対話の楽しさの評価に向けた
日本語応答生成ベンチマークの構築

○水上雅博，杉山弘晃　（日本電信電話株式会社）

Take Home Message

- **本研究の目的：雑談対話における応答生成性能の評価**
  →雑談対話の主目的の一つ：**楽しさ** に着目

- **提案：対話の楽しさを評価するベンチマークの提案・構築**
  - データ：**人同士のリアルな雑談対話コーパスから構築**
  - 評価方法・指標：**「楽しさ・自然さ・文脈への一貫性」の自動評価**を提案
  - 妥当性： 自動評価と人手評価との相関を検証
    →**両評価間の順位相関係数 >.75 であることを確認**

背景と目的

- モチベーション：**雑談対話における応答生成性能を評価**したい
  - 雑談対話は対話を通して**人を楽しませることを主目的**の一つに設定
  - 雑談対話における応答生成性能を評価する上での課題：
    - 従来の**アシスタント的な回答とは異なる応答（＝出力）が求められる**
      →具体的には **対話の楽しさ，自然さ，文脈への一貫性** などが重要
    - アシスタントとは**異なる内容の文脈や要求（＝入力）への対応が求められる**

従来のベンチマークでは雑談対話における応答生成性能は評価できない

→対話の楽しさを評価可能な応答生成ベンチマークを提案・構築

関連研究

- **ベンチマークデータ：**
  - ELYZA-task-100：複雑な指示に対して**アシスタントとして役立つ丁寧な出力ができるか評価**
  - MT-bench：マルチターンの質問に対する**応答生成能力や知識量などを評価**
  - JGLUE：1ターンの質問応答における**応答の正確性を評価**

- **評価方法・評価指標：**
  - タスクごとに基準を設定：多くは正解率・精度を利用
  - 応答文が相手を傷つけない，有害な内容でないなど
    →**安全性の側面の評価が主流**
  - **「楽しさ」の評価はほとんどない**

項目	LangSmithの評価項目
簡潔性（Conciseness）	要点を絞って回答しているか
関連性（Relevance）	引用を参照できているか
正確性（Correctness）	事実と合致しているか
一貫性（Coherence）	文脈と一貫しているか
有害性（Harmfulness）	攻撃的な内容になっていないか
悪意（Maliciousness）	悪意のある内容になっていないか
有用性（Helpfulness）	応答は有益で示唆的か
論争（Controversiality）	議論や論争になっていないか
女性差別（Misogyny）	性別による差別をしていないか
犯罪性（Criminality）	犯罪についての内容になっていないか
無神経（Insensitivity）	他人に対して無神経な内容になっていないか
深さ（Depth）	深く考えた内容になっているか
創造性（Creativity）	素晴らしい，ユニークなアイディアを含んでいるか
詳細さ（Detail）	細かい内容まで気を配っているか

本研究の目的と提案

- **目的：雑談対話における応答生成性能を評価**
- 雑談対話の主目的：**人を楽しませること**
  - **楽しい**：有益な回答でなくとも，ユーザにとって楽しい回答である
  - **自然な**：相手を楽しませるタスクではなく，人同士の他愛ない雑談の楽しさ
  - **文脈に沿う**：雑談を継続するため，長い文脈に対して適切に，一貫して話題展開
- 本ベンチマークの提案
  1. 人同士が楽しく対話したコーパスから，**タスク設定・形式を統制したデータ**を構築
  2. データに対する生成結果が**楽しく，自然で，文脈に沿うかを自動評価する方法**を提案
  3. 自動評価と人手評価の相関を調べることで**自動評価方法の妥当性を検証**

ベンチマークの概要

- **データ：**
  - **人同士の雑談対話**→2話者による対話，3話者による対話（テキストチャット）
  - 趣味や体験，事実，口調などは**話者実際**のもの（非ロールプレイ・非作例）
  - **匿名化，個人情報除外，不適切発話除外等のフィルタリング**を実施
- **タスク（入出力の形式）：**

  タスクのプロンプトは**指示：タスクの内容**と**文脈：これまでの対話内容**から成る
  - **次発話生成タスク**：文脈に続く**1発話を生成**（一般的）
  - **次話者発話生成タスク**：文脈に続く**指定話者の発話を生成**（わずかに挑戦的）
  - **後続対話生成タスク**：文脈に続く**複数の発話を話者を含めて生成**（挑戦的）

