行動認識の粒度アライメントに基づく予定の履行認識

藤田一天 [1]　河野誠也 [2,1]　吉野幸一郎 [3,2,1]
[1] 奈良先端科学技術大学院大学　情報科学領域　[2] 理化学研究所 GRP
[3] 東京科学大学　情報理工学院
fujita.kazuma.fm0@is.naist.jp　seiya.kawano@riken.jp
koichiro.yoshino@riken.jp

概要

キャプショニングなどを用いた状況認識により，動画内のユーザ行動を自然言語で表現することが可能になった．しかし，このように認識されたユーザー行動を事前に定義された行動予定と比較して履行判断に用いようとする場合，テキストで表現される行動の粒度が問題となる．そこで，本研究では言い換えモデルを用いてテキストで認識した動画内のユーザー行動を言い換え，行動予定に対してアライメントを取るモデルを提案する．言い換えによって生成された行動とあらかじめ定義された予定行動を自動評価手法によって比較し，履行が判断できる閾値を設定することで予定の履行を適切に評価できることを確認した．

1　はじめに

現代社会では，個人や組織が予定行動を達成するために，計画を練り，それを実践するプロセスが重要視される．このプロセスにおいて，予定行動を適切に管理し，実行状況を評価することは不可欠である．個人が行動を達成したかどうかは，動画からのキャプショニング技術などを用いることにより，自然言語で把握することが可能になった．これにより行動を計画通り履行したかどうかを確認できるようになることが期待される．

ただし，キャプショニングにより認識されたユーザー行動を人が作成した予定行動とを比較し，自動的に履行を判断することは容易ではない．主な理由として，両者の粒度の違いが挙げられる．同じ内容を説明してもその粒度が異なったり，あるいは行動の一部のみを説明しているキャプションが得られるようなことが起こり，予定の履行判断位は困難が生じる．

図 1: 動画内のユーザー行動のキャプショニングと予定行動の比較（行動の粒度が異なるため，履行判断が困難）

そこで本研究では，まず言い換えモデルによってキャプショニングで得られたユーザの複数の行動を予定と同等の粒度に言い換える．さらに，言い換えた結果を予定行動との比較を意味的に行い，履行判断を行うための手法を検討する．この中で，様々な評価指標を検証し，またそれらを履行判断に用いる時の閾値について検討する．

2　タスク

2.1　データ:Ego4D

Ego4D[3] は，動画内の人間の行動をシステムが説明するタスクを目的としたデータセットであり，人間の一連の動作を詳細にキャプショニングした結果が含まれる．基本的には，動画内の情報を基に人間の行動を漏れなく記述し，個々の動作を適切に説明するようにキャプションが付与されている．本研究では Ego4D で付与されている詳細なキャプション文をキャプショニングで得られる行動結果と仮定

し，これらを人間が設定した予定行動とを比較することで，適切な履行判断を行うことを目的とする．

2.2 要約による曖昧性解消

Ego4D によって生成されるキャプション文では，ユーザー行動が非常に細かい粒度で記述されている[8]．本研究では，このように粒度が細かく記述されたユーザー行動と人間が設定した予定行動の比較を行う際，表現方法の不十分さが引き起こす課題に着目している．この課題に対応するため，ユーザー行動の言い換えを適用するアプローチを採用した．本研究は，言い換えを通じて細分化された行動間のアライメントを取り，予定行動との比較を可能にする．具体的には，Ego4D のデータセットに含まれる粒度の細かいキャプション文を大規模言語モデル[1)]により要約して用いる．

2.3 履行判断の自動化

要約により動作のキャプショニング結果を抽象化することで，ユーザが作成した行動予定との比較を可能にする．ただし，実際に動作が履行されたかどうかを自動判断するためには，何らかの自動化指標が必要である．

2.4 タスク設定

ここまで説明した要約と履行判断について，本研究では以下のタスクを設定する．

履行判断タスク　動画内のユーザー行動を要約したキャプション文と予定行動を比較し，粒度が揃っているか，さらに履行したと判断できるほど同じ意味を持つかという 2 つの軸で自動的に評価することを目的としたタスクである．この評価を実現するために，適切な評価指標と閾値を見つけることが重要である．

