B-3

事実正誤判定が不要な文の検出に向けたデータセットの収集と分析

亀井遼平[1], 塩野大輝[1], 赤間怜奈[1,2], 鈴木潤[1,2]

[1]東北大学, [2]理化学研究所

言語処理学会第30回年次大会(NLP2024) @神戸国際会議場, 2024/3/13

Fai Lab

TOHOKU NLP LAB

与えられた知識に基づいていることが良い対話の条件か？

- LLMの出力の事実性の担保が課題、知識に基づく対話応答への取り組み
- FaithDial[Dziri+, 2022]
  - Wizard of Wikipedia[Dinan+, 2018]に応答中に知識に含まれない情報が含まれていれば幻覚ラベル付与 書き換え を実施

history	knowledge	original_response		response
["I love candy, what's a good brand?"]	Dylan's Candy Bar is a chain of boutique candy shops …	Dylan's Candy Bar is a great brand of candy Hallucination	書き換え →	I don't know how good they are, but Dylan's Candy Bar has a chain of candy shops in various cities.

与えられた知識に基づいていることが良い対話の条件か？

- LLMの出力の事実性の担保が課題，知識に基づく対話応答への取り組み
- FaithDial[Dziri+, 2022]
  - Wizard of Wikipedia[Dinan+, 2018]に応答中に知識に含まれない情報が含まれていれば幻覚ラベル付与 書き換え を実施

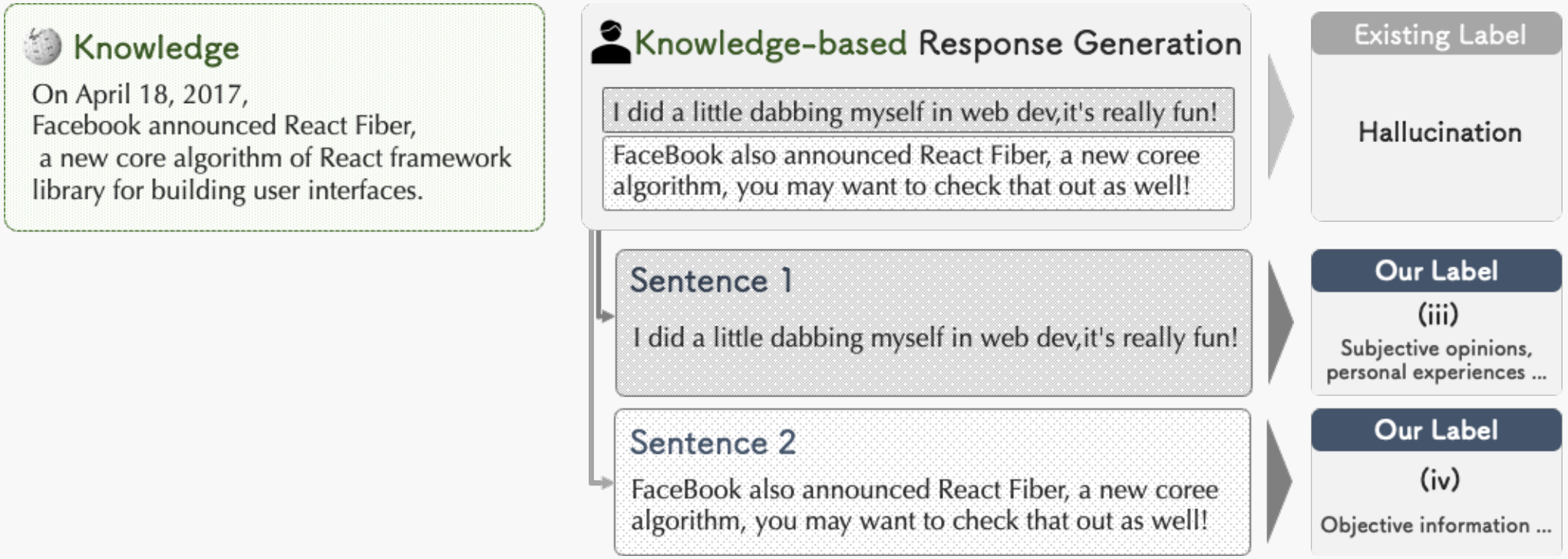
history	knowledge	original_response		response
["I love candy, what's a good brand?"]	Dylan's Candy Bar is a chain of boutique candy shops …	Dylan's Candy Bar is a great brand of candy 事実正誤判定不要	**書き換え** →	I don't know how good they are, but Dylan's Candy Bar has a chain of candy shops in various cities.

