家族日常対話における計画情報抽出

沼屋征海 [1] 佐藤魁 [1] 吉田倖 [1] 亀井遼平 [1]
所年雄 [2] 滑川登 [2] 濱口泰時 [2] 上村崇 [3] 乾健太郎 [4,1,5] 坂口慶祐 [1,5]
[1] 東北大学 [2] トランスコスモス株式会社
[3]epiST Ventures 株式会社 [4]MBZUAI [5] 理化学研究所
numaya.ikumi.t4@dc.tohoku.ac.jp

概要

本研究では，家庭内の日常対話に現れる計画情報を対象とした新たな計画情報抽出タスクを提案し，それに対応する「Tohoku-Transcosmos 家庭内会話計画コーパス」を構築・公開する[1)]．提案タスクは，対話の進行に伴う計画情報の追加・更新を発話単位で逐次的に扱う点を特徴とする．コーパス構築では，子どもを含む家庭内対話を音声収録・文字起こしし，計画情報リストの操作を人手で注釈付けした．大規模言語モデルを用いた予備実験では一定の性能が確認されたが，行動の粒度や子どもによる曖昧表現の解釈に関する課題が明らかになった．

1 はじめに

Siri や Alexa などの音声アシスタントの家庭への浸透と，大規模言語モデル (LLM) の発展を背景として，日常対話に含まれる計画情報を理解・抽出する技術への関心が高まっている [1, 2, 3, 4]．例えば，メール内容から ToDo 項目を抽出する手法 [5] や，オンラインカレンダーを参照して指定された予定を自動生成する手法 [6] など，今後の計画に関わる情報を自動管理する枠組みが提案されており，特に計画に関する話題は，人間同士の対話において頻繁に現れることが指摘されている [7, 8]．

対話研究においては，日常対話を収集したデータセットが多く提案され [9, 10, 11]，注釈付けを行うことで対話理解を試みる研究が進められているが [12, 13, 14]，家庭における日常対話に現れる計画情報をどのように定義し，抽出・評価するかというタスク設計やデータセットは，体系的に整備されているとは言い難い．

そこで本研究では，家庭における日常対話に現れる計画情報の抽出を対象とした新たなタスクを提案し，注釈付き対話データセットを構築・公開する．計画情報抽出タスクは，対話の進行に伴い，発話ごとに計画情報の追加や更新を行う点を特徴とする (図 1，表 1)．計画情報抽出タスクに対応するデータセットとして，本研究では「**Tohoku-Transcosmos 家庭内会話計画コーパス**」を構築した．本コーパスでは，子どもを含む家庭内の日常対話を対象に音声収録および文字起こしを行い，各対話に対する計画情報リストを整理した上で，タスク定義に基づく発話ごとのリスト操作を人手で注釈付けしている．大規模言語モデルを用いた予備実験の結果，計画情報の抽出において一定程度の性能が確認された一方で，抽出すべき行動の粒度の判断や，子どもによる曖昧表現の解釈が求められる場面では課題が残ることが明らかとなった．

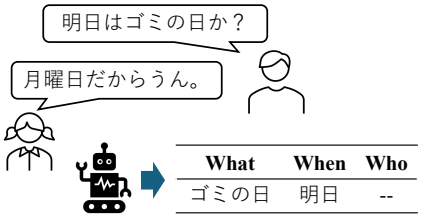

図 1 LLM による家族対話からの行動・予定抽出．

2 計画情報抽出タスク

本研究では，家庭内対話における計画情報を自動抽出するタスクを新たに提案する．このタスクでは，言及された計画情報を一覧にした計画情報リストを作成することを目的とする．

ここで，計画情報の要素として行動と予定をそれぞれ以下のように定義する：

- **行動**：同居家族の行動や行為に焦点のあたる未来の出来事
- **予定**：同居家族の行動や行為が行われる日時に焦点のあたる未来の出来事

1) https://github.com/tohoku-nlp/tt-cfcp

表 1　計画情報リスト例.

