大規模言語モデルによるイベント知識グラフからのマルチターン few-shot 実況生成手法の検討

辻村 有輝　江上 周作　浅田 真生　石垣 達也　福田 賢一郎　高村 大也
産業技術総合研究所
{tsujimura.res,s-egami,masaki.asada,ishigaki.tatsuya,
ken.fukuda,takamura.hiroya}@aist.go.jp

概要

本研究ではイベント知識グラフからの実況テキスト生成に取り組む．タスク設定はシナリオの進行に合わせて実況を生成できるようマルチターン形式で定義する．グラフ情報はテキスト形式へ変形し，大規模言語モデルを用いて few-shot プロンプトによる生成を行う．また，我々は既存のイベント知識グラフを備えた動画データセットに対して，日本語の実況音声とテキストを新たに付与することで 4 種のデータ形式を備える動画実況データセットを構成し実験に利用した．結果として，詳細なグラフ情報の利用により BLEU スコアの向上を確認したものの，言語モデルの最大系列長制限のため正解実況で言及されることのある情報の一部は入力中に含められなかった．データセットは一般公開されている．[1)]

1　はじめに

様々な自然言語処理タスクでの活用可能性から，大規模言語モデルに対する研究が盛んに取り組まれている [1, 2, 3, 4, 5, 6, 7, 8]．特に最近では，一層の活用範囲拡大のため，自然言語テキスト以外の形式のデータを扱う能力を兼ね備えたマルチモーダル大規模言語モデルの開発が試みられている [9]．

実世界にはソースコードのリポジトリやウェブページの DOM ツリー，専門知識データベースといった，グラフ構造で表現される情報が多数存在する．高いグラフ理解能力を兼ね備えた大規模言語モデルの実現は，これらの情報の利活用のために重要である．言語モデルでグラフを扱う研究も取り組まれているものの，言語モデルに比べてグラフエンコーダーの学習データ規模が限られていたり [10]，グラフ構造を直接扱わず画像として解釈する [9] など，未だ発展途上である．

本研究ではグラフからのテキスト生成タスク手法の検討を行うことで，グラフ情報のより高度な理解能力を兼ね備えた大規模言語モデル実現への足掛かりとする．また我々は，仮想空間上でシミュレートされた CG 動画と動画中に発生したイベントを表現するイベント知識グラフからなる既存データセット [11] に，動画内容を説明する実況音声とその文字起こしテキスト，並びにその動画の概要テキストを新たに付与することで，動画，グラフ，音声，テキストの 4 種の形式のデータを備えるマルチモーダル動画実況データセットを構成した．本研究ではこのうちイベント知識グラフと実況の文字起こしテキストを用いて，グラフからの実況テキスト生成タスクに取り組む．

本研究ではグラフからの実況テキスト生成をマルチターン形式のタスクとして定義し，これを解くための大規模言語モデルによるマルチターン実況テキスト生成手法を提案する．言語モデルでのグラフ情報の利用のため，グラフをテキスト形式へ変形 [12] し，few-shot プロンプトで入力を構成する．実験では与えるイベント情報の詳細化により BLEU スコアの向上を確認した一方で，言語モデルの最大系列長制限からイベント知識グラフ中の一部の情報のみの利用に留めなければならず，オブジェクト間の位置関係のような，実況アノテーション上で言及されることがある情報を入力に含められなかった．

2　関連研究

大規模言語モデルは大規模コーパスを用いて学習された巨大なニューラル言語モデルであり，与えられた入力文脈を通じて追加の学習なしにタスクに適応できる [6, 7, 8]．大規模言語モデルにタスクを実行させるときに与える入力文脈はプロンプトと呼ば

1） https://kirt.airc.aist.go.jp/corpus/ja/VirtualHomeCommentary

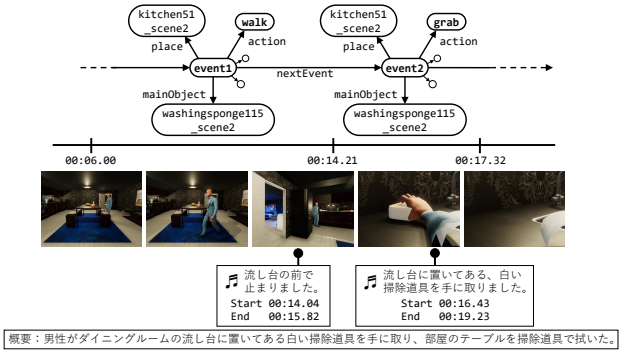

図 1　本研究で利用した CG 動画実況データセット．動画，イベント知識グラフ，実況音声とその書き起こしテキスト，動画概要説明テキストから構成される．簡単のためノードや関係名は省略して表記した．