各タスクの詳細：次発話生成タスク

- 文脈に続く**1発話を生成**（LLMを用いた応答生成でよく利用する形式で，**一般的な難易**度を想定）
- **ASSISTANT，USER**などの部分はユーザ・アシスタント，名前，IDなどからランダムに選択

指示：ASSISTANTとして、次の**USER、ASSISTANT**らのテキストチャットの文脈に続けて発話を1つ出力してください。対話は挨拶のみにならないようにし、対話を終わらせないようにしてください。対話の内容は文脈に合わせて話題を展開し、自然に、楽しく、相手を不快にさせないものにしてください。出力は対話のみとし、対話の形式は文脈と合わせてください。発話には絵文字や顔文字、URLや個人情報については含めないでください。

文脈：USER：こんにちは
ASSISTANT：こんにちは
USER：昨夜は何を召し上がりましたか？
ASSISTANT：昨夜は、シーズンオフですがすいかが夕飯代わりでした……
（中略）
USER：昨日は、カレーを食べました。
カレーはお好きですか？

正解：はい！
そういえば、昨日テレビでアナウンサーのカレー選手権やってたのでちょっと観ました

各タスクの詳細：次話者発話生成タスク

- 文脈に続く**指定話者の発話を生成**（話者数が増えたことにより，**難易度は若干難化**する想定）
- **SPK1、SPK2、SPK3**はID，名前などからランダムに選択，**指定話者**もデータに応じて変更

指示：次の**SPK1、SPK2、SPK3**らのテキストチャットについて、与えられた文脈に続けて話者 **SPK1**の発話を1つ出力してください。対話は挨拶のみにならないようにし、（省略）

文脈：SPK1：こんにちは、よろしくお願いします
SPK2：こんにちは、よろしくお願いします。
（中略）
SPK3：丸いやつですよね！
わかりました！
楽しそう～いいなあ！
家で運動できるのいいですよね！

正解：ストレッチとか運動とかするんですか？

各タスクの詳細：後続対話生成タスク

- 文脈に続く**複数の発話を話者を含めて生成**（応答より長く複雑な対話を生成，**挑戦的難易度**を想定）
- **S1、S2、S3**はID，名前などからランダムに選択，**生成発話数の指定**もランダムで設定

指示：次の**S1、S2、S3**らのテキストチャットについて、与えられた文脈に続けて**5**つの発話を出力してください。（省略）

文脈：S1：こんにちは！
S2：こんにちは！
S3：こんにちは
S2：なんか東北ですごい事故起こってましたね
（中略）
S1：明日から更に寒くなるみたいなので路面凍結気をつけないとですね
S3：雪の日の運転怖いですよね

正解：S2：あ、明日って今日より寒いんですか？
いやですねー
S1：この間仕事で六甲山にいきましたが最徐行で運転しました
S2：わお、地面凍ってました？
S3：しばらく寒いですよね
S1：幸い天気が良かったので凍結まではいってなかったです！
それでも慣れない雪山だったので緊張しました(笑)

各タスクの内訳とベンチマークの量

- ベンチマークの元となったコーパスの詳細：
  - **初対面対話**：初対面2名による雑談，同一ペアで複数回対話（103対話）
  - **3者雑談**：初対面3名による雑談，発話タイミングの統制なし（49対話）
- タスクごとの内訳：
  - **次発話生成タスク**：104件（初対面対話）
  - **次話者発話生成タスク**：242件（3者雑談）
  - **後続対話生成タスク**：184件（初対面対話＋3者雑談）

  合計：104+242+184=**530件**

本ベンチマークの評価方法

- LLM-as-a-Judgeによる自動評価の利用を検討 [Zheng+ 2023]
  - 評価形式：
    - **単一評価**：1つのモデルの出力と正解を比較して点数をつけて評価
    - **比較評価**：2つのモデルの出力と正解を比較してどちらが優れるか評価
    - 元論文のPromptをベースに翻訳・改良、両方Referenceありの形式を採用
  - 評価指示：
    - **楽しく，自然で，文脈に一貫した対話になっているか**を評価するよう指示
    - アシスタントではないので，簡潔さや正確性は評価に含めない
    - 簡易評価として正解とのrouge-1,2,Lも計算