ID	Status	What	When	Where	Who	Why	How
行動 0	未確定	歯の検診	–	–	母	–	–
予定 0	確定	ゴミの日	明日	–	–	–	–

2.1　計画情報項目の定義

計画情報リストでは，情報抽出の先行研究に従い [15, 16]，「いつ・どこで・誰が・何を・なぜ・どのように」に対応する 5W1H の観点を用いる．計画情報リストの例を表 1 に示す．リストは ID，Status，What，When，Where，Who，Why，How から構成される．Status は，各計画情報の状態を表し，**未確定/確定/終了/削除済み** の 4 値をとる．**未確定**は，対話内で計画情報が提案または検討されている段階など，実施が合意されていない状態を指す．**確定**は，断定的な発話や家族間の同意により，実施が合意された状態を指す．また，計画情報が実施済みであることが示された状態を**終了**，対話内で不要であることが示された状態を**削除済み** とする．

2.2　タスクの流れ

計画情報リストは発話ごとに情報の追加や更新が行われる．ここで，発話ごとの計画情報リストに対する操作として**追加/更新**を定義する．**追加**では，新たな行動や予定に対応する計画情報を追加する．**更新**では，既存の計画情報の内容や空欄要素を更新する．発話からリスト操作を抽出するための推論手順を図 2 に示す．この図から，リスト操作を以下の 3 段階に分けて定義する：

1. **計画情報種別判定**：発話が**行動**または**予定**に関連するかを分類する．(青色部分)
2. **操作種別判定と紐づけ**：行うべき操作を**追加**または**更新**で判定し，**更新**の場合は対象となる ID を特定する．(紫色部分)
3. **情報抽出**：操作に必要な情報を発話から抽出する．(橙色部分)

本タスクでは，このような流れに従い計画情報リストを作成することを目指す．

3　データセット

3.1　対話収集

幼児または小中学生のいる世帯を対象に協力者を募集し，2 世帯を対象として合計 3 対話を録音した．各世帯は録音したデータについて個人情報を保護することを条件に研究目的で使用，公開することに同意している．同居家族が集まる状況で自然な対話を収録するため，話題を指定せず食事の時間帯で録音を実施した．話者の年齢情報は付録 A に示す．次に，人手または LLM によって収集した音声の文字起こしを行った．1 対話は人手で，残り 2 対話は `Gemini-1.5-pro` [17] によって文字起こしした後に人手で修正を行った．また，笑い声は [laughs]，話し声で聞き取れなかったものは [inaudible]，その他非言語的な音声は [noise] のタグとして表記した [18].

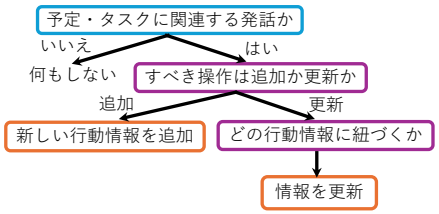

図 2　リスト操作に関する情報抽出の推論手順.

表 2　基本統計量.

対話数	3 対話
全発話数	1573 発話
発話あたりの文字数	12.59 文字
リスト操作数	91

3.2　注釈付け

各対話に一名ずつの日本語母語話者を注釈者として割り当て，第 2 節で定義したタスク設計に従い注釈付けを行った．注釈付け後，人手で確認を行い，内容に影響しない表記や形式上の不整合を定義に基づいて修正した．こうして得られた注釈付き対話データの例は付録 B に示す．

3.3　統計情報

構築したデータセットの基本統計量を表 2 に示す．本データセットは 3 対話 (計 1573 発話) からなり，発話長は平均 12.59 文字と短い．実際に，口語表現や省略，指示表現が多く見られる．また，リスト操作は 91 件であり，発話全体の約 5.8%に相当する．すなわち，「操作が必要な発話」が相対的に少ない不均衡な設定である点に留意する．

4　実験

第 2.2 節において 3 段階で定義した計画情報抽出に対する LLM の能力を検証するため，各段階に対応した実験・評価を行う．

表 3 各実験における評価スコア．F1 は F1 スコア，Acc は Accuracy，sim はコサイン類似度を表す．

	実験 1 (§ 4.1)	実験 2 (§ 4.2)		実験 3 (§ 4.3)	
文脈	(F1)	2.1 (Acc)	2.2 (Acc)	3.1 (F1)	3.2 (sim)
0	0.548	0.738	0.659	0.588	0.874
1	0.576	0.655	0.902	0.589	0.870
2	0.597	0.786	0.951	0.597	0.867
3	0.589	0.762	0.976	0.597	0.868
ベース	0.332	0.443	0.287	0.346	–

共通設定　全実験共通の設定として，モデルは GPT-5 [19] を用い，0-shot のプロンプトを入力することにより実験を行った．入力には，計画情報抽出の対象となる発話 (ターゲット発話) に加えて，直前の過去 k 発話 (文脈発話) ($k \in \{0,\ldots,3\}$) を与えた．また，ラベル分布に基づくランダム予測ベースライン (ベース) を 100 回試行した平均として算出した．さらに，ターゲット発話時点における計画情報リストなど，入力に必要な上流段階の情報は人手の注釈を用いて与えた．そのため，上流段階の誤りが下流段階に伝搬する影響は本実験では扱わない．

