サプライザルを用いた対話話題誘導の評価とその分析

吉田快 [1,2]　吉野幸一郎 [3,2,1]
[1] 奈良先端科学技術大学院大学　[2] 理研ガーディアンロボットプロジェクト
[3] 東京科学大学
yoshida.kai.yf1@is.naist.jp　yoshino.k.ai@m.titech.ac.jp

概要

推薦や説得、情報収集などシステムが事前に与えられた対話目標を達成するために対話を進める対話タスクを総括して目標指向対話システムと呼ぶ。これら目標指向対話の従来の研究では、目標にたどり着くための話題の切り替えに焦点が当てられてきた。一方で、ユーザに誘導感無く、目標を気づかれることなくタスクを完了することも、自然な対話体験とシステムの目標達成の両立をする上では重要である。本研究ではこれら発話の誘導感と目標の気づかれやすさを、外部言語モデルを用いたサプライザルにより定量化する。これらサプライザルベースの評価によって発話選択を行うシステムの対話実験を行った結果、発話の誘導感を減らすことができることが示された。また対話実験の分析の結果、提案する評価と人の感じる誘導感には相関があること、対話の進行につれて目標に対するサプライザルが低下していくことが確認された。

1　はじめに

対話システムの研究において、対話の中で達成されるべき目標は重要な役割を担ってきた。特に、対話システムがあらかじめ特定の話題や目的を持ち、それに向けて対話を進行させる枠組みは、目標指向対話システム（Target-Guided Conversation）として、タスク指向・非タスク指向の双方で広く研究が進められている。例えばタスク指向の分野においては、商品の推薦や情報提供 [1–3]、ユーザの説得 [4, 5]、対話を通じたユーザ情報の獲得 [6] 等が行われている。非タスク指向の目標指向対話では、システム自身が話したい話題を持ち、対話の流れを自然に操作しながらその話題へと誘導することが目的とされることが多い [7–14]。これらの研究では、現在の話題から目標話題へ至る話題遷移の系列を設計・推論することに重点が置かれてきた。

目標指向対話においては、対話システムが持つ目標とユーザの対話に対するエンゲージメントの双方を考慮しながら対話を進行することが重要とされる [1–6, 8, 12–16]。ただし、不自然な誘導や話題転換はしばしばユーザエンゲージメント低下の要因となる。ユーザエンゲージメントを損なわないためには、過去の対話文脈に対して自然で、なおかつ徐々にシステムの目的へ向かうようシステム自身の発話を選択していく必要がある。言い換えれば、対話履歴に対して自然で、目標となる話題への誘導が露骨に透けて見えてこない発話選択が必要ということである。

本研究では、この課題に対してサプライザル理論 [17, 18] に着目する。サプライザル理論では、与えられた文脈に対して予測されにくい語や文ほど高いサプライザルを持つとされる。一方で、文脈に対して発話のサプライザルが低いことは、その発話が文脈において自然で予測しやすいことを表す。そのためサプライザル理論は、人間の予測する発話と実際の発話のずれを定量化する指標として利用されている。対話文脈に対して自然な発話は予測されやすく、低いサプライザルを持つ一方で、ユーザの予測から大きく外れた発話は高いサプライザルを持ち、違和感を持たれやすい。

本研究では、過去の発話に対し自然で (サプライザルが低く)、対話目標に対する隠蔽性が高い (サプライザルが高い) 対話応答選択を実装し、そのユーザ評価を報告する。また、実際に提案法により選択されたシステム発話の性質について分析を行った結果を報告する。

2　目標指向対話のタスクと評価

本研究では、非タスク指向対話においてあらかじめ与えられた目標発話（target）に自然に到達するよう対話を誘導することを目指す。特に本研究では、ユーザとの対話において以下の 2 点を同時に満たす