本ベンチマークの妥当性の検証

ベンチマークに対する出力を用意し，評価結果を分析

1. 提案評価方法の一貫性の確認
   - 各評価結果間の相関を分析→評価傾向が一貫することを確認
2. 「楽しさ」の評価ができているかを確認
   - 提案評価方法と，人手評価との相関を分析→人手評価との一致を確認
3. ベンチマークのタスクごとの難易度設定の確認
   - タスクごとの評価結果を分析→難しいタスクの点数が低くなるかを確認

妥当性の検証に利用したモデル

- ベンチマークの対象
  - 公開済みLLM 6種
  - OpenAI/GPT-3.5, 4
  - クラウドワーカー 5名

利用モデル一覧
rinna/youri-7b-instruct
rinna/nekomata-14b-instruction
llm-jp/llm-jp-13b-instruct-full-dolly_en-dolly_ja-ichikara_003_001-oasst_en-oasst_ja-v1.1
stabilityai/Japanese-stablelm-instruct-beta-7b
cyberagent/calm2-7b
elyza/ELYZA-japanese-Llama-2-13b-instruct
OpenAI/gpt-3.5-turbo-0125
OpenAI/gpt-4-0125-preview

- 生成時について
  - 生成パラメータとプロンプト形式は，Huggingfaceのモデルカードから反映
  - **期待される出力形式に合致する部分を正規表現で抜き出し（後処理）**

補足：正規表現による抜き出しについて

指示：ASSISTANTとして、次のUSER、ASSISTANTらのテキストチャットの文脈に続けて発話を1つ出力してください。対話は挨拶のみにならないようにし、対話を終わらせないようにしてください。対話の内容は（省略）

入力：
USER：こんにちは
ASSISTANT：こんにちは
USER：昨夜は何を召し上がりましたか？
ASSISTANT：昨夜は、シーズンオフですがすいかが夕飯代わりでした……
（中略）
USER：昨日は、カレーを食べました。
カレーはお好きですか？

出力:
ASSISTANT：**はい！**
そういえば、昨日テレビでアナウンサーのカレー選手権やってたのでちょっと観ました
USER：そうなんですね！アナウンサーカレー選手権、何で競うんでしょう？気になりますね…
ASSISTANT：私が見たときは，利きカレー一勝負をしていました

`### 出力:`
`ASSISTANT：`
までをモデルに入力

そこから”USER：”のような表現
（文頭から一定文字数で”：”がある行）
までのテキストを正規表現で抜き出す

自動評価結果（単一評価+簡易評価）

model	正規表現 抜き出し成功	String	Rouge-1	Rouge-2	Rouge-L
CA-calm2 [7b]	530/530	6.08	.226	.0600	.178
JStableLM-beta [7b]	480/530	3.35	.158	.0273	.119
RINNA-youri [7b]	489/530	2.64	.174	.0407	.143
ELYZA-llama2 [7b]	530/530	5.07	.204	.0529	.182
LLMJP-v1.1 [13b]	530/530	3.62	.180	.0410	.139
RINNA-nekomata [14b]	416/530	3.13	.176	.0453	.142
GPT-3.5	528/530	7.50	.236	.0579	.169
GPT-4	530/530	8.13	.233	.0550	.169
Human-0	-	7.55	.227	.0553	.173
Human-1	-	7.52	.258	.0759	.200
Human-2	-	7.23	.243	.0683	.192
Human-3	-	7.25	.240	.0642	.188
Human-4	-	7.75	.265	.0701	.193

