サッカー実況中継を付加的情報の提供という側面から見る

森雄一郎[1] 前川在[1] 小杉哲[1] 船越孝太郎[1]
高村大也[2] 奥村学[1]
[1] 東京工業大学 [2] 産業技術総合研究所
{moriy, maekawa, kosugi, funakoshi, oku}@lr.pi.titech.ac.jp
takamura.hiroya@aist.go.jp

概要

本稿では，スポーツ実況における付加的情報の提供という側面に焦点を当てる．これまでの実況生成に関連する研究は，主要なイベントが起こった際映像上の内容を細かく記述するタスクの解決を志向している．しかし，現実のスポーツ実況者は，主要なイベントを取り上げるだけでなく，映像に関連する付加的な情報を提供し，観客の知的好奇心を満たす．本研究では，大規模言語モデルを用いて実際のサッカーの試合中継における実況コメントを分析し，付加的情報の提供を担うコメントの割合を調査する．そして，分析により，付加的情報が現実の実況に多く含まれていることを示すとともに，今後の実況生成の研究において付加的情報を考慮する必要があることを主張する．

1 はじめに

スポーツにおける実況者の役割は，観客に対してスポーツの魅力を多角的に伝えることにある．特にサッカーのような動的なスポーツでは，実況者は試合の重要なアクションを情熱的に説明し，試合中の出来事を様々な角度から捉えることで観客の興奮を高め続ける．このような実況の重要性にもかかわらず，プロの実況者が不足していることから，見逃し配信やアマチュアスポーツの録画など，広範な領域においてその魅力を十分に伝えきれていない実情がある．そこで，自然言語生成技術を用いた実況生成の自動化が有望な解決策として提案されている [1, 2]．

従来の研究は，主に試合映像から発話タイミングを推定し，対応する実況文の生成を行う Dense Video Captioning (DVC) [3] タスクの一つとして実況生成に取り組んできた [4, 5]．これらは主に映像上の主要なイベントの記述に焦点を当てているのに対して，

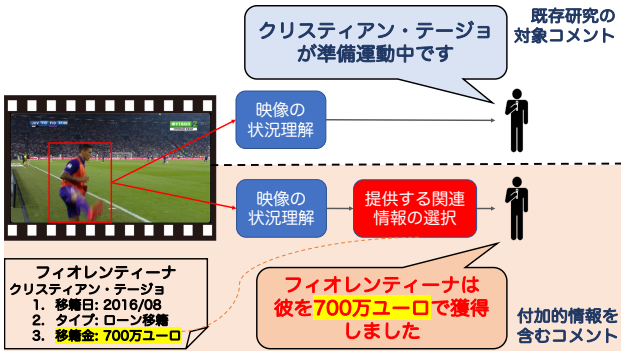

図 1 実況の例．上のコメントは映像の説明を行っており，既存研究で扱われている．これに対して，下のコメントが本研究で扱うコメントであり，資料を参照しつつ映像に関連する付加的な情報を提供している．

本研究では「付加的な情報の提供」という側面に焦点を当てる．現実の実況者は，主要なイベントを取り上げるだけでなく，選手のプロフィールや試合のボール保持率といった，映像に関連する付加的な情報を提供し，観客の知的好奇心を満たすという重要な役割も果たす．例えば，図 1 下の「フィオレンティーナは彼を 700 万ユーロで獲得しました」という実況者のコメントは，フィオレンティーナが劣勢の中，クリスティアン・テージョが交代指示を待つ状況で，実況者が発話したものである．そして，これは「700 万ユーロの価値のあるクリスティアン・テージョの活躍と，それによるフィオレンティーナの逆転への期待感を高める」という重要な役割を果たす．このように，付加的情報の提供という側面は実況においてなくてはならないものである．