4.1　計画情報種別判定

ここでは，「**行動**または**予定**への関連性を正しく判定できるか」(図 2 青色部分) を検証する．

実験設定　実験 1 では，**行動**/**予定**/None (計画情報を含まない) の 3 クラス分類精度を評価した．正解ラベルは，ターゲット発話に付与された行動に関する操作が 1 つ以上存在する場合を**行動**，予定に関する操作が 1 つ以上存在する場合を**予定**，いずれも存在しない場合を None とした．なお，1 発話に**行動**と**予定**の両方の操作が付与されている場合は，**行動**を正解ラベルとした．評価指標は，3 クラスの F1 スコアを平均した Macro-F1 スコアを用いた．

結果　実験結果を表 3 (実験 1) に示す．文脈数によらず，ベースラインを超える F1 スコアが示された．また，ターゲット発話以前の文脈発話を入力に含むことで精度が向上した．一方で，F1 スコアはすべての条件で 0.6 を下回っており，ベースラインは超えるが実用上は不十分であると考えられる．

4.2　操作種別判定/紐づけ

ここでは「既存の計画情報リストに対して行うべき操作種別を判定できるか」(図 2 紫色部分) を検証する．具体的には，以下の二つの実験を行った．

- **実験 2.1**: 適切な操作 (**追加**/**更新**) を正しく選択できるかを評価する．
- **実験 2.2**: **更新**操作に該当する場合に，更新対象の ID と正しく紐付けられるかを評価する．

実験設定　実験 2.1 では，注釈が付与された発話のみをターゲット発話とし，文脈発話とターゲット発話時点の計画情報リストを同時に入力として与え，適切なリスト操作種別を出力するようモデルに指示した．複数操作が付与された発話では，全操作を正確に出力できた場合を正解とし，**追加**/**更新**/**追加**+**更新**/... のような多クラス分類を評価した．分布は，**追加**のみおよび**更新**のみが同程度である一方，複数操作は極めて少ない (詳細は付録 C)．そこで本実験では，付与された操作集合との完全一致精度 (Accuracy) を評価指標として用いた．実験 2.2 では，**更新**が付与された発話のみを対象とし，実験 2.1 と同様の入力を与えた．評価指標は各発話について正解 ID との完全一致精度 (Accuracy) とした．

結果　実験結果を表 3 (実験 2) に示す．文脈発話数によらず，実験 2.1，2.2 両方においてベースラインを上回る精度を示した．また実験 2.2 では，文脈発話数を増やすことで他の実験と比較して大きく性能向上が確認された．

4.3　情報抽出

ここでは，「リスト操作時のターゲット発話に対する情報抽出がどの程度正確に行われるか」(図 2 橙色部分) を検証する．具体的には，以下の二つの実験を行った．

- **実験 3.1**: 各項目に対する人間と LLM の更新タイミングの一致度を評価する．
- **実験 3.2**: リストの意味的類似度を評価する．

実験設定　実験 3.1，3.2 では，操作種別と，その種別が**更新**であった場合の対象 ID が明らかになった上で，ターゲット発話に対して Status，5W1H の情報抽出を LLM を用いて行った．実験 3.1 では，各発話について LLM が更新した項目集合 (例：What, When, Who など) を予測結果とし，各項目ごとの F1 スコアを平均した Macro-F1 スコアを評価指標として用いた．すなわち，項目の更新タイミングのみを評価した．実験 3.2 では，リスト操作後における人手と LLM による行動リストとの内容比較を行った．具体的には，リスト内の全項目を結合して一つの文字列とし，文埋め込み同士の

表 4　文脈発話数 $k = 0$ における ID 紐づけの誤答例.

区分	話者	発話
文脈発話	父	“そういえばお友達呼ぶ日決めたの”
ターゲット	娘	“24 になりそう”

表 5　実験 1 における分類性能 (文脈発話数 $k = 2$).

