
GPT-4	0.194	0.162	0.232

IRS

表 4.18: 正解率, F 値 (IRS)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
ベースライン	0.485 (267/550)	0.484	0.673
GPT-3	0.520 (286/550)	0.581	0.774
GPT-3.5	0.536 (295/550)	0.563	0.743
GPT-4	0.698 (384/550)	0.700	0.845

表 4.19: JS divergence(IRS)

	O,T,X	O,T+X	O+T,X
ベースライン	0.424	0.284	0.279
GPT-3	0.382	0.231	0.250
GPT-3.5	0.374	0.237	0.233
GPT-4	0.273	0.157	0.169

表 4.20: Mean squared error(IRS)

	O,T,X	O,T+X	O+T,X
ベースライン	0.231	0.264	0.256
GPT-3	0.205	0.204	0.224
GPT-3.5	0.198	0.210	0.207
GPT-4	0.132	0.121	0.136

Live Competition IRS

表 4.21: 正解率, F 値 (Live Competition IRS)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
ベースライン	0.425 (102/240)	0.469	0.689
GPT-3	0.450 (108/240)	0.248	0.696
GPT-3.5	0.408 (98/240)	0.218	0.698
GPT-4	0.629 (151/240)	0.698	0.796

表 4.22: JS divergence(Live Competition IRS)

	O,T,X	O,T+X	O+T,X
ベースライン	0.465	0.296	0.324
GPT-3	0.427	0.267	0.218
GPT-3.5	0.432	0.270	0.234
GPT-4	0.316	0.192	0.185

表 4.23: Mean squared error(Live Competition IRS)

	O,T,X	O,T+X	O+T,X
ベースライン	0.253	0.273	0.301
GPT-3	0.236	0.239	0.188
GPT-3.5	0.240	0.243	0.206
GPT-4	0.155	0.157	0.148

Live Competition MMK

表 4.24: 正解率, F 値 (Live Competition MMK)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
ベースライン	0.821 (184/224)	0.000	0.278
GPT-3	0.616 (138/224)	0.080	0.435
GPT-3.5	0.746 (167/224)	0.000	0.492
GPT-4	0.844 (189/224)	0.250	0.286

表 4.25: JS divergence(Live Competition MMK)

	O,T,X	O,T+X	O+T,X
ベースライン	0.228	0.212	0.055
GPT-3	0.331	0.272	0.106
GPT-3.5	0.260	0.221	0.046
GPT-4	0.208	0.192	0.053

表 4.26: Mean squared error(Live Competition MMK)

	O,T,X	O,T+X	O+T,X
ベースライン	0.103	0.164	0.026
GPT-3	0.176	0.234	0.082
GPT-3.5	0.128	0.174	0.017
GPT-4	0.087	0.140	0.024

Live Competition MRK

表 4.27: 正解率, F 値 (Live Competition MRK)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
ベースライン	0.607 (136/224)	0.098	0.400
GPT-3	0.536 (120/224)	0.350	0.622
GPT-3.5	0.621 (139/224)	0.083	0.364
GPT-4	0.598 (134/224)	0.491	0.647

表 4.28: JS divergence(Live Competition MRK)

	O,T,X	O,T+X	O+T,X
ベースライン	0.346	0.290	0.137
GPT-3	0.367	0.268	0.115
GPT-3.5	0.320	0.278	0.090
GPT-4	0.332	0.242	0.119

表 4.29: Mean squared error(Live Competition MRK)

	O,T,X	O,T+X	O+T,X
ベースライン	0.172	0.252	0.104
GPT-3	0.195	0.227	0.079
GPT-3.5	0.157	0.238	0.051
GPT-4	0.166	0.195	0.084

Live Competition TRF

表 4.30: 正解率, F 値 (Live Competition TRF)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
ベースライン	0.695 (178/256)	0.292	0.397
GPT-3	0.723 (185/256)	0.063	0.474
GPT-3.5	0.746 (191/256)	0.245	0.268
GPT-4	0.734 (188/256)	0.507	0.690

表 4.31: JS divergence(Live Competition TRF)

	O,T,X	O,T+X	O+T,X
ベースライン	0.290	0.245	0.122
GPT-3	0.261	0.225	0.086
GPT-3.5	0.266	0.235	0.131
GPT-4	0.251	0.184	0.122