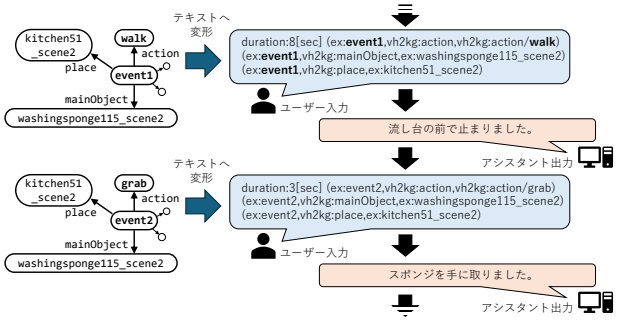

図 2　マルチターン形式によるグラフからの実況テキスト生成手法の概観．各ターンではこれまでの経過を元に，最新のイベントに対する実況を生成する．

れ，タスク説明や，few-shot 事例と呼ばれる参考入出力例，実際に解きたいタスク事例などを記述し，その続きを言語モデルとして生成させることでタスクを解く．

3　イベント知識グラフからの実況生成

3.1　データセット

我々は，仮想環境 VirtualHome-AIST [13] 上でシミュレートされた CG 動画データセット [11] に対し，日本語実況音声とその書き起こしデータ，および動画の概要を説明するテキストを人手で追加付与することにより，動画，グラフ，音声，テキストの 4 種の形式のデータを備えたデータセットを構成し，実験で使用した．データセット中に含まれる事例を図 1 に示す．各 CG 動画には別々のシナリオに対する人間エージェントによる動作の様子が描画されている．エージェントは一つのシナリオ内で複数の動作を連続して行い，一つ一つの動作の継続時間は数秒程度である．各シナリオごとに，行われた一連の動作や環境内のオブジェクトの情報を表すイベント知識グラフも提供される．グラフ内では一つの動作が一つのイベントノードとして存在し，イベントの連なりからシナリオ全体が表現される．グラフは RDF 形式のデータであり，ノードや関係は一意な URI で表現される．

実況データは人手で付与された日本語の実況音声と，その書き起こしテキストから構成される．実況音声は動画のみに基づいて人手で作成され，アノテーターは動画中で何が起こっているかを把握できるような実況を行うこと，また人間の動作以外でも気になった点について言及するよう指示された．書き起こしテキストには各発話ごとのその開始と終了のタイムスタンプも含まれる．概要テキストも動画のみに基づいて人手で作成されており，その動画中で何が起きたかを説明する 50 文字程度のテキストである．

3.2　タスク設定

本研究では 3.1 節のデータセットのうちイベント知識グラフと実況の文字起こしテキストを用いて，グラフからの実況生成タスクに取り組む．ここではシナリオの進行に合わせた実況生成が行えるよう，マルチターン形式でタスクを設定する．

実況システムには実況対象シナリオ中の各イベントを表現するサブグラフが発生した順に連続して与えられ，その都度そのイベント発生中に発話が開始された実況のテキストを出力することとする．元々のデータセット中では実況の発話は文単位でタイムスタンプが付与されており，一つのイベント発生中に複数文の実況が発話されている場合は連結して一つの実況区切りとする．

3.3　実況生成手法

大規模言語モデルを用いたマルチターン形式によるグラフからの実況テキスト生成手法を提案する．図 2 に実況生成手法の概観を示す．生成はイベントの発生ごとに行うマルチターン方式とすることで，シナリオ進行に合わせた実況生成を行う．グラフ情報はテキスト形式へ変形し入力とする．言語モデルの追加の学習は行わず，他シナリオと実況の例を与える few-shot プロンプトによる生成を行う．