本研究では，付加的情報の提供が実況の重要な側面であることを示すために，サッカー試合中継の実況コメント中に，どのくらい付加的情報が含まれるかを調査する．そのため，我々は大規模言語モデルを用いた few-shot learning を活用し，付加的情報を含むか否かを各コメントにラベル付けした実況コメントデータセットを構築する．具体的には，我々は

SoccerNet-v2 [6] に含まれるサッカー実況中継の音声を書き起こし，実況コメントを得る．得られた実況コメントは非常に数が多いため，人手でラベル付けするのは，時間的・金銭的コストが大きいという問題がある．そこで，大規模言語モデルを使用し，低コストでラベル付けを実行する．このラベル付きデータセットは，付加的情報の提供を含む実況生成システム構築に利用できる．最後に，我々は大規模言語モデルにより付与されたラベルを集計し，その割合を算出することで，付加的情報が現実のサッカー実況中継に多く含まれていることを明らかにする．さらに，付加的情報を含むコメントの時間帯別・イベント別の使用率を調査し，構築したラベル付きデータセットの特徴を捉える．調査により，使用率が時間帯や関連するイベントの種類で異なることが分かった．

2 関連研究

2.1 実況生成

スポーツ実況生成の研究は，サッカー [5, 1, 2, 7]，野球 [8]，ゲーム映像 [4] などの領域を対象に行われている．実況生成の研究は主に，テキストやデータを用いる手法と映像を用いる手法，または両方を使う手法がある．Kubo ら [1] の SNS の書き込みを使う手法や Taniguchi ら [2] のデータを使う手法は，実況者が観戦している映像を用いずに実況文を生成している．最近は，急速に発展している Vision and Language の技術を用いて，映像を入力とする実況文生成手法も提案されている．例えば，Ishigaki ら [4] は，レーシングゲームの映像と追跡データを入力として，(1) 発話タイミングの推定と (2) 実況文の生成を行うタスクを提案している．Mkhallati ら [5] は，SoccerNet-v2 [6] に収録されているサッカー実況中継の試合映像のみを入力とする実況生成タスクを提案している．さらに，Mkhallati らはこのタスク用に SoccerNet-Caption という新しいデータセットを構築した．このデータセットには，サッカーの試合映像に基づいたキャプションが含まれている．映像を入力として用いる実況生成の研究 [4, 5, 8] は，DVC と呼ばれるタスクから着想を得ている．DVC とは，映像中の重要なイベントを検出し，各イベントの説明文を生成するタスクである．そして，DVC タスクを発展させる形で，Qi ら [9] は選手やチーム，プレーに関する知識を補強するための知識ベースを用いたサッカー実況文生成を提案している．これらの研究は映像上のイベントの説明に焦点を当てているのに対して，本研究では映像だけでは得られない付加的な情報の提供という側面に焦点を当てる．

2.2 実況発話のラベリング

発話ラベリングは，音声や発話テキストに対して，話者の意図や発話内容のラベルを予測するタスクであり，これまで多くの応用研究が行われた．例えば，音声通話の発話意図推定 [10] やメールスレッドを対象とした発話テキストの分類 [11]，会議や日常会話における発話行為のタグ付け [12] がある．最近では，実況に関連する研究も存在し，上田ら [13] はレーシングゲームの実況発話内容ラベルの推定手法を提案した．本研究では，「付加的情報」という新たな視点に基づいて発話ラベリングを行うため，既存の発話ラベルセットを利用せず，独自の基準を設定した（詳細は 3.1 節で述べる）．

発話ラベリングは，分類もしくは系列タグ付けとして定式化されており，近年はラベル付きデータセットを用いてニューラルネットワークを学習する手法が用いられる [13]．本研究では，発話ラベリングを分類問題として捉え，大規模言語モデルの few-shot learning [14] によりラベル付けを行う．

3 データセット構築

本節では，ラベル付き実況コメントデータ構築の手順を述べる（概要を図 2 に示す）．我々は，3.2 節の実況コメント収集と 3.3 節の自動ラベリングを行い，最終的に表 1 に示すような 428,834 件のラベル付きコメントデータを作成した．