- 魅力的な対話のためには，与えられた知識の提供だけでなく，共感や提案，意見の主張も必要 [Huang+, 2020]

  →事実性の判定の前に，そもそも事実正誤判定が必要・不要を判定する

本研究の説明

- 対話における事実正誤判定が不要な応答を検出するタスクの提案
  - 学習・評価データセット（Dialogue Dataset annotated with Fact-Check-needed label, DDFC)の作成
  - Llama 2のベースラインで，約88%の正解率で検出可能であることを確認



DDFC データセットの構築の準備

- 目標
  - 応答中の各文にラベルを振り，事実正誤判定が不要な文と必要な文を区別

- アノテーションをする基となるデータセット
  - FaithDialのoriginal_responseを文単位に分割

- 文タイプのラベルの種類
  - (i)同意/不同意・相槌
  - (ii)提案・アドバイス
  - (iii)主観的な意見・個人的な経験/考え/感情
  - (iv)客観的な情報

DDFC データセットの構築の準備

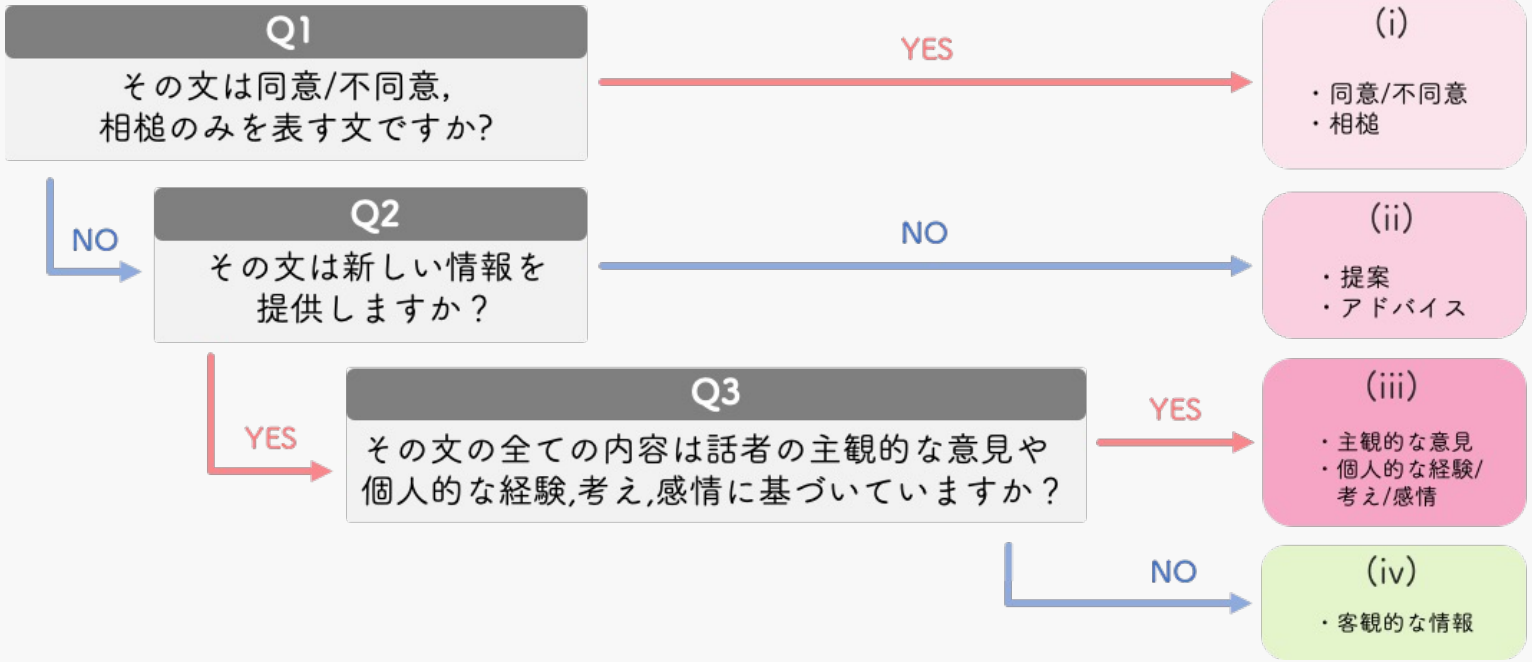
- 目標
  - 応答中の各文にラベルを振り，事実正誤判定が不要な文と必要な文を区別