言語モデルへ与える入力プロンプトは，前から順にタスク説明部，few-shot 事例の記述部，対象シナリオの記述部の三つから構成される．入力プロント

の先頭部分を構成するタスク説明部には，実況生成を行う旨をシステムプロンプトの形で記す．使用する言語モデルが固有の先頭システムプロンプトを持つ場合，タスク説明の前にそれを付け加える．

few-shot 事例の記述部では，ランダムに選んだシナリオ中の全イベントのテキスト形式の入力と，対応する実況テキストが参考例として対話形式で提示される．1 回の対話ターンは入力である一つのイベントと応答である実況 1 区切りからなり，シナリオ中の各イベントのターンの繰り返しによってシナリオ全体の実況が構成される．各イベントの情報はグラフ情報から作成したテキストで表現される．具体的には，イベントの継続時間を示す “duration:N[sec]” (N は継続時間を整数に丸めた値) に，対象イベントに関するグラフ中のトリプルを “(主語, 述語, 目的語)” の形で並べたものを連結したテキストであり，ユーザープロンプトの形で記す．イベント知識グラフ中のトリプル数は非常に多いため，言語モデルの最大系列長の都合から，トリプル列の記述には特に重要なトリプルとして対象イベントを主語とするトリプルの一部のみを使用する．各イベントに対応する実況はそのイベントの発生中に発話が開始された実況文であり，アシスタントプロンプトの形で記す．参考例となるシナリオの数はハイパーパラメータであり，1 シナリオあたり 1 shot と数える．各シナリオの始まりには言語モデルがシナリオの区切りを捉えやすくなるようにシステムプロンプトの形で “新しい状況の開始” のテキストを付け加える．

対象シナリオの記述部は few-shot 事例と同一のフォーマットであるが，実況生成対象のイベントに対するグラフ情報の記載とそれに対するアシスタントターン開始プレフィックスまでで留め，その続きの生成により実況生成を行う．

その他の手法の詳細は付録 A に記載した．

4 実験と結果

4.1 実験設定

サブグラフ上のトリプルすべてを利用すると言語モデルの最大系列長を超過するため，入力には特に重要なトリプルとしてイベントを主語とする一部の関係の組み合わせを 5 パターン用意しそれぞれ実験を行った．表 1 に使用した関係トリプルの組み合わせを示す．実験で使用したモデルの一覧を表 2 に示す．URI 中の共通部分は接頭辞で置き換え短縮した．また，URI 中にシナリオの識別名が含まれる場合はその部分を除去した．Few-shot 事例として与えるシナリオ数は 0, 1, 3, 5 シナリオの 4 パターンで実験を行った．1 ターンの最大生成トークン数は 256 トークンとし，最大に達した場合は強制的にターン終了とした．生成はイベントごとに個別に行い，過去イベントに対応する実況は常に正解の実況を入力プロンプトで使用した．

表 1 入力に使用した関係トリプルの組み合わせ．“action” は動作内容を，“mainObject” は動作の目的語を，“targetObject” は追加の目的語を表す．“place”, “to”, “from” はいずれも場所を表す．

action + mainObject
action + mainObject + targetObject
action + mainObject + place + from + to
action + mainObject + targetObject + place + from + to
(全ての関係)

評価は文字レベル BLEU-4 と単語レベル BLEU-4 を用いた．単語レベルのスコアについては異なる辞書を用いた単語分割により 2 パターン計測した．単語分割は MeCab によって行い，辞書には IPA 辞書および neologd 辞書を使用した．一つのターンの実況ごとに 1 文書として扱った．

4.2 結果

各モデルについて最も良い単語レベル BLEU スコアを得た設定とそのスコアを表 2 に示す．本研究では 3 パターンの BLEU-4 スコアを計測しているが，最高のスコアとなる設定は多くのモデルで BLEU スコアの計測方法によらなかったため，ここでは IPA 辞書による単語レベル BLEU スコアが最大となった設定とそのスコアのみを表示している．実施した実験結果の詳細は付録 B に記載した．

Llama 3.1 の指示チューニング済みモデルは非日本語特化でありながら，Swallow モデルを除き今回実験した日本語特化モデルのうち最も BLEU スコアがよい Calm3 モデルと同等の性能となっており，Llama 3.1 モデルに継続事前学習を施し日本語性能が向上した Swallow の指示チューニング済みモデルは実験中最高の BLEU スコアを記録した．