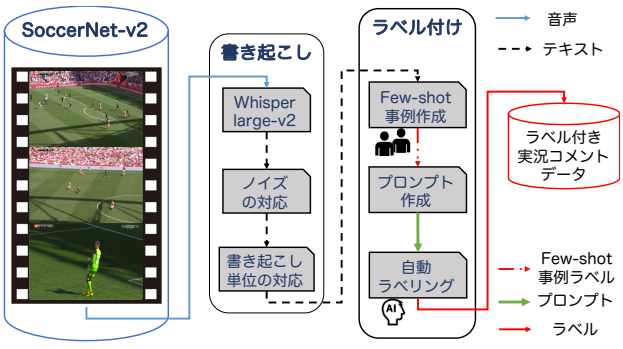

図 2 データセット構築の概要図．

3.1 付加的情報の定義

まず，本研究における付加的情報を定義する．付加的情報とは「実況者が見ている映像からは分から

表 1　構築したラベル付きコメントデータの例．試合情報は試合の日付，開始時間，チーム名，最終スコアを含む．付加的情報かは，実況テキストが付加的情報を含むならば Yes，含まないならば No をとる．

項目	例
試合情報	{date: 2015-08-29, time: 19-30, team_1: Bayern Munich, team_2: Bayer Leverkusen, score: 3 - 0}
発話区間	14:06 – 14:22
実況テキスト	It's a game that we brought you here on BT Sport and it was a stunning performance from Roger Schmid's side to see off the Italians from 1-0 down in the first leg.
付加的情報か	Yes

ない，外部から必要な情報」とする．我々は，全ての実況コメントに対して，付加的情報が含まれるか否かという 2 値分類をおこなった．

3.2　実況コメント収集

本研究では，SoccerNet-v2 [6] をデータソースとして用いる．SoccerNet-v2 には，500 試合の実況音声付きサッカー中継映像があり，そのうち約 80 %が実況者による発話を含む．実況は，英語・スペイン語を中心とする 12 種類の言語により提供される．我々は，SoccerNet-v2 の映像のうち，実況者の発話を含む試合を対象とした．対象とした試合の映像に含まれる実況音声を Whisper large-v2 [15] で書き起こす．Whisper large-v2 を用いることで，多言語の音声の書き起こしと，対応する発話時間の推定を同時に行うことができる．なお，Whisper large-v2 の利用にあたり (1) ノイズ と (2) 書き起こし単位の問題が発生した．付録 A にそれぞれの詳細と対処法を示す．

3.3　大規模言語モデルを用いた自動ラベリング

我々は，3.2 節で収集されたコメントを，大規模言語モデルの few-shot learning で自動ラベリングした．モデルに与えるプロンプト，利用したモデルやパラメータ等の設定を以下で述べる．

プロンプト　我々は，モデルへの指示と人手で作成した few-shot 事例，対象のコメントでプロンプトを構成した．ラベリング例や対象コメントには，参照すべき文脈情報として，(1) 対象の前 2 件のコメント，(2) 試合チームと日付・最終スコアを与えた．ラベリング例には分類した理由も与えており，大規模言語モデルの出力時も同じく理由を述べるように指示した．図 3 にプロンプトの例を示す．なお，few-shot 事例は，人手で分類した 100 件の実況コメントの比率に合わせ，付加的情報を含むコメント 2 件，含まないコメント 5 件で構成した[1)]．