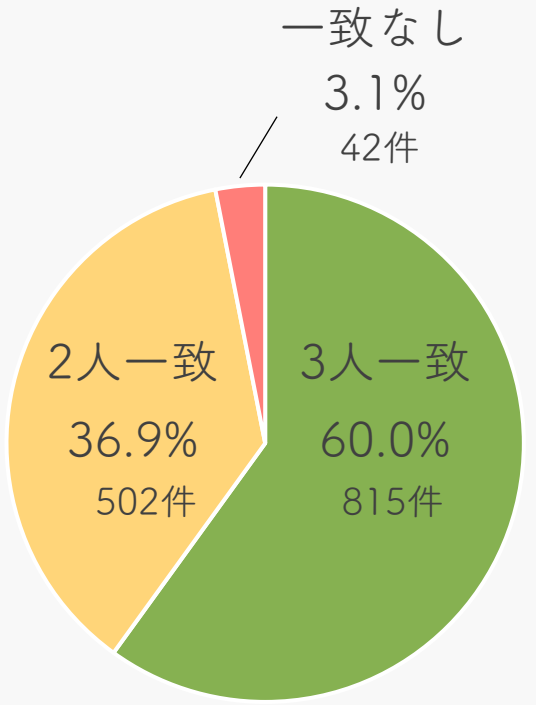
- アノテーションをする基となるデータセット
  - FaithDialのoriginal_responseを文単位に分割

- 文タイプのラベルの種類
  - (i)同意/不同意・相槌
  - (ii)提案・アドバイス → 事実性判定が不要
  - (iii)主観的な意見・個人的な経験/考え/感情
  - (iv)客観的な情報 → 事実性判定が必要

クラウドソーシングによるアノテーション

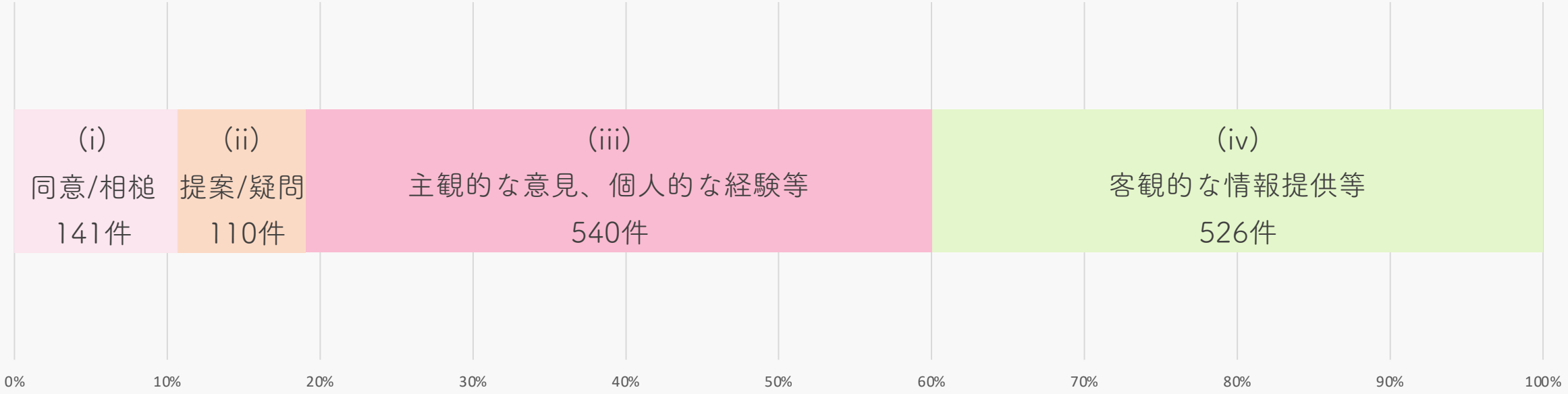
- ワーカのタスク：与えられた1文についてのYES/NO質問に最大3問答える
  - 1359文にアノテーション、1文あたり3人のワーカを割り当て
  - 3人のワーカのラベル一致なしの文はデータセットから除外





DDFC のラベル分析

- FaithDialでは71.4%の応答がHallucinationという判定
- DDFCでは約60%が"事実正誤判定が不要"な応答
  - (iii)主観的な意見、個人的な経験等 が多いのは対話の特徴．WoW作成時の「関連する知識を楽しく魅力的な方法で提示すること」という指示が影響しているか