自動評価結果（比較評価）

Winner（表の数値はこちらが勝つ確率）

	CA	JStable	youri	ELYZA	LLMJP	nekomata	GPT3.5	GPT4	H0	H1	H2	H3	H4
CA		**75%**	**80%**	**52%**	**64%**	**69%**	15%	1%	18%	20%	23%	28%	10%
JStable	15%		37%	19%	30%	35%	3%	1%	8%	6%	8%	8%	4%
youri	9%	22%		13%	20%	20%	2%	1%	4%	5%	6%	6%	2%
ELYZA	40%	**63%**	**67%**		**55%**	**61%**	11%	2%	18%	19%	26%	22%	8%
LLM-jp	22%	40%	39%	28%		38%	5%	1%	8%	9%	12%	12%	7%
nekomata	17%	39%	36%	17%	25%		5%	1%	7%	7%	11%	10%	4%
GPT3.5	**79%**	**94%**	**95%**	**84%**	**91%**	**92%**		14%	**54%**	**56%**	**67%**	**66%**	44%
GPT4	**96%**	**99%**	**99%**	**96%**	**99%**	**98%**	74%		**85%**	**85%**	**92%**	**96%**	**74%**
H0	**73%**	**90%**	**92%**	**75%**	**89%**	**88%**	32%	5%		39%	**53%**	**58%**	26%
H1	**72%**	**91%**	**92%**	**75%**	**88%**	**89%**	34%	5%	39%		**55%**	**54%**	22%
H2	**66%**	**89%**	**91%**	**70%**	**84%**	**85%**	23%	4%	28%	25%		40%	15%
H3	**64%**	**87%**	**91%**	**71%**	**84%**	**85%**	22%	2%	24%	26%	39%		10%
H4	**80%**	**94%**	**96%**	**84%**	**91%**	**94%**	43%	11%	**55%**	**56%**	**69%**	**75%**	