粒度アライメントタスク　行動予定と比較する上で，大規模言語モデルに与える最適なプロンプトを見つけることを目的としたタスクである．

3 実験

本節では，履行を自動的に判断する評価指標と閾値決定の方法と粒度アライメントを取るための要約

1) ChatGPT（GPT-3.5 Turbo）を使用した

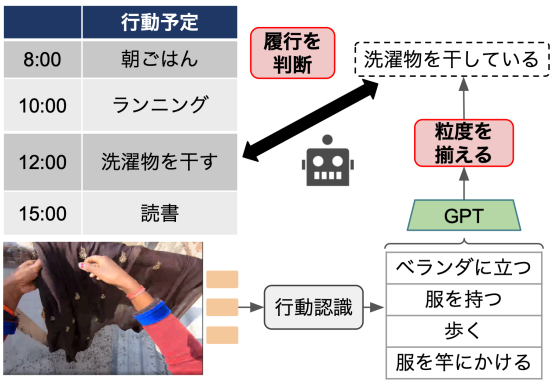

図 2: 提案モデルの概要図．行動予定と比較困難だったキャプション文を言い換えることで比較を可能にする．

作成のプロセスを述べる．

3.1 提案モデル

提案するモデルの構成を図 2 に示す．本モデルは，Ego4D データセットに含まれる粒度の細かいキャプション文を予定行動と比較し，履行判断を行うことを目的として設計されている．モデルの設計では，粒度の細かいキャプション文を予定行動と比較可能な形に言い換えるため，キャプション文を要約する手法を採用している．具体的には，まず Ego4D から抽出したキャプション文の系列を大規模言語モデルに入力し，要約を生成する．その後，生成された要約文と行動予定との比較を行う．これにより，テキストで認識された粒度が細かい複数の行動が統合され，予定行動に基づいた履行判断を可能にする．また，要約プロセスを通じて生成される文の正確性を保証するため正答率を測定し，その結果に基づいてモデルの性能を最適化する．

3.2 テストセット

本研究では，履行判断を自動的に行う評価指標とその閾値を決定するためのデータとして，動画データ 150 件のうち 50 件を開発データとして利用した．一方，残りの 100 件をテストデータとして用いた．テストデータは，設定した評価指標を用いて履行判断の精度を評価するためのデータであり，開発データで調整した閾値や指標の妥当性を確認する目的で使用する．開発データおよびテストデータに対して

は，すべてアノテータの作業によって正解となる回答を作成した．

3.3 履行判断評価指標と閾値

履行判断タスクでは，粒度アライメントに用いることができる評価指標の検討を行う．実験は以下の手順により実施する．

- Step1: Ego4D に含まれる動画をもとに，アノテータ代表者が予定行動を作成する．
- Step2: 動画に付与されているユーザー行動のキャプション文を ChatGPT を用いて要約する．
- Step3: 要約された行動キャプション文と予定行動の履行判断をアノテータ代表者が評価する．
- Step4: 類似度を評価できる評価指標を用いて Step3 と同様の履行判断を行う．
- Step5: 人と近い判断ができる評価指標とその閾値を見つける．

3.4 予定行動のデータ収集

本研究では予定行動データが存在しないため，新たにデータを人手で作成する．具体的には， Ego4D データセット内に含まれる動画を対象とし，その内容を基に予定行動を作成した．この方法により， 細分化されたキャプション文と予定行動を比較する枠組みが構築可能となる．しかし，同一の動画に対しても作成者ごとに異なる解釈が生じる可能性がある．例えば，同じ行動を「買い物に行く」と記述するか「スーパーで食材を選ぶ」と記述するかは，作成者の視点や解釈に依存する．これにより，データの一貫性や信頼性に影響を与える可能性がある点が課題として挙げられる．また，履行判断を行うことができる評価指標と閾値を決定する実験においても，人に近い判断ができるものを特定することが目標である．そのため，予定行動と要約文を用いて人が履行判断を行う必要があるが，この準備段階においても主観的なばらつきを緩和する必要がある．本研究ではこれらの課題を考慮し，予定行動データの信頼性と実用性を確保する．

表 1: アノテータ間の主観性評価

項目	文平均類似度	判断一致度	サンプル数
評価結果	0.8066	88%	50
中央値	0.7678	-	50
標準偏差	0.0418	-	-

表 2: 各評価指標の AUC 比較結果

要約例	行動予定例	評価指標	AUC
キッチンに立つ	料理をする	BLEU	0.82
		METEOR	0.85
		SentenceBERT	0.91
		BERT Score	0.84

3.5 予定行動データのばらつき

予定行動の作成において，アノテータ間の主観的なばらつきが課題として挙げられる．本研究では複数名のアノテータに予定行動の作成を行ってもらい，そのばらつきについて評価した．具体的には，BERT ベースの類似度スコアを用いて，別々のアノテータが作成した予定行動文の類似度を定量的に評価した．また，履行判断における一致度は，異なるアノテータが同じ判断をした割合で測定した．これにより，予定行動の記述および判断が主観的でないことを示すことを目指した．結果を表 1 に示す．評価の結果，代表者が作成した予定行動や判断が，他のアノテータと比較して主観的ではないことが確認された．本研究は，データ構築における公平性と一貫性を担保する基盤を提供するものであり，表 1 に示す通り予定行動作成の基準化に向けた重要なステップとなる．