指標	精度	再現率	F1	ラベル数
None	0.979	0.956	0.967	1489
行動	0.264	0.378	0.311	47
予定	0.439	0.617	0.513	37

コサイン類似度を用いた．文埋め込みモデルには sentence-bert-base-ja-mean-tokens-v2[2] を用いた．

結果　実験結果を表 3 (実験 3) に示す．実験 3.1 では，項目の更新タイミングの一致度はベースラインを上回るとともに，文脈発話数の増加に伴い F1 スコアは緩やかに上昇する傾向を示した．実験 3.2 では，いずれの文脈数においてもコサイン類似度は 0.86 を上回り，人間と LLM が生成したリスト内容の意味的類似度は高い値を示した．

4.4　分析

発話文脈による精度向上　実験 2.2 において，ターゲット発話に加えて文脈発話を入力することで更新操作における ID の紐づけ精度が大きく向上した．これは，文脈発話が**更新**操作を補助していることが精度向上の一因であると考えられる．文脈発話の利用が正しいラベルの予測に寄与した事例を，表 4 に示す．この例では，文脈発話に対する回答がターゲット発話となっているため，“24” が “お友達を呼ぶ日” を指していることが推測できる．一方，文脈発話を与えない場合には参照先が欠落するため，正しい ID の予測が困難となる．このように**更新**操作では文脈に依存した計画情報の紐づけが要求される場合が多いため，文脈発話を入力することで ID の紐づけ精度が向上したと考えられる．

計画情報種別判定における過検出　次に，他の実験と比較して人手注釈との一致率が低かった実験 1 についてより詳細な分析を行う．表 5 に最も精度の高かった条件 (文脈発話数 $k = 2$) での実験 1 のクラス別スコアを示す．**行動**および**予定**はいずれも再現率が比較的高かった．これは，LLM が計画情報を含む発話を一定程度検出できていることを示唆する．一方で，精度は再現率を下回り，**行動**では 0.264 と低いことから，過検出の傾向を示唆する．

2)　https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

表 6　行動の細かさの解釈に関する不一致の例.

話者	発話	正解	予測
母	“はるちゃん 2 時 40 分下校でしょう”	**予定**	**予定**
娘	“下校時刻になったら速攻で帰ってくる”	None	**行動**

表 7　子どもの発話に対する解釈の不一致の例

話者	発話	正解	予測
父	“すごいじゃん。”	None	None
息子	“別の塾に転職しましょう”	None	**行動**

行動の細かさによる不一致　過検出が生じる理由として，人間と LLM で抽出すべき**行動**の基準が異なることが考えられる．人手による正解ラベルが None である発話に対して，LLM が**行動**と分類した例を表 6，7 に示す．表 6 では，下校の予定に関する発話の後，その予定に付随した “下校時刻になったら帰って来る” という行動が言及されている．人間の判断では細かい行動であることから除外されたが，LLM は実行可能な行動として拾い上げ，**行動**と判定したと考えられる．このことから，人手の注釈におけるリスト化する行動の細かさと，モデルが解釈する範囲に差があることが示唆された．

子どもの発話の解釈による不一致．　さらに表 7 では，“別の塾に転職しましょう” という表現が子どもによる冗談・誇張を含む不確実な提案であり，人手では None として扱われた．一方で，モデルは表層的な提案表現を判断材料とし，**行動**として分類したと考えられる．このことから，人間は話者の情報や裏の意図を踏まえて分類を行っているのに対し，本設定における LLM はそのような情報を十分に捉えられていない可能性が示唆される．

5　おわりに

本研究では，家庭内日常対話で現れる計画情報を抽出するタスクを新たに提案し，計画情報を注釈付けした対話データセットを新たに構築した．LLM による予備実験の結果，一定程度の人手注釈との一致が見られた一方で，抽出すべき行動の細かさや子どもによる曖昧表現の解釈を要するケースでは不一致が生じた．本研究では，人手注釈間の一致度については検証しておらず，不一致が LLM の精度に起因するものか，抽出基準や内容の曖昧性等によるものかは明らかでない．今後は複数の注釈者による注釈を付与し，人手注釈間の一致度を用いた分析を予定している．

謝辞

本研究は，トランスコスモス株式会社，東北大学共創イニシアティブ株式会社による共創プロジェクトの一環として実施されたものである．注釈付け作業担当者に深く感謝の意を表する． なお，研究内容および結論は，著者らの学術的見解に基づくものであり，関係組織の公式見解を示すものではない．

参考文献

[1] Yixuan Hou, Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. SOVA-Bench: Benchmarking the Speech Conversation Ability for LLM-based Voice Assistant. In **Interspeech 2025**, 2025.

[2] Satoshi Akasaki and Manabu Sassano. Detecting ambiguous utterances in an intelligent assistant. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track**, 2024.

[3] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. **Frontiers of Computer Science**, Vol. 18, No. 6, p. 186345, 2024.

[4] Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. Agentic reasoning: A streamlined framework for enhancing LLM reasoning with agentic tools. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2025.

[5] Sudipto Mukherjee, Subhabrata Mukherjee, Marcello Hasegawa, Ahmed Hassan Awadallah, and Ryen White. Smart to-do: Automatic generation of to-do items from emails. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, 2020.