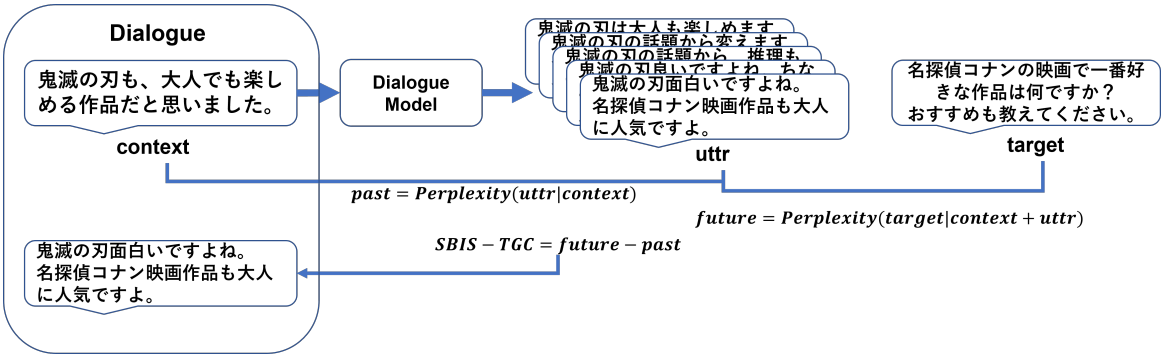

図 1　SBIS-TGC による発話評価と発話選択の例

ことを研究目標とする：

1. 誘導感を与えず、対話を進行させること
2. あらかじめ与えた目標話題に到達すること

この目的に対し、本研究では各発話候補の誘導感を自動評価する指標を用い、これを用いて発話選択を行う対話システムを用いた対話実験を実施することで、上記 2 点への影響を確認する。

2.1　SBIS-TGC

サプライザル理論に基づいた、目標に対する発話の誘導感を評価する **S**urprisal **B**ased **I**nduction **S**core for **T**arget-**G**uided **C**onversation (SBIS-TGC) を提案する。目標指向対話において、ユーザの予測から外れたサプライザルの高い発話は、唐突感を与え、対話における誘導の意図をユーザに悟らせる可能性が高いと仮定する。またシステムの目標が現在の対話文脈から容易に想像できる場合は対話目標のサプライザルが低い状態になると考えられる。つまり**対話履歴に対する発話のサプライザルを下げ、同時に対話履歴と発話に対する目標のサプライザルを上げる**ことが今回の研究目標達成に繋がるという仮定を持つ。SBIS-TGC は現在の文脈 (context) とその発話候補 (uttr)、事前に与えられた目標 (target) を用いてそれらを計算する。入力文 $x = (x_1, x_2, \ldots, x_t)$ に対する、定式化した誘導度スコア SBIS-TGC は式 (5) で計算できる。

$$\text{Surprisal}(X) = -\frac{1}{t}\sum_{i}^{t} \log_\theta(x_i|x_{<i}), \quad \text{where } x_i \in X \tag{1}$$

$$\text{Perplexity} = \exp\{\text{Surprisal}(X)\} \tag{2}$$

$$\text{past} = \text{Perplexity}(\text{uttr}|\text{context}) \tag{3}$$

$$\text{future} = \text{Perplexity}(\text{target}|\text{context} + \text{uttr}) \tag{4}$$

$$\text{SBIS} - \text{TGC} = \text{future} - \text{past} \tag{5}$$

ここで、$P_\theta(x)$ はパラメータ θ を持つ言語モデルが入力文 X に与える尤度である。言語モデルが与える尤度から計算できる平均サプライザルは、*uttr* が *context* に対してどれほど予測困難かを定量化する指標である。*past* は発話 *uttr* の対話履歴 *context* に対するサプライザルを計算することで、現在の対話文脈に対する発話の予測可能性、つまり発話の誘導感を導出する。*future* は現在の対話状態 *context* + *uttr* によって目標 *target* が予測されやすいかを表す。*future* は大きいほどよいものであるのに対し、*past* は小さいほどよいため、最終的な評価値は *future* − *past* として計算を行う。

2.2　発話選択の実装と実験

本研究では、複数の異なる性質を持つ対話誘導発話候補を大規模言語モデルベースの発話候補生成器に出力させる。それら候補を、SBIS-TGC による評価値を用いて選択し、応答として用いる雑談対話システムを実装する。雑談におけるシステムの目標は最初の話題と直接関わりのない話題に対話の中でユーザに気づかれないように言及することである。以下ではこの実装と、実際にどのように選択を行ったか、またこのシステムが期待通り機能しているかを確認するための評価設計について説明をする。

2.2.1　システムの概要

対話実験では雑談のみをするシステム (open) と文脈に対して 5 種類の異なる目標誘導発話候補を選択するシステム (baseline、proposed) を実装し、計 3 種類のシステムの比較を行う。誘導発話候補の生成・選択を行うシステムの枠組みを図 1 に示す。この際 proposed のシステムは誘導スコア SBIS-TGC が最も低いものを選択するのに対し、baseline は生成された発話候補からランダムに選択を行う。発話候補の