実験1：事実正誤判定が不要な文を検出できるか

- タスク
  - 事実性の判定が必要になるか否かの二値分類
    - 入力：1文単位の対話応答
    - 出力：事実正誤判定が不要かどうかの予測
- 学習・評価データセット
  - 計1317件を、1,000件の学習データと317件の評価データに分割
- 使用したモデル
  - GPT-3.5, GPT-4, Llama 2 Chat 7B, DeBERTa v3 large, RoBERTa large, BERT large
  - 追加学習の学習設定についてはAppendixに記載
- 評価指標
  - 正解率, 適合率, 再現率, F1値

適合率 低 再現率 高：判定が必要なものにも不要と言ってしまう

適合率 高 再現率 低：判定不要なのに中々不要と言わない

実験1：追加学習が効果的

- Llama $2_{\text{Chat 7B}}$に追加学習を施したものの精度が一番高い

モデル名	アーキテクチャ	パラメータサイズ	追加学習	正解率	適合率	再現率	F1 値
GPT-3.5	デコーダ	非公開	×	57.73	58.17	**96.74**	72.65
GPT-4	デコーダ	非公開	×	57.73	58.99	89.13	71.00
$\text{Llama 2}_{\text{Chat 7B}}$	デコーダ	7B	×	58.99	58.60	100.0	73.90
$\text{Llama 2}_{\text{Chat 7B}}$	デコーダ	7B	✓	**88.33**	**91.53**	88.04	**89.75**
$\text{DeBERTa v3}_{\text{large}}$	エンコーダ	434M	✓	86.75	85.83	81.95	83.85
$\text{RoBERTa}_{\text{large}}$	エンコーダ	355M	✓	84.23	87.39	72.93	79.51
$\text{BERT}_{\text{large}}$	エンコーダ	335M	✓	83.28	80.77	78.95	79.85

実験1：デコーダは適合率が低く再現率が高い

- デコーダモデル（追加学習なし）では、ほとんど"事実判定不要"と予測

モデル名	アーキテクチャ	パラメータサイズ	追加学習	正解率	適合率	再現率	F1 値
GPT-3.5	デコーダ	非公開	×	57.73	58.17	**96.74**	72.65
GPT-4	デコーダ	非公開	×	57.73	58.99	89.13	71.00
Llama $2_{\text{Chat 7B}}$	デコーダ	7B	×	58.99	58.60	100.0	73.90
Llama $2_{\text{Chat 7B}}$	デコーダ	7B	✓	**88.33**	**91.53**	88.04	**89.75**
DeBERTa $v3_{\text{large}}$	エンコーダ	434M	✓	86.75	85.83	81.95	83.85
$\text{RoBERTa}_{\text{large}}$	エンコーダ	355M	✓	84.23	87.39	72.93	79.51
$\text{BERT}_{\text{large}}$	エンコーダ	335M	✓	83.28	80.77	78.95	79.85

実験1：エンコーダモデルは適合率が高く再現率が低い

- 正解率は同程度．適合率が高く、再現率がやや低い

モデル名	アーキテクチャ	パラメータサイズ	追加学習	正解率	適合率	再現率	F1 値
GPT-3.5	デコーダ	非公開	×	57.73	58.17	**96.74**	72.65
GPT-4	デコーダ	非公開	×	57.73	58.99	89.13	71.00
Llama $2_{\text{Chat 7B}}$	デコーダ	7B	×	58.99	58.60	100.0	73.90
Llama $2_{\text{Chat 7B}}$	デコーダ	7B	✓	**88.33**	**91.53**	88.04	**89.75**
DeBERTa v3_{large}	エンコーダ	434M	✓	86.75	85.83	81.95	83.85
$\text{RoBERTa}_{\text{large}}$	エンコーダ	355M	✓	84.23	87.39	72.93	79.51
$\text{BERT}_{\text{large}}$	エンコーダ	335M	✓	83.28	80.77	78.95	79.85