[6] Donghyeon Kim, Jinhyuk Lee, Donghee Choi, Jaehoon Choi, and Jaewoo Kang. Learning user preferences and understanding calendar contexts for event scheduling. **Proceedings of the 27th ACM International Conference on Information and Knowledge Management**, 2018.

[7] Lauren M. Papp. Topics of marital conflict in the everyday lives of empty nest couples and their implications for conflict resolution. **Journal of Couple & Relationship Therapy**, Vol. 17, No. 1, pp. 7–24, 2018.

[8] Daena J. Goldsmith and Leslie A. Baxter. Constituting relationships in talk: A taxonomy of speech events in social and personal relationships. **Human Communication Research**, Vol. 23, No. 1, pp. 87–114, 2006.

[9] Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. Design and evaluation of the corpus of everyday Japanese conversation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, 2022.

[10] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: telephone speech corpus for research and development. In **Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1**, 1992.

[11] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, 2017.

[12] Mai Omura, Hiroshi Matsuda, Masayuki Asahara, and Aya Wakasa. UD_Japanese-CEJC: Dependency relation annotation on corpus of everyday Japanese conversation. In **Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, 2023.

[13] Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito. Construction and analysis of a multimodal chat-talk corpus for dialog systems considering interpersonal closeness. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, 2020.

[14] Alex C. Fang, Harry Bunt, Jing Cao, and Xiaoyue Liu. Collaborative annotation of dialogue acts: Application of a new ISO standard to the switchboard corpus. In **Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data**, 2012.

[15] Yang Cao, Yangsong Lan, Feiyan Zhai, and Piji Li. 5w1h extraction with large language models. In **2024 International Joint Conference on Neural Networks (IJCNN)**, 2024.

[16] Hiromi Narimatsu, Hiroaki Sugiyama, Masahiro Mizukami, and Tsunehiro Arimoto. Chat agents respond more empathetically by using hearsay experience. **Frontiers in Robotics and AI**, Vol. 10, , 2023.

[17] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. https://arxiv.org/abs/2403.05530.

[18] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. In **Interspeech**, 2018.

[19] OpenAI. Gpt-5 system card. Technical report, OpenAI, 2025.

A　話者情報

表 8 に話者の年齢情報を示す．参加した子どもは幼稚園児から小中学生のように言語の発達状況が異なっている．

表 8　話者の年齢情報 (対話 2 と対話 3 は同一話者).

対話	父	母	子ども
対話 1	30 代	30 代	息子：幼稚園児
対話 2–3	40 代	40 代	娘：中学生，息子：小学生

B　対話例

表 9，10 にデータセットに収録されている対話例及び付与された注釈を示す．表 9 の対話では，タスク 0 が**追加**された後，Status および What が対話を通じて更新されている．表 10 の対話は，“鼻血” から “そろばん”，“健康診断” というように対話トピックが素早く切り替わる例であり，家庭内日常対話の特徴を表している．

表 9　データセットに収録された対話及び注釈の例 1．紙面の都合上，5W1H は What と Who のみ.

対話データ		注釈				
話者	発話	ID	操作	Status	What	Who
母	“新聞作るの？”	タスク 0	**追加**	未確定	新聞を作る	息子
息子	“うん”	タスク 0	**更新**	確定	–	–
父	“ふーん。”	–	–	–	–	–
母	“ふーん。”	–	–	–	–	–
母	“クラスで作るの？”	–	–	–	–	–
息子	“広報委員。”	タスク 0	**更新**	–	広報委員で新聞を作る	–

表 10　データセットに収録された対話及び注釈の例 2．紙面の都合上，5W1H は What と Who のみ.

データ		注釈				
話者	発話	ID	操作	Status	What	Who
息子	“今日朝鼻血出た”	–	–	–	–	–
母	“そろばん”	–	–	–	–	–
娘	“休もうぜ”	–	–	–	–	–
息子	“ダメー”	–	–	–	–	–
父	“明日？”	–	–	–	–	–
娘	“休んでも月 1 の”	–	–	–	–	–
息子	“明日お母さん健康診断だ 9 時以降食べちゃだめなんじゃないの今日”	タスク 3	**追加**	確定	健康診断	母

C　操作種別の分布

第 4.2 節で用いた操作種別ごとの頻度を表 11 に示す．**追加**と**更新**の頻度は同程度であるが，複数操作は極めて少ない．

表 11　操作種別ごとの頻度.

追加	更新	追加+更新	追加+追加+追加	更新+更新+更新	更新+更新
42	37	2	1	1	1