全モデル共通の傾向として，基本的に shot 数が多い時に BLEU スコアが高く，またサブグラフの情報も action と mainObject 関係のトリプルだけでなく追加の情報も与えたほうが良い結果となった．例外として，LLM-jp-3 モデルはサブグラフとして与える情報を増やしたり shot 数を増加させると最大入力長を

表 2 各モデルの IPA 辞書による単語レベル BLEU 最高となった設定とそのスコア．Swallow と指示チューニングありの Qwen の最大系列長はベースとなったモデルのものである．“place”，“from”，“to” 関係はまとめて “location” と記した．

モデル	日本語特化	モデルサイズ	最大入力長	指示Tuning	入力 使用した関係トリプル	入力 shot 数	単語レベル BLEU-4
LLM-jp-3 [1]	✓	13B	4,096	✓	action+mainObject+targetObject	1	18.69
Calm3 [2]	✓	22B	16,384	✓	action+mainObject+targetObject	5	23.31
Calm2 [3]	✓	7B	32,768	✓	action+mainObject+targetObject+location	3	16.34
Llama 3.1 Swallow v0.2 [4, 5]	✓	8B	131,072		action+mainObject+targetObject+location	5	20.68
	✓	8B	131,072	✓	action+mainObject+targetObject+location	5	**26.11**
Llama 3.1 [6]		8B	131,072		action+mainObject+location	5	17.48
		8B	131,072	✓	action+mainObject+targetObject+location	5	23.13
Qwen2.5 [7]		7B	131,072		全て	3	17.01
		7B	131,072	✓	全て	5	20.15
Qwen2 [8]		7B	131,072		全て	5	15.29
		7B	131,072	✓	action+mainObject+targetObject	5	17.07

表 3 最高の BLEU スコアとなった設定の指示チューニング済み Swallow モデルにおけるシナリオ “clean_kitchentable1_scene2” に対する出力．イベント発生順に入力に用いられた情報とそれに対する実況を示している．このシナリオでは全イベントに述語が “kitchen51_scene2” の place 関係も存在し，入力として与えられている．

発生順	継続時間	関係ごとの述語 action	mainObject	targetObject	実況テキスト モデル出力	正解
1	6 秒	walk	kitchentable129_scene2		男性がダイニングキッチンを歩いています．テーブルの前で止まりました．	男性が部屋を歩いています．ダイニングルームです．
2	8 秒	walk	washingsponge115_scene2		歩いています．流し台の前で止まりました．	流し台の前で止まりました．
3	3 秒	grab	washingsponge115_scene2		スポンジを手に取りました．	流し台に置いてある，白い掃除道具を手に取りました．
4	8 秒	walk	kitchentable129_scene2		歩いています．ダイニングルームのテーブルの前で止まりました．	歩きだしました．
5	19 秒	wipe	washingsponge115_scene2	kitchentable129_scene2	テーブルの上を拭き始めました．	部屋の中央辺りにあるテーブルの前で止まりました．テーブルを掃除道具で拭いています．拭き終わりました．

超過し出力の崩壊を引き起こすため，shot 数が 1 のとき最も高い BLEU スコアとなっている．

指示チューニングを施したモデルは全てベースとなったモデルより BLEU スコアが高くなった．これは今回用いたプロンプトにタスクの指示が含まれており，指示チューニングがタスク理解に有利に働いた結果と考えられる．また，関係トリプルを並べた非自然言語テキストに続いて実況という自然言語テキストが出現する文脈は，事前学習で遭遇する一般的なテキストではあまり現れないと考えられるが，指示チューニング学習時に遭遇するソースコード理解の質問などはより近い形の入力であり，その点でも有利である可能性がある．

Swallow の指示チューニング済みモデルで最高の BLEU スコアを記録した設定における実際の出力例を表 3 に示す．表中の出力の列の通り，モデルはグラフ情報に基づく実況を生成できていることが確認された．多くの実況では action や mainObject 関係が重要な情報源となる一方で，最初の実況にある位置情報や最後の実況の道具の情報など，他のトリプルが必要になるケースも存在する．今回の実験では言語モデルの最大系列長制限のためオブジェクト間の位置情報に関するトリプルは利用できなかったが，最後のイベントの正解実況のようにその情報が言及されるケースもあり，言語モデルでの効率の良いグラフ処理方法の検討は今後の課題である．

5 おわりに

本研究では大規模言語モデルによるグラフからの実況テキスト生成タスクに取り組んだ．実験では我々が構成したテキスト，グラフ，動画，音声の 4 種の形式のデータを備えたマルチモーダル動画実況データセットを使用した．結果として詳細なグラフ情報を用いることで BLEU スコアの向上が確認された．今後の課題として言語モデルにおけるグラフ情報利用の効率化が挙げられる．

謝辞

この成果は, 国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の助成事業 (JPNP20006) による支援および産総研政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」の結果得られたものである.