表 1 手法ごとの目標達成率と目標を予測されずに達成できた対話数。proposed の右の数字は目標達成できたセッションのみを集計対象とした数

手法	目標達成率 (%)	予測無し (%, ↑)	キーワードマッチング (%, ↑)	LLM マッチング (%, ↑)
open	0	82.3	100	100
baseline	**100**	30.4	46.1	39.2
proposed	70.6	**39.2/38.9**	**70.6/62.5**	**47.1/44.4**

生成には大規模言語モデル (LLM) を用いる。

2.2.2 実験の概要

対話実験では 102 名の評価者が open と baseline、proposed の 3 種類のシステムに対してそれぞれ 7 ターンの対話を行う。誘導を行う baseline と proposed に関しては目標に到達し次第、雑談プロンプトに切り替わる。この際、目標達成はキーワードマッチングと LLM による判定の 2 つを使用している。目標達成判定に用いたプロンプトは付録 A に示す。

今回、システムの目標は特定の話題を反映した発話とし、ランダムに選択された話題を用いて目標発話を生成した。目標発話として用いる話題は、Wikipedia における 2024 年 5 月の閲覧数上位 300 記事をもとに選定した。まず、不適切な話題を人手で除外した上で、残った 287 件からランダムに 10 件を選出する。その後、これらの話題が広く知られているかどうかを LLM により判定し、選択されたものを最終的な目標話題として用いた。話題リストと目標発話生成に使用したプロンプトは付録 B に示す。また、各発話生成と目標話題選択、目標発話生成には ChatGPT [1] を用い、SBIS-TGC の計算のための言語モデルには sarashina-2.2-3b を用いた。 [2] [3]

2.2.3 評価指標

人手評価では、対話者が各対話の終了後に発話ごとに 3 段階の唐突感スコアの付与と、対話を通じて「システムが話したがっていると感じた話題 (誘導先)」の記述を行う。発話ごとの唐突感スコアは「唐突でない、人によっては唐突に感じる、多くの人が唐突と感じる」の 3 種である。誘導先記述のインストラクションは付録 C に示す。

誘導先予測評価の集計には、ルールベースのキーワードマッチングと LLM によるマッチングの 2 種類を用いた。LLM マッチングに用いたプロンプトは付録 C に示す。

1) gpt-4.1-mini-2025-04-14
2) https://huggingface.co/sbintuitions/sarashina2.2-3b
3) これは、オンプレミスでの計量モデルを用いることで、生成された発話候補に対して速やかに SIBS-TGC の評価値を付与し、対話のライブ感を維持するためである。

3 実験結果

まず、各 102 件の対話実験中に open は 714 発話を、baseline では 135 発話と open を 579 発話、proposed では 422 発話と open を 292 発話生成していた。proposed の評価では対話目標を達成して open に移行するまでによりターン数を掛けている一方、baseline はランダムな選択を行うためより早く open へと移行したことが見て取れる。

次に目標達成率と目標を予測されずに達成できた数をそれぞれ表 1 に示す。この際、「予測無し」はユーザがシステムの誘導先を予測できなかった数、「キーワードマッチング」は目標話題と予測された話題のキーワードマッチングによって一致しなかった割合、LLM マッチングは LLM を用いた一致の判定で一致と判定されたなかった割合を表す。また、proposed のみ対話目標未到達の対話が含まれているため、左に全体を対象とした際の割合を、右に未達成のものを除いた際の割合を示す。

まず、目標達成率に着目すると baseline は 100%の対話で目標に到達していたのに対し、proposed では 70.6%の目標達成率となっている。baseline の目標達成率が proposed より 30%近く高いのは、level 4 や 5 の発話が明確に target に言及する発話であるため、強制的に対話目標に到達したと判定されるためである。次に、目標を予測されずに目標達成できた対話セッション (LLM マッチングと予測無し、キーワードマッチング) に注目すると、これら全てで proposed は baseline と比べて目標を予測されずに対話目標の言及に成功している。また、目標未達成の対話を除いた際の割合についても proposed が目標を予測されずに対話目標の言及に成功していることが確認できる。

次に手法事の唐突感評価を図 2 に示す。目標未達成の場合を含め proposed は baseline に「唐突でない」が倍以上出やすく、「多くの人が唐突と感じる」に関して倍以上出にくい結果となっている。これは SBIS-TCG によって誘導感のない発話が選択できた