実験2：学習データの量によって分類精度がどのように変わるか

- 使用したモデル
  - Llama 2 Chat 7B（デコーダ）、DeBERTa v3 large （エンコーダ）

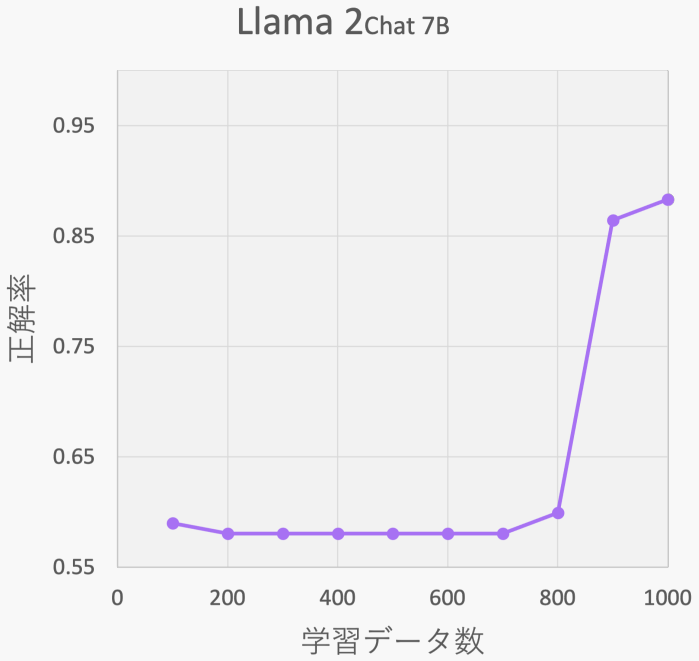
- 学習データ
  - データ量を{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000}として実験
  - 追加学習の学習設定についてはAppendixに記載

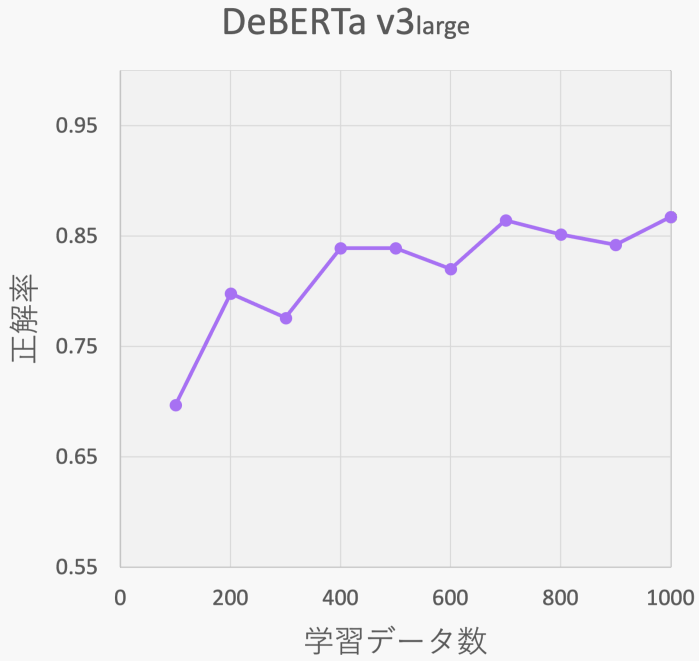
- 評価指標
  - 正解率

実験2：結果

- Llama 2ではデータ数が800件を超えてから顕著に正解率が増加
- DeBERTaではLlama 2と比べて緩やかな正解率が増加

→さらなるデータ収集によって分類精度の向上の可能性あり





実験2：定性分析

- データ量少で上手く予測できず，データ量多で予測できた例
  - 固有名詞や年代があるときに"事実正誤判定が必要"と予測できるようになっている傾向

ex.1) It's pretty cool, his last name is ''Dumile'' hence the DOOM part and he borrows imagery from the Comic character Doctor Doom

ex.2) It was first documented all the way back to 1481.

ex.3) Toews is great, he was the third overall pick in the 2006 NHL draft.

まとめと今後の方向性

- 本研究では対話における事実正誤判定が不要な応答を検出するタスクを提案
  - 学習・評価データセット (Dialogue Dataset annotated with Fact-Check-needed label, DDFC)の作成
  - Llama 2のベースラインで，約88%の正解率で検出可能であることを確認

- データセットの拡充
  - 実験2より，データを増やすことでモデルの精度が向上する可能性大

- 対話応答システムへの適用
  - 事実正誤判定前に判定不要な文を検出することで魅力度と事実性の担保が両立できるか

Appendix

追加学習の学習設定

parameter	encoder	decoder
Number of epochs	5	2
Global batch sizes	64	32
Optimizer	AdamW	AdamW
Learning rate	5.0×10^{-4}	5.0×10^{-5}
Scheduler	cosine	cosine
Max length	256	$1,024$

Table 3: Fine-tuning settings