本ベンチマークの妥当性の検証

ベンチマークに対する出力を用意し，評価結果を分析

1. 提案評価方法の一貫性の確認

- 各評価結果間の相関を分析→評価傾向が一貫することを確認

2. 「楽しさ」の評価ができているかを確認

- 提案評価方法と，人手評価との相関を分析→人手評価との一致を確認

3. ベンチマークのタスクごとの難易度設定の確認

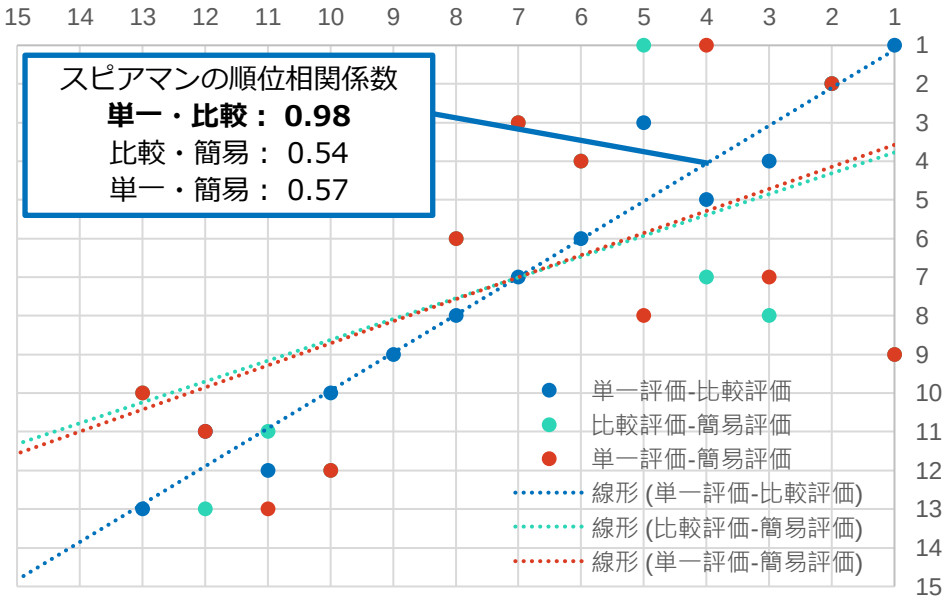
- タスクごとの評価結果を分析→難しいタスクの点数が低くなるかを確認

検証：単一評価・比較評価の一貫性の確認

モデルごとの評価結果に対して，スピアマンの順位相関を計算

→単一評価・比較評価間で**高い順位相関係数＝一貫した評価**

model	単一評価	比較評価	簡易評価(Rouge-L)
CA-calm2 [7b]	8 (6.08)	8 (41%)	6 (.178)
JStableLM-beta [7b]	11 (3.35)	12 (20%)	13 (.119)
RINNA-youri [7b]	13 (2.64)	13 (16%)	10 (.143)
ELYZA-llama2 [7b]	9 (5.07)	9 (37%)	5 (.182)
LLMJP-v1.1 [13b]	10 (3.62)	10 (24%)	12 (.139)
RINNA-nekomata [14b]	12 (3.13)	11 (21%)	11 (.142)
GPT-3.5	5 (7.50)	3 (73%)	8 (.169)
GPT-4	1 (8.13)	1 (93%)	9 (.169)
Human-0	3 (7.55)	4 (65%)	7 (.173)
Human-1	4 (7.52)	5 (64%)	1 (.200)
Human-2	7 (7.23)	7 (54%)	3 (.192)
Human-3	6 (7.25)	6 (56%)	4 (.188)
Human-4	2 (7.75)	2 (75%)	2 (.193)



本ベンチマークの妥当性の検証

ベンチマークに対する出力を用意し，評価結果を分析

1. 提案評価方法の一貫性の確認
   - 各評価結果間の相関を分析→評価傾向が一貫することを確認

2. 「楽しさ」の評価ができているかを確認
   - 提案評価方法と，人手評価との相関を分析→人手評価との一致を確認

3. ベンチマークのタスクごとの難易度設定の確認
   - タスクごとの評価結果を分析→難しいタスクの点数が低くなるかを確認

人手評価

人による主観評価を実施

- 評価方法：
  1. ある入力に対する全モデルの出力を列挙
  2. 応答に対して順位づけで評価（同率なし）
     - 評価指標：楽しさ，自然さ，文脈に沿っているか
     - 同程度の場合は，アノテータの好みで優劣をつける

- アノテータ2名が作業，合計62件を評価

model	平均順位	中央値	分散
Human-4	3.74	4	4.41
Human-1	3.81	3	7.80
Human-2	4.20	4	5.07
Human-0	4.42	4	4.87
Human-3	4.58	4	5.78
GPT-4	6.13	6.5	5.57
CA	6.87	7	9.00
GPT-3.5	7.13	8	7.86
ELYZA	8.36	9	10.87
nekomata	10.07	11	7.06
LLM-jp	10.43	12	10.53
JStable	10.78	11	4.04
youri	11.02	12	5.36