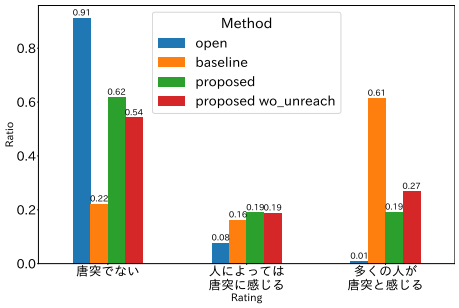

図 2　手法別の唐突感

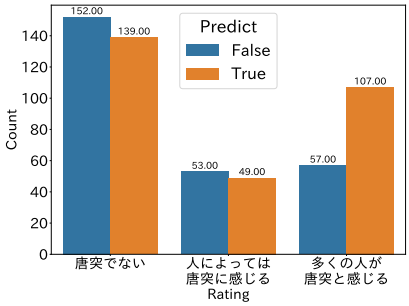

図 3　唐突感ごとの誘導先予測の成否

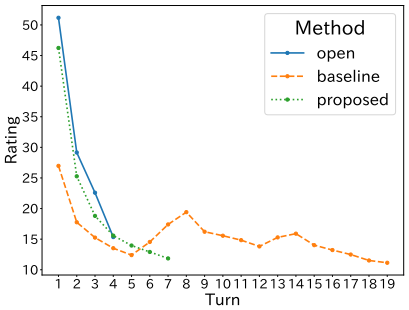

図 4　ターンごとの future の遷移

ことを示唆している。

唐突感ラベルとユーザによる話題誘導先予測成否 (失敗する方がよい) との関係を図 3 に示す。誘導を行った baseline と proposed のみを対象とし、誘導先予測の成功判定には LLM の判定結果を用いている。「唐突でない」に関しては、誘導先予測が成功した True に比べ、失敗した False の件数が多い。「多くの人が唐突と感じる」に関しては、True が False の 2 倍程度の数となっている。これは、発話に唐突感があるほどユーザが誘導先の予測に成功してしまうことを示している。そのため、ユーザに発話の唐突感を与えないことが誘導を感じさせないために重要であるという本研究の前提を指示する結果が得られたと考えられる。

4　分析

SBIS-TGC における past と future のそれぞれが我々の仮説通りの役割を担っていたか確認をする。つまり、past の仮説である「対話文脈に対してサプライザルの高い発話が唐突感を与える」こと、future の仮説である「現在の対話文脈に対し目標のサプライザルが低い状態はユーザからのシステム目標の予測を容易にすること」について、どの程度うまく機能していたかを分析する。

まず past と future のそれぞれについて、発話ごとの各スコアと唐突感スコアの相関を計算する。この際、「唐突でない」「人によっては唐突に感じる」、「多くの人が唐突と感じる」のそれぞれに 1 から 3 の値を割り当て、計算を行う。結果、各スコア間のピアソンの相関係数について past は 0.418、future は-0.338 であった。これは past、future いずれの仮設についても支持する結果となっている。

次に、対話ターンごとの future の遷移を図 4 に示す。対話ターンの増加につれ、future の値が減少していることが確認できる。open が比較的横ばいなのに対し、baseline と proposed では open より高い値で始まり、ターン経過によって open より急激に減少している。また、ターン経過によりスコアが減少していくことは、誘導の継続によって目標が予測されやすくなっていることを示唆している。

5　まとめ

本研究では、目標指向対話においてサプライザル理論が誘導感と誘導先を悟らせないことに有効であるという仮説を検証するため、対話実験によりその確認を行った。SBIS-TGC によって発話選択を行うシステムを用いた対話実験では、SBIS-TGC による発話選択によってシステムの発話の唐突感が減少することが確認できた。また、目標誘導におけるサプライザルと誘導感には相関関係が見られた。さらに、対話ターンが進むにつれ、目標のサプライザルが下がるなど、サプライザルによって目標への到達状態を評価できることが示唆された。一方で、目標達成率と誘導感、ユーザに目標を気づかれないことにはトレードオフの関係があることが判明した。

倫理に関する表明

本研究ではユーザに誘導感を与えずシステム自身の目標を達成する技術を対象としている。これは目標指向対話システムにおける自然な対話体験を可能にするものであるが、その一方で、不適切に使用された場合にはユーザの自律性を損ねたり、ユーザの行動を操作する手段として使われるリスクも存在する。特に説得対話システムのようにユーザの行動変容を促す目標を設定する場合、対話システムの正体の開示、個人情報や心理的傾向の分析に対する同意、応答の適切性・非差別性の確保 といった原則が重視される [5]。

謝辞

本研究の一部は科研費 23K24910 の支援を受けた。

References

[1]Koichiro Yoshino, Yu Suzuki, and Satoshi Nakamura. “Information Navigation System with Discovering User Interests”. In: **Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue**. 2017, pp. 356–359.