表 4.32: Mean squared error(Live Competition TRF)

	O,T,X	O,T+X	O+T,X
ベースライン	0.144	0.208	0.095
GPT-3	0.125	0.183	0.058
GPT-3.5	0.124	0.197	0.104
GPT-4	0.119	0.137	0.099

Live Competition ZNK

表 4.33: 正解率, F 値 (Live Competition ZNK)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
ベースライン	0.518 (116/224)	0.330	0.664
GPT-3	0.545 (122/224)	0.496	0.794
GPT-3.5	0.469 (105/224)	0.317	0.482
GPT-4	0.580 (130/224)	0.619	0.817

表 4.34: JS divergence(Live Competition ZNK)

	O,T,X	O,T+X	O+T,X
ベースライン	0.381	0.263	0.212
GPT-3	0.354	0.200	0.202
GPT-3.5	0.416	0.340	0.213
GPT-4	0.337	0.192	0.178

表 4.35: Mean squared error(Live Competition ZNK)

	O,T,X	O,T+X	O+T,X
ベースライン	0.197	0.229	0.183
GPT-3	0.183	0.157	0.172
GPT-3.5	0.218	0.315	0.184
GPT-4	0.172	0.149	0.146

4.7.3 データセット全体をまとめて集計した結果

データセットの結果をまとめて1つとしたときのラベル一致系統の評価結果を表 4.36 に示す.

表 4.36: 正解率, F 値 (マイクロ平均法)

	正解率	F 値 (O,T,X)	F 値 (O,T+X)
先行研究 (best)	0.5841	0.5387	0.7254
ベースライン	0.5312	0.4081	0.6564
GPT-3	0.5405	0.4394	0.7069
GPT-3.5	0.5490	0.3820	0.6366
GPT-4	0.6448	0.6125	0.7844

データセットの結果をまとめて1つとしたときの分布距離系統の評価結果を表 4.37 と表 4.38 に示す.

表 4.37: JS divergence(マイクロ平均法)

	O,T,X	O,T+X	O+T,X
先行研究 (best)	0.0947	0.0601	0.0612
ベースライン	0.3881	0.2641	0.2327
GPT-3	0.3683	0.2446	0.1815
GPT-3.5	0.3637	0.2606	0.1797
GPT-4	0.3048	0.1917	0.1664

表 4.38: Mean squared error(マイクロ平均法)

	O,T,X	O,T+X	O+T,X
先行研究 (best)	0.0463	0.0635	0.0507
ベースライン	0.2032	0.2313	0.2074
GPT-3	0.1944	0.2091	0.1512
GPT-3.5	0.1896	0.2275	0.1491
GPT-4	0.1498	0.1493	0.1353

4.7.4 評価指標ごとの上限値

正解率，適合率，再現率，F 値の上限値を表 4.39 と表 4.40 に示す.

表 4.39: 正解率，適合率 (X)，再現率 (X)，F 値 (X) の上限値

	正解率	適合率 (X)	再現率 (X)	F 値 (X)
先行研究 (best)	0.7884	0.7296	0.7328	0.7305
提案手法 (GPT-4)	0.8438	0.6513	0.7560	0.6997

表 4.40: 適合率 (T+X)，再現率 (T+X)，F 値 (T+X) の上限値

	適合率 (T+X)	再現率 (T+X)	F 値 (T+X)
先行研究 (best)	0.9028	0.7396	0.8128
提案手法 (GPT-4)	0.9104	0.8403	0.8453

JS divergence(JSD) と Mean squared error(MSE) の上限値を表 4.41 と表 4.42 に示す.

表 4.41: JS divergence の上限値

	JSD(O,T,X)	JSD(O,T+X)	JSD(O+T,X)
先行研究 (best)	0.0495	0.0254	0.0239
提案手法 (GPT-4)	0.2357	0.1566	0.0565

表 4.42: Mean squared error の上限値

	MSE(O,T,X)	MSE(O,T+X)	MSE(O+T,X)
先行研究 (best)	0.0196	0.0215	0.0142
提案手法 (GPT-4)	0.1061	0.1211	0.0279