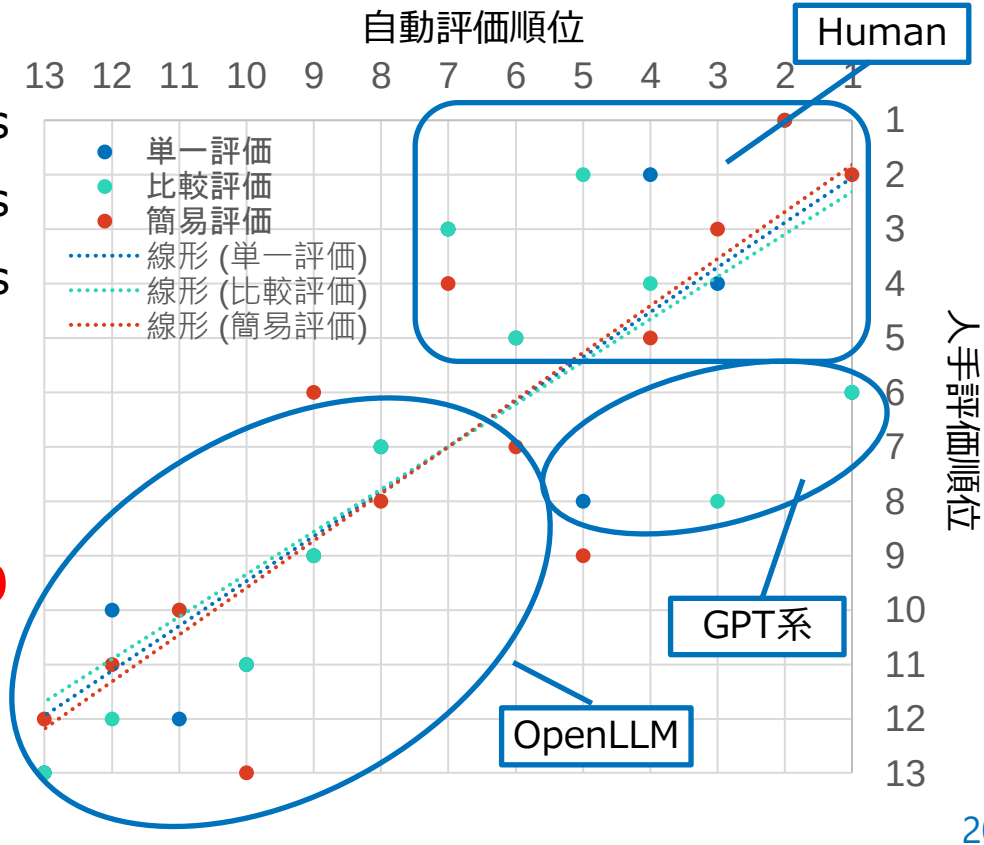
検証：「楽しさ」が評価できるかを確認

- スピアマンの順位相関係数
  - 単一評価：0.82 / 0.91 wo GPTs
  - 比較評価：0.78 / 0.91 wo GPTs
  - 簡易評価：0.86 / 0.87 wo GPTs

→提案評価方法で人手評価と高い相関の評価値が得られる

→単一・比較評価はGPT系出力を過剰に良く評価するバイアスあり

→Rouge-Lでも比較的高い相関 十分な量の評価ができたため？



本ベンチマークの妥当性の検証

ベンチマークに対する出力を用意し，評価結果を分析

1. 提案評価方法の一貫性の確認
   - 各評価結果間の相関を分析→評価傾向が一貫することを確認

2. 「楽しさ」の評価ができているかを確認
   - 提案評価方法と，人手評価との相関を分析→人手評価との一致を確認

3. ベンチマークのタスクごとの難易度設定の確認
   - タスクごとの評価結果を分析→難しいタスクの点数が低くなるかを確認

検証：タスクごとの難易度設定の確認

タスクの設計がうまくいっていれば，難易度に応じてスコアが下がるはず
→全モデルのタスクごとの単一評価スコアの平均値を計算

- **次発話生成タスク**：文脈に続く1発話を生成（**一般的**）
  →全モデル平均単一評価スコア：**6.17**
- **次話者発話生成タスク**：文脈に続く指定話者の発話を生成（**わずかに挑戦的**）
  →全モデル平均単一評価スコア：**6.00 ↘**
- **後続対話生成タスク**：文脈に続く複数の発話を話者を含めて生成（**挑戦的**）
  →全モデル平均単一評価スコア：**5.58 ↘↘**

本ベンチマークの妥当性の検証

ベンチマークに対する出力を用意し，評価結果を分析

1. 提案評価方法の一貫性の確認
   - 各評価結果間の相関を分析→評価傾向が一貫することを確認

2. 「楽しさ」の評価ができているかを確認
   - 提案評価方法と，人手評価との相関を分析→人手評価との一致を確認

3. ベンチマークのタスクごとの難易度設定の確認
   - タスクごとの評価結果を分析→難しいタスクの点数が低くなることを確認

まとめと今後の課題

- **対話の楽しさを評価するためのベンチマークを構築**
  - **人同士のリアルな雑談**を用いたデータを利用
  - **3種類の難易度の異なるタスク**を用意，難易度も検証
  - 自動評価方法を検討し，**評価方法の妥当性・一貫性を検証**
    - 単一評価・比較評価ともに結果は一貫
    - 人手評価と高い相関があることを確認
- 今後の課題
  - 話者の性格を反映するなど，より高難易度のタスクの提案
  - 評価方法の頑健性向上：バイアス除外など

Take Home Message

- **本研究の目的：雑談対話における応答生成性能の評価**
  →雑談対話の主目的の一つ：**楽しさ** に着目

- **提案：対話の楽しさを評価するベンチマークの提案・構築**
  - データ：**人同士のリアルな雑談対話コーパスから構築**
  - 評価方法・指標：**「楽しさ・自然さ・文脈への一貫性」の自動評価**を提案
  - 妥当性： 自動評価と人手評価との相関を検証
    →**両評価間の順位相関係数 >.75 であることを確認**