[2]Raymond Li et al. “Towards Deep Conversational Recommendations”. In: **arXiv:1812.07617** (2018).

[3]Kun Zhou et al. “Towards Topic-Guided Conversational Recommender System”. In: **Proceedings of the 28th International Conference on Computational Linguistics**. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4128–4139.

[4]Takuya Hiraoka et al. “Reinforcement Learning of Cooperative Persuasive Dialogue Policies using Framing”. In: **Proceedings of coling 2014, the 25th international conference on computational linguistics: technical papers**. 2014, pp. 1706–1717.

[5]Xuewei Wang et al. “Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good”. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5635–5649.

[6]Shiki Sato et al. “Proactive User Information Acquisition via Chats on User-Favored Topics”. In: **arXiv:2504.07698** (2025).

[7]Marine Riou. “A Methodology for the Identification of Topic Transitions in Interaction”. In: **Discours** 16 (2015). Published online on September 9, 2015. Accessed on December 4, 2024.

[8]Jianheng Tang et al. “Target-Guided Open-Domain Conversation”. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5624–5634.

[9]Wenquan Wu et al. “Proactive Human-Machine Conversation with Explicit Conversation Goal”. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3794–3804.

[10]Karin Sevegnani et al. “OTTers: One-turn Topic Transitions for Open-Domain Dialogue”. In: **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**. Online: Association for Computational Linguistics, Aug. 2021, pp. 2492–2504.

[11]Jingxuan Yang, Si Li, and Jun Guo. “Multi-Turn Target-Guided Topic Prediction with Monte Carlo Tree Search”. In: **Proceedings of the 18th International Conference on Natural Language Processing (ICON)**. National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLPAI), Dec. 2021, pp. 324–334.

[12]Prakhar Gupta, Harsh Jhamtani, and Jeffrey Bigham. “Target-Guided Dialogue Response Generation Using Commonsense and Data Augmentation”. In: **Findings of the Association for Computational Linguistics: NAACL 2022**. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 1301–1317.

[13]Yosuke Kishinami et al. “Target-Guided Open-Domain Conversation Planning”. In: **Proceedings of the 29th International Conference on Computational Linguistics**. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 660–668.

[14]Yang Deng et al. “Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration”. In: **Findings of the Association for Computational Linguistics: EMNLP 2023**. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10602–10621.

[15]Anqi Liu et al. “MTGP: Multi-turn Target-oriented Dialogue Guided by Generative Global Path with Flexible Turns”. In: **Findings of the Association for Computational Linguistics: ACL 2023**. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 259–271.

[16]Yang Deng et al. “Proactive Conversational AI: A Comprehensive Survey of Advancements and Opportunities”. In: **ACM Trans. Inf. Syst.** 43.3 (Mar. 2025).

[17]John Hale. “A Probabilistic Earley Parser as a Psycholinguistic Model”. In: **Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies**. 2001.

[18]Roger Levy. “Expectation-Based Syntactic Comprehension”. In: **Cognition** 106.3 (2008), pp. 1126–1177.

A　誘導発話と雑談発話生成に用いたプロンプト

誘導発話生成と雑談発話生成に用いたプロンプトを以下に示す。この際、{context}は対話文脈、{target_topic}は目標話題で置き換えている。また、誘導発話に用いたプロンプトは Riou ら [7] の話題転換の分析を参考に level 3 から 5 を作成し、雑談に近いものを level 1 と 2 として追加したものである。

誘導発話生成に用いたプロンプト

あなたはユーザを目標話題（「{target_topic}」）に導く対話発話を作る AI です。
以下のユーザとの対話文脈に対して自然かつ、5 種類の誘導を反映した発話を生成してください。
ただし、各発話では、必ず以下の 4 つの要素を守ってください。
【1】指定された誘導レベルに応じて **{target_topic}** への誘導を行ってください。
【2】各発話はそれぞれ異なる内容であり、*****50 文字以内*****である必要があります。
【3】**ユーザの最後の発話に対する自然な発話**を含めます。
【4】ただし誘導によって発話が不自然になる場合は、発話の自然さを優先してください。