第5章 考察

本章では、本研究で用いたデータセットを例に考察を行う。

5.1 節では、データセットごとの F 値の違いについて考えられる原因を記述する。5.2 節では、データセットごとのラベルの偏りの例を記述する。

5.1 データセットごとの F 値性能の調査

本研究では、GPT に対して過去の DBDC のデータセットを One-shot learning と Few-shot learning のデモンストレーション例とし、第 4 回目である DBDC4 の評価データに対してラベル付与を行った。その際、F 値の性能がデータセットによって大きく異なることが確認できた。その原因として、データセットごとのラベルの偏りが考えられる。そのため、正解ラベル ‘X’ の割合が少ないようなデータセットでは、F 値の振れ幅が大きくなっている。

5.2 データセットごとのラベルの偏り

表 4.3 を見ての通り、DBDC4 データセットでは ‘O’、‘T’、‘X’ のラベル分布が比較的散らばっていることが分かる。対して、表 4.4 の Live Competition MMK データセットや Live Competition TRF データセットではラベルの偏りが確認でき、Live Competition MMK データセットに関しては、正解率が最も高いが F 値が最も低い値であることが確認できる (表 4.24)。

そこで、Live Competition MMK データセットの一部を例とし、5.2.1 節と 5.2.2 節にそれぞれラベル付与が適切な例と不適切な例を記述する。

5.2.1 ラベル付与が適切な対話履歴

ラベル付与が適切に行われた例として、Live Competition MMKに含まれる対話履歴を以下に示す。

入力:

S0: 昭和からタイムスリップしてきた10才女子モモコです!! ピンク・レディーと沢田研二のファンです!! あなたは何年生まれですか? 何ってお呼びすれば良いですか? 私は1968年生まれのモモコです!!

U1: 私は未来からタイムスリップしてきた2100年生まれのタコです。

S2: 2100年生まれのタコさん!! 覚えた!! 突然だけどクイズ!! ヒントは私の好きな人です ピンク・レディーのデビューのきっかけとなった懐かしのTV番組といえは?

U3: 紅白歌合戦。

S4: 残念!! どんどん行こう!” なんて世間はあなたを気にする” これ、沢田研二のなんていう曲かわかる?

U5: ブーメランかな。

S6: 難しかったかな? どんどん行こう! 沢田研二が「ボギー ボギー」って歌ってたけど、誰のことか知ってる?

U7: それは野口五郎だね。

S8: 難しかったかな? どんどん行こう!” 花園で眠れぬこともあるんだよ” これ、沢田研二のなんていう曲かわかる?

U9: 楽園でどうだ!

S10: 難しかったかな? どんどん行こう! おっ!? このジャンルだと正解0問、不正解4問だね 次のジャンルに進む?

U11: 別のジャンルで頼む。

S12: ”ある時真面目な医学生” これ、ピンク・レディーのなんていう曲かわかる?

U13: UFOじゃないかな。

S14: 惜しい〜 どんどん行こう! ピンク・レディーの曲『Last Pretender』を作曲した元 YMO のメンバと言えは?

U15: 小室哲哉じゃない。

S16: 残念!! どんどん行こう! ”寝顔にキスでもしてあげたいけど/そしたら一日旅立ちが延びるだろう” これ、沢田研二のなんていう曲かわかる?

U17: 寂しい一人旅じゃないかな。

S18: 難しかったかな? どんどん行こう! ピンク・レディーの曲で、「指名手配」を意味する曲名は?

U19: 今度こそ UFO でしょ。

S20: 難しかったかな? どんどん行こう! おっ!? このジャンルだと正解 0 問、不正解 8 問だね 次のジャンルに進む?

U21: 次に進んでください。

S22: ”しばらくお色気さようなら” これ、ピンク・レディーのなんていう曲かわかる?

U23: 私は 18 歳の女の子。

S24: 惜しい～ どんどん行こう! ”小指に食い込む指輪を見つめ” これ、沢田研二のなんていう曲かわかる?

U25: ONE LOVE で当たってくれ。

S26: 難しかったかな? どんどん行こう! ピンク・レディーの素人時代のデュオの名前は? ヒント: 食べ物の名前。ただしキャンディーズじゃないよ

U27: イルカ。

S28: 残念!! どんどん行こう! 沢田研二のニックネームと言えは?