参考文献

[1] 大規模言語モデル研究開発センター. LLM-jp-3 1.8B・3.7B・13B の公開, 2024. https://llmc.nii.ac.jp/topics/post-707/.

[2] Ryosuke Ishigami. CyberAgentLM3, 2024. https://huggingface.co/cyberagent/calm3-22b-chat.

[3] Ryosuke Ishigami. CyberAgentLM2, 2023. https://huggingface.co/cyberagent/calm2-7b-chat.

[4] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[5] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[6] Llama Team. The llama 3 herd of models, 2024. arXiv:2407.21783.

[7] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. arXiv:2412.15115.

[8] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. arXiv:2407.10671.

[9] Yihe Deng, Chenchen Ye, Zijie Huang, Mingyu Derek Ma, Yiwen Kou, and Wei Wang. Graphvis: Boosting LLMs with visual knowledge graph integration. In **The Thirty-eighth Annual Conference on Neural Information Processing Systems**, 2024.

[10] Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. GreaseLM: Graph REASoning enhanced language models. In **International Conference on Learning Representations**, 2022.

[11] Shusaku Egami, Takanori Ugai, Mikiko Oono, Koji Kitamura, and Ken Fukuda. Synthesizing event-centric knowledge graphs of daily activities using virtual space. **IEEE Access**, Vol. 11, pp. 23857–23873, 2023.

[12] Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. Investigating pretrained language models for graph-to-text generation. In Alexandros Papangelis, Pawe l Budzianowski, Bing Liu, Elnaz Nouri, Abhinav Rastogi, and Yun-Nung Chen, editors, **Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI**, pp. 211–227, Online, November 2021. Association for Computational Linguistics.

[13] Shusaku Egami, Takanori Ugai, Swe Nwe Nwe Htun, and Ken Fukuda. Vhakg: A multi-modal knowledge graph based on synchronized multi-view videos of daily activities. In **Proceedings of the 33rd ACM International Conference on Information and Knowledge Management**, CIKM '24, p. 5360–5364, New York, NY, USA, 2024. Association for Computing Machinery.

表 4　実験を行った設定全ての BLEU スコア．各設定について単語レベル BLEU-4，IPA 辞書での単語分割による単語レベル BLEU-4，neologd 辞書での単語分割による単語レベル BLEU-4 をスラッシュで区切って表示した．太字のスコアは各モデルごとの最高の BLEU スコアである．