- level 1（自然な発話）：ユーザの発話に対して自然な発話をする。ただし、可能であれば「{target_topic}」の周辺知識に近づくような唐突さのない発話にする。
- level 2（自然な発話と誘導）：ユーザの発話に対して自然な発話をする。その後、可能であれば現在の話題から「{target_topic}」に関連する話題の提案を行う。
- level 3（想起と誘導）：対話中の過去の話題から「{target_topic}」へ遷移しやすい話題を見つけ、遷移を試みる。
- level 4（承認と誘導）：ユーザの発話に肯定的に答えた上で、明確に話題を切り替え「{target_topic}」に関する新たな話題を提供する。
- level 5（直接的な誘導）：現在の対話の流れを切り、明確に「{target_topic}」について話し始める。

出力は以下の形式でお願いします。
- level 1: level 1 の発話
- level 2: level 2 の発話
- level 3: level 3 の発話
- level 4: level 4 の発話
- level 5: level 5 の発話

対話履歴
{context}

雑談発話 (open) 生成に用いたプロンプト

以下の対話履歴に続く自然な発話を生成してください。
また、話の流れに沿ってユーザが興味を持ちそうな話題を選び、自然な会話の流れを維持してください。
ただし質問の多用は避け、現在の話題に対するユーザの反応が良くなければ現在の話題を止めて新しい話題を考え、切り替えることも考慮してください。
1 発話は 40 語以内に収めるように意識してください。
また出力は発話のみでお願いします。

対話履歴
{context}

B　目標発話生成に用いたプロンプトと話題リスト

システムの目標発話を生成するために使用したプロンプトと、実際に対話実験で用いられた話題語を以下に示す。

目標発話生成に使用したプロンプト

あなたは話題:「{target_topic}」についてユーザと話したい AI です。対話の流れの中でユーザに「{target_topic}」について話かけるための発話を考えてください。発話は 20 語以内でお願いします。

目標に用いられた話題リスト

大谷翔平, 鎌倉殿の 13 人, 名探偵コナンのアニメエピソード一覧, 日本航空 123 便墜落事故, BTS_(音楽グループ), SPY × FAMILY, 呪術廻戦, 鎌倉幕府, 三浦春馬, 名探偵コナン映画作品, 木村拓哉, 鋼の錬金術師, Twitter, トップガン_マーヴェリック, 志村けん, 名探偵コナン, 第二次世界大戦, 鬼滅の刃, ワイルド・スピードシリーズ, 名探偵コナン_(アニメ), 新垣結衣, シン・ウルトラマン, イチロー, 連続テレビ小説, 米津玄師, 錦鯉_(お笑いコンビ), 第一次世界大戦, キングダム_(漫画), X_JAPAN, ONE_PIECE, 名探偵コナンの登場人物, 斎藤工, 乃木坂 46, ウルトラマン, ハリー・ポッターシリーズ, 三種の神器, HUNTER × HUNTER, 菅田将暉, ウクライナ, イーロン・マスク, 織田信長, ゴールデンカムイ, ジョジョの奇妙な冒険, トップガン_(映画), 岸田文雄, 梨泰院クラス, かぐや様は告らせたい〜天才たちの恋愛頭脳戦〜, マイファミリー, アメリカ合衆国, HIKAKIN, 暴太郎戦隊ドンブラザーズ, 東京卍リベンジャーズ, ハリー・ポッターシリーズの登場人物一覧, Hey!_Say!_JUMP, YouTube, 小栗旬, ウィキペディア, 北条義時, 知床観光船沈没事故, 崖の上のポニョ, 綾瀬はるか, King_&_Prince, 五等分の花嫁, Mr.Children, 庵野秀明

C　評価判定に用いたプロンプト

対話参加者に与えたシステムの目標予測の評価指示文と、ユーザの予測がシステム目標と一致しているかを判定するプロンプトをそれぞれ以下に示す。

対話誘導先記述の評価

システムが話したい話題を持っていると感じた場合はそれを記入してください (例: 野球、ポケモン、阿部寛など)。ない場合は空欄でお願いします。

ユーザの誘導先予測とシステムの目標の一致判定に用いたプロンプト

以下はシステムが対話目標としていた話題です。
{target_topic}
この際、{feel_topic}は目標話題とある程度同じものだと言えますか？同じ場合、「True」を違う場合は「False」を出力してください。
出力は「True」「False」のみでお願いします。