U29: ジュリー。

S30: 当たり～ 頭いい～ おっ!? このジャンルだと正解 1 問、不正解 11 問だね 次のジャンルに進む

出力:

S0:O S2:O S4:O S6:O S8:O S10:O S12:O S14:O S16:O S18:O S20:O S22:O S24:O
S26:O S28:O S30:O

このデータセットはクイズ対話を行うシステムから収集されており、システムがクイズを出題し、ユーザがそれに対して答えるといった形式のデータセットとなっている。上記の例では、システムに対しユーザ側がシステム側の想定した回答を続けており、破綻はこの対話では起こっていないとアノテータが判断している。GPT-4 に対してこの対話履歴にラベル付与を行った結果、全ての発話箇所にラベル‘O’を出力しており、この対話では破綻が起こっていないと判断している。

5.2.2 ラベル付与が不適切な対話履歴

Live Competition MMK に含まれる別の対話履歴を示す.

入力:

S0: 昭和からタイムスリップしてきた 10 才女子モモコです!! ピンク・レディーと沢田研二のファンです!! あなたは何年生まれですか? 何ってお呼びすれば良いですか? 私は 1968 年生まれのモモコです!!

U1: 私は 1988 年生まれのららです!

S2: んん?? よく分かんなかったから、yamada さんって呼ぶことにするね。決定!! 突然だけどクイズ!! ヒントは私の好きな人です ”ある時真面目な医学生” これ、ピンク・レディーのなんていう曲かわかる?

U3: UFO という曲ですか?

S4: 残念!! どんどん行こう! ピンク・レディーの素人時代のデュオの名前は? ヒント: 食べ物の名前。ただしキャンディーズじゃないよ

U5: んー、考えたのですが分らないです。

S6: 残念!! どんどん行こう! ”年上の女 美し過ぎる” これ、沢田研二のなんていう曲かわかる?

U7: 沢田研二って誰ですか?

S8: 難しかったかな? どんどん行こう! ピンク・レディーの曲『Last Pretender』を作曲した元 YMO のメンバと言えよ?

U9: モモコさんは歌手にとっても詳しいんですね! 歌手以外に好きなものはありますか?

S10: 惜しい~ どんどん行こう! おっ!? このジャンルだと正解 0 問、不正解 4 問だね 次のジャンルに進む?

U11: 次のジャンルに進みます!

S12: ”ある時アラブの大富豪” これ、ピンク・レディーのなんていう曲かわかる?

U13: ピンクレディーについて詳しくないので、分らないです。

S14: 難しかったかな? どんどん行こう! ”空っぽよ心はうつろよ何もないわ” これ、ピンク・レディーのなんていう曲かわかる?

U15: 分かりません。

S16: 残念!! どんどん行こう! 三浦百恵 (=山口百恵) が作詞、沢田研二が作曲した
アン・ルイスのヒット曲といえば?

U17: 歌についてはよくわからないんです。他のジャンルに変更しませんか?

S18: 残念!! どんどん行こう! ”小指に食い込む指輪を見つめ” これ、沢田研二
のなんていう曲かわかる?

U19: 沢田研二の曲も全く知らないんです。

S20: 残念!! どんどん行こう! おっ!? このジャンルだと正解0問、不正解8問だ
ね 次のジャンルに進む?

U21: 今度こそ次のジャンルに進みたいです!

S22: ピンク・レディーのデビューシングルといえば?

U23: またピンクレディーですか? 私はその世代ではないんです。

S24: 残念!! どんどん行こう! 沢田研二の「TOKIO」の作詞者は?

U25: すみません、全く分かりません。

S26: 惜しい～ どんどん行こう! ”フラミンゴみたいひょいと一本足で” これ、
ピンク・レディーのなんていう曲かわかる?

U27: ピンクレディと沢田研二の問題がたくさんあるんですね!

S28: 惜しい～ どんどん行こう! 志村けんが沢田研二の曲名とピンクレディーの
曲名を混ぜ合わせたギャグをやったけど、その名前は?

U29: 志村けんって懐かしいなー。

S30: 惜しい～ どんどん行こう! おっ!? このジャンルだと正解0問、不正解12問
だね 次のジャンルに進む?