モデル	関係トリプル	shot 数			
		0	1	3	5
llm-jp-3-13b-instruct	action+mainObject	22.95/14.58/14.45	28.73/18.60/18.63	28.45/18.63/18.62	17.90/10.92/10.88
	action+mainObject+targetObject	23.17/14.68/14.56	**28.82/18.69/18.70**	27.94/18.21/18.21	16.71/10.04/9.96
	action+mainObject+location	26.02/16.18/16.03	28.42/18.19/18.16	17.96/13.60/13.60	5.29/3.18/3.21
	action+mainObject+targetObject+location	26.23/16.34/16.18	28.56/18.33/18.28	19.67/13.30/13.30	6.29/2.93/2.94
	全て	9.33/7.86/7.65	14.11/6.82/6.78	0.47/0.00/0.00	0.00/0.00/0.00
calm3-22b-chat	action+mainObject	23.39/15.59/15.36	30.26/19.75/19.68	32.06/22.02/21.93	32.58/22.79/22.46
	action+mainObject+targetObject	23.66/15.81/15.56	30.52/20.28/20.20	32.34/22.49/22.38	33.06/**23.31/22.91**
	action+mainObject+location	25.48/16.72/16.45	30.13/19.62/19.52	32.11/21.70/21.45	**33.13**/22.86/22.47
	action+mainObject+targetObject+location	25.64/16.77/16.50	30.12/19.49/19.39	32.41/22.19/22.02	33.12/23.11/22.70
	全て	17.27/12.26/12.06	29.99/19.19/19.08	31.95/21.87/21.52	32.38/21.85/21.54
calm2-7b-chat	action+mainObject	21.55/12.83/12.87	25.03/15.32/15.42	25.53/15.72/15.78	25.79/15.90/15.92
	action+mainObject+targetObject	21.71/12.87/12.93	25.17/15.43/15.49	25.68/15.89/15.94	25.96/16.17/16.23
	action+mainObject+location	22.49/12.81/12.85	24.68/14.59/14.67	25.59/15.58/15.63	25.82/15.88/15.98
	action+mainObject+targetObject+location	22.86/13.11/13.18	25.02/14.80/14.92	**26.05/16.34/16.43**	26.03/16.22/16.30
	全て	13.67/8.96/8.96	22.53/12.10/12.11	24.32/14.19/14.19	24.44/14.32/14.40
Llama-3.1-Swallow-8B-v0.2	action+mainObject	24.19/16.67/16.56	26.24/17.52/17.05	27.50/18.91/18.44	27.13/18.61/18.04
	action+mainObject+targetObject	24.28/16.71/16.58	25.83/17.41/16.97	28.54/19.71/19.20	26.81/18.54/17.99
	action+mainObject+location	23.95/16.39/16.19	25.02/16.75/16.31	29.18/20.47/**19.78**	29.11/20.32/19.41
	action+mainObject+targetObject+location	24.07/16.57/16.38	24.81/16.57/16.09	29.07/20.27/19.50	**29.59/20.68**/19.77
	全て	21.94/15.85/15.61	26.18/17.55/17.02	25.87/17.71/17.04	25.55/17.57/16.78
Llama-3.1-Swallow-8B-Instruct-v0.2	action+mainObject	27.25/18.30/18.10	31.29/20.97/20.41	33.86/23.29/22.64	34.02/23.04/22.41
	action+mainObject+targetObject	27.73/18.70/18.51	31.71/21.45/20.92	34.68/24.12/23.53	35.62/24.59/23.96
	action+mainObject+location	28.72/19.21/18.95	33.46/22.91/22.29	35.40/24.50/23.73	36.12/25.25/24.30
	action+mainObject+targetObject+location	28.93/19.42/19.15	33.82/23.25/22.68	35.93/24.98/24.25	**37.12/26.11/25.22**
	全て	22.71/15.04/14.75	34.39/23.99/23.40	35.13/24.44/23.69	35.68/24.99/24.16
Llama-3.1-8B	action+mainObject	17.22/11.31/11.15	21.24/13.72/13.22	23.22/15.01/14.56	23.45/16.41/15.72
	action+mainObject+targetObject	17.50/11.61/11.43	19.72/12.77/12.29	22.74/14.83/14.37	22.40/15.23/14.53
	action+mainObject+location	14.89/9.75/9.53	21.59/14.10/13.50	23.42/15.57/14.91	**24.46/17.48/16.66**
	action+mainObject+targetObject+location	14.85/9.80/9.58	21.45/14.27/13.71	22.85/15.15/14.51	23.92/16.54/15.72
	全て	17.61/12.24/12.01	20.78/13.70/13.26	21.59/14.45/13.73	22.63/14.79/14.07
Llama-3.1-8B-Instruct	action+mainObject	26.27/16.52/16.08	29.87/18.97/18.39	31.69/20.97/20.29	33.87/22.54/21.76
	action+mainObject+targetObject	26.40/16.60/16.21	29.79/19.13/18.58	31.97/21.20/20.54	33.51/22.50/21.77
	action+mainObject+location	25.42/16.23/15.88	30.14/19.48/18.83	31.62/20.90/20.11	**34.02/23.13/22.24**
	action+mainObject+targetObject+location	25.31/16.13/15.80	30.08/19.57/18.91	32.38/21.68/20.96	33.95/22.90/21.99
	全て	25.66/16.71/16.37	30.69/20.00/19.76	31.56/21.03/20.33	33.19/22.06/21.24
Qwen2.5-7B	action+mainObject	18.12/11.07/10.89	20.29/12.39/12.01	20.69/12.69/12.30	22.16/13.84/13.26
	action+mainObject+targetObject	18.20/11.26/11.09	20.67/12.60/12.23	20.60/12.70/12.27	23.00/14.37/13.84
	action+mainObject+location	19.71/12.46/12.15	21.98/13.75/13.31	22.93/14.41/13.74	23.00/14.62/13.92
	action+mainObject+targetObject+location	19.67/12.44/12.13	21.83/13.75/13.30	22.56/13.97/13.35	23.17/14.88/14.15
	全て	16.43/9.95/9.66	23.74/14.63/14.20	**26.92/17.01/16.28**	25.76/16.80/16.04
Qwen2.5-7B-Instruct	action+mainObject	16.85/9.91/9.71	25.02/15.15/14.73	27.52/17.12/16.60	28.30/17.72/17.24
	action+mainObject+targetObject	17.13/10.16/10.01	25.13/15.37/14.98	27.69/17.26/16.77	28.66/18.19/17.69
	action+mainObject+location	16.49/9.81/9.60	25.85/15.84/15.43	28.55/17.85/17.30	29.96/18.81/18.24
	action+mainObject+targetObject+location	16.73/10.06/9.84	26.00/16.03/15.53	28.96/18.32/17.76	30.11/19.33/18.71
	全て	15.14/8.45/8.23	26.05/16.05/15.49	29.80/19.09/18.45	**31.74/20.15/19.45**
Qwen2-7B	action+mainObject	15.16/9.78/9.52	20.23/12.39/12.00	20.18/12.17/11.77	21.56/13.25/12.75
	action+mainObject+targetObject	15.32/9.92/9.67	21.85/13.44/13.02	21.03/12.67/12.24	21.48/13.19/12.72
	action+mainObject+location	14.99/9.87/9.56	20.35/12.30/11.84	22.08/13.42/12.80	23.37/14.62/13.91
	action+mainObject+targetObject+location	14.77/9.69/9.39	20.72/12.59/12.10	22.98/14.11/13.48	23.83/14.88/14.14
	全て	10.30/7.17/6.91	19.87/12.11/11.68	22.55/13.71/13.17	**24.59/15.29/14.62**
Qwen2-7B-Instruct	action+mainObject	19.02/10.26/9.95	22.31/12.68/12.35	25.42/15.14/14.70	26.95/16.58/16.08
	action+mainObject+targetObject	18.94/10.29/10.00	22.50/12.77/12.43	25.38/15.08/14.65	27.69/**17.07/16.55**
	action+mainObject+location	16.92/9.14/8.89	22.04/12.30/11.95	25.37/15.34/14.94	26.83/16.48/15.99
	action+mainObject+targetObject+location	17.14/9.35/9.10	22.50/12.49/12.17	25.46/15.34/14.99	27.05/16.50/16.03
	全て	15.93/8.86/8.68	23.30/13.39/13.06	26.67/16.11/15.59	**28.08**/17.03/16.44