3.6 最適な評価指標と閾値の決定

最適な評価指標と閾値を決定するため，文の類似度を定量評価できる 4 つの評価指標（BERT Score・SentenceBERT・ BLEU・METEOR）[10, 6, 7, 1] を用いて履行判断を行った．それぞれの評価指標を用い，ランダムに設定された閾値のもとで比較を行い，50 件のデータを対象に積み上げ式で評価を実施した．積み上げ式では評価対象データを 1 件ずつ増やし，その過程で精度の変化を観察した．

表 3: プロンプト別履行判断の正答率比較

	正答率
要約せずに履行判断	0%
何をしているか一言で教えてください	78%
要約してください 出力例) 食事をする	88%

4 実験結果

4.1 履行判断ができる評価指標

人に近い履行判断ができる評価指標を知るために，人が行った履行判断と各指標の結果を比較し，Precision[4] を基準としてもっとも良い自動評価指標を調査した．その結果を表 2 に示す．AUC[5] が一番高い，すなわち安定して人に近い判断ができている SentenceBERT が本タスクにおける最適な評価指標であることが明らかになった．閾値の選定については，Precision と Recall のバランス [2] を考慮し，両者のグラフが交差する点を採用した．

4.2 ベースラインと提案モデル比較

履行判断の正答率 (Acc) をモデルの評価に用いた (式 1).

$$Acc = \frac{cnt}{100} \tag{1}$$

cnt は回答セットの 100 件の人手判断と予測判断が完全一致した回数である．

表 3 に，実際に履行判断を試みた際の正答率を示す．その結果，出力例を交えたプロンプト [9] が最も高い精度を示したことが確認された．特に，出力例を予定行動と同程度の粒度で提供することで，ChatGPT が行動キャプション文を生成する際，予定行動に近い粒度で正確な要約を行えるようになったと考えられる．考察として，プロンプト設計がモデルの精度向上に大きく寄与する可能性が示唆された．特に，具体的かつ明確な出力例を用いることで，モデルが生成するテキストの関連性と正確性が高まることが確認された．今後の研究では，出力例の多様化や他の手法との組み合わせによって，さらに汎用性の高いプロンプト設計が可能になると期待される．

5 おわりに

本研究では，予定行動と認識されたユーザー行動の比較において，粒度を考慮した履行判断の評価指標を提案し，キャプショニング技術を活用した要約モデルを構築した．Ego4D データセットを用いた実験により，主観的なばらつきを抑えつつ，自動的な履行判断が可能であることを示した．

謝辞

本研究の一部は JST が JPMJPR24TC の支援を受けた．本研究において多大なるご支援をいただきました皆様に，心より感謝申し上げます．貴重なアドバイスをいただいた指導教員の先生方に深く感謝いたします．さらに，実験データの収集や解析に協力してくださった皆様にも感謝申し上げます．本研究を進めるにあたり，技術的なサポートをいただいた方々にも感謝の気持ちを伝えたいと思います．

参考文献

[1] Patrick Behm, Paul Benoit, Alain Faivre, and Jean-Marc Meynadier. Meteor: A successful application of b in a large project. In **International Symposium on Formal Methods**, pages 369–387. Springer, 1999.

[2] Kendrick Boyd, Kevin H Eng, and C David Page. Area under the precision-recall curve: point estimates and confidence intervals. In **Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13**, pages 451–466. Springer, 2013.

[3] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pages 18995–19012, 2022.

[4] Michael Kane. The precision of measurements. **Applied measurement in education**, 9(4):355–379, 1996.

[5] Shahzad Ali Khan and Zeeshan Ali Rana. Evaluating performance of software defect prediction models using area under precision-recall curve (auc-pr). In **2019 2nd International Conference on Advancements in Computational Sciences (ICACS)**, pages 1–6. IEEE, 2019.

[6] Guangyuan Piao. Scholarly text classification with sentence bert and entity embeddings. In **Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25**, pages 79–87. Springer, 2021.

[7] Matt Post. A call for clarity in reporting bleu scores. **arXiv preprint arXiv:1804.08771**, 2018.

[8] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. **Advances in Neural Information Processing Systems**, 36, 2024.

[9] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. **arXiv preprint arXiv:2302.11382**, 2023.

[10] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. **arXiv preprint arXiv:1904.09675**, 2019.