出力:

S0:O S2:O S4:O S6:O S8:O S10:T S12:O S14:O S16:O S18:T S20:T S22:O S24:T
S26:T S28:X S30:T

この対話システムは前述の通り、ユーザ側がクイズの回答を行っているという前提で対話が行われている。つまり、あらかじめ決められた型に従って出力を行うパターンマッチング型の対話システムとなっている。そのため、ユーザの回答内容として参照するのは正解となる文字列が含まれているかだけであり、その後の対話にはユーザの情報は用いられていない。

U9やU17のユーザ発話箇所を見ての通り、ユーザ側がシステムに対し雑談対話を試みていることがわかる。それに対し、システム側の発話箇所であるS10やS18では、直前のユーザの発話内容を考慮せずシステムが次のクイズを出題している。2.1節でも記述したように、対話がスムーズに行えない場合や続けることが困難だと判断される発話内容は、対話システムの仕様に関係なくアノテータによってラベル‘T’や‘X’が付与される。そのためこの対話履歴では、ラベル‘T’と‘X’の付与が難しい発話が多くなってしまい、結果としてGPT-4のラベル付与精度も下がってしまった。

以下にGPT-4が同様の対話に対してラベル付与を行った結果を示す。

出力:

S0:O S2:T S4:O S6:O S8:O S10:T S12:O S14:O S16:O S18:T S20:T S22:T S24:T
S26:T S28:T S30:T

第6章 おわりに

近年、対話システムや対話を機能の1つとして取り入れているサービスが増えている。このような対話システムが出力する文そのものは、人間が話すように自然で流暢なものとなってきている。しかし、対話システムと人間の対話において、文脈の整合性が保てない、間違った情報を出力するなどシステム側が論理的に破綻している内容を出力する場合がある。

そこで本研究では、LLMの1つであるGPTの汎化性能に着目した。この言語モデルは、推論時に1つもしくは少量のタスクのデモンストレーションを提示するOne-shot learningとFew-shot learningを適用することによって性能向上が期待できる。このことから、対話履歴とラベル付与例をOne-shotとFew-shotの提示例としてGPTにプロンプトする手法を提案した。

実験では、先行研究のデータを対象に先行研究の結果と提案手法の比較を行った。評価指標はラベルの一致率と分布距離の2つがあり、ラベルの一致率は先行研究と比べ提案手法が勝ったが、分布距離の評価指標では劣るということが分かった。

今後の課題を以下にまとめる。

- ラベルの基準設定の見直し
- 新たなデータセットとしてGPTとの対話履歴を学習・評価データとする
- 提示例として与えるデモンストレーションの選定及びプロンプトの改善

謝辞

本研究を進めるにあたり, 研究の説明や論文の書き方など様々なご指導を頂きました鳥取大学工学部電気情報系学科自然言語処理研究室の村田真樹教授に心から御礼申し上げます. また, 本研究を進めるにあたり, 御指導, 御助言を頂きました, 村上仁一准教授に心から御礼申し上げます. また, 同じ班に所属されていた自然言語処理研究室の皆様へ心から感謝の気持ちと御礼を申し上げたく謝辞にかえさせていただきます.

参考文献

- [1] 東中竜一郎, 船越孝太郎, 小林優佳, 稲葉通将. 対話破綻検出チャレンジ. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 75 回 (2015/10), pp. 27–32. 一般社団法人 人工知能学会, 2015.
- [2] 東中竜一郎, 船越孝太郎, 稲葉通将, 荒瀬由紀, 角森唯子. 対話破綻検出チャレンジ 2. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 78, pp. 64–69, 2016.
- [3] R. Higashinaka, K. Funakoshi, M. Inaba, Y. Tsunomori, T. Takahashi, and N. Kaji. Overview of dialogue breakdown detection challenge 3. *Dialog System Technology Challenges Workshop (DSTC6)*, 2017.
- [4] Ryuichiro Higashinaka, Luis FD ' Haro, Bayan Abu Shawar, Rafael E Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, João Sedoc. Overview of the dialogue breakdown detection challenge 4. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pp. 403–417. Springer, 2021.
- [5] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *preprint*, 2018.
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.
- [7] Openai api. <https://platform.openai.com/docs/guides/text-generation>.
- [8] Chatgpt. <https://chat.openai.com/>.

- [9] 東中竜一郎, 船越孝太郎. Project next nlp 対話タスクにおける雑談対話データの収集と対話破綻アノテーション. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 72 回 (2014/12), p. 08. 一般社団法人 人工知能学会, 2014.