A　手法の詳細

ここでは本文中で説明を省略した手法の詳細として，最終的なプロンプトの構成方法と生成の終了判定について説明する．言語モデルへ与えるタスク説明は “[タスク] 与えられた状況に対する実況を生成してください。状況は現在起こっているイベントの継続時間と、それを表現する関係トリプルの列によって与えられます。一つの状況でイベントは連続して複数回発生します。” とした．指示チューニングが施されていないモデルを利用する場合は，システム，ユーザー，アシスタントプロンプトそれぞれを 1 行ごとに記述し，その先頭に担当の表示として “[SYSTEM]”，“[USER]”，“[ASSISTANT]” のテキストを付け加える．指示チューニングを施したモデルを利用する場合は，各モデルの対話テンプレートに記載されたプレフィックスに従う．一般的に指示チューニングを施したモデルはターン終了を表す特殊トークンをもってアシスタントターンの終了を表現するが，本研究で使用するプロンプトでは実況テキストは常に 1 行で完結するため，終了トークンのほか改行文字の出現でもアシスタントターン生成の終了とみなす．指示チューニングが施されていないモデルでは改行文字の出現をもって生成の終了とする．

B　詳細な実験結果

表 4 に本研究で実施した全ての設定についての BLEU スコアを示す．各設定について単語レベル BLEU-4，IPA 辞書での単語分割による単語レベル BLEU-4，neologd 辞書での単語分割による単語レベル BLEU-4 をスラッシュで区切って表示している．多くのモデルについて，最高のスコアとなる設定は BLEU スコアの計測方法によらない．