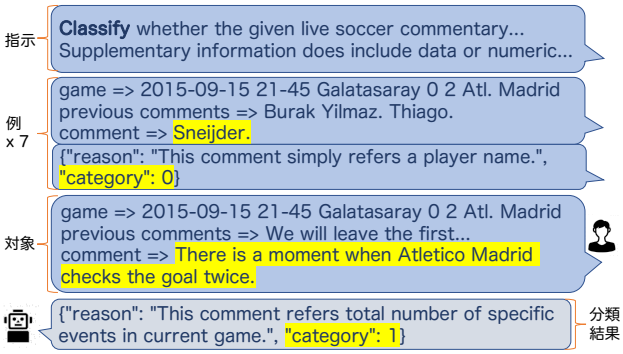

図 3　プロンプトと大規模言語モデルによる応答の例．応答は，ラベリング対象のコメントに対するラベルとその理由で構成される．

大規模言語モデルの設定　我々は，トークンあたりの料金と性能のバランスを考慮し，gpt-3.5-turbo-1106 [16] を用いた．デコードのパラメータは，再現性を考慮して temperature=0 のみ指定し，残りはデフォルト値を用いた[2)]．自動ラベリングに掛かったコストは約 300USD である．

4　分析

本節では，3 節の自動ラベリングの性能を評価し，大規模言語モデルにより得られたラベルの信頼性を確認する．また，データセット構築によって得られたラベルを集計して，付加的情報が含まれるコメントの割合を確認し，現実のサッカー実況には付加的情報の提供という重要な側面もあることを定量的に示す．さらに，付加的情報を含むコメントの時間帯別の使用率を調査し，構築したラベル付きデータセットの特徴を捉える．最後に，擬似的な主要イベントのラベルを用いて，実況者が付加的情報を提供するタイミングについて分析する．

自動ラベリングの評価　大規模言語モデルによる自動ラベリングの信頼性を担保するために，無作為にサンプリングした実況コメント 100 件について，2 人のアノテータにより人手で付与したラベル[3)] と

1)　few-shot 事例は，4 節の評価用サンプルにおける gpt-3.5-turbo の分類性能を最大化するように選択した．本研究における自動ラベリングは，あくまでデータセットの分析と整備のためなので，この事例選択方法は妥当であると考える．

2)　https://platform.openai.com/docs/api-reference/chat/create

3)　人手ラベルの妥当性を確保するため，2 人のアノテータがそれぞれラベリングを行った．2 人が付けたラベルにおける Cohen の κ 係数 [17] は 0.798 であった．ラベルがアノテータ

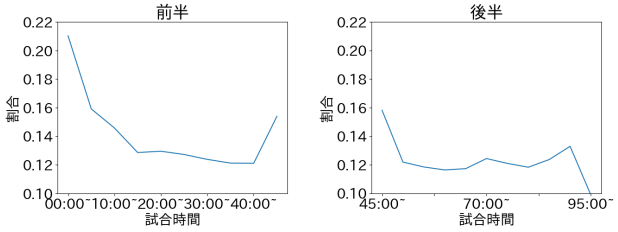

図 4　時間帯別の付加的情報の使用率．ラベルの割合は 5 分間隔で集計したため，グラフは前・後半それぞれ，アディショナルタイムを含めて 10 個の点で構成される．

の Accuracy, Precision, Recall, F_1 を評価した．このサンプル 100 件には，人手で付加的情報を含むと判断されたコメントが 14 件，含まないと判断されたコメントが 86 件含まれている．評価結果は，Accuracy が 0.97，Precision が 1.0，Recall が 0.75，そして F_1 が 0.86 であった[4]．この結果から，大規模言語モデルは高い精度で実況コメントをラベリングできていることがわかる．

ラベルの割合　付与されたラベルごとの実況コメント数を算出すると，「付加的情報を含む」コメントの数は 56,718 で，全体の 13 %であることがわかった．一方で，「付加的情報を含まない」コメントの数は 370,396 で，全体の 86 %である．さらに，「その他」のカテゴリーには 1,718 件のコメントが含まれており，これは全体の 1 %未満である[5]．この 13% という値は，付加的情報を含むコメントが無視できない割合で存在することを示している．つまり，現実の実況において付加的情報の提供という側面も重要であることが分かった．

時間帯別の使用率　本研究では，実況コメントが時系列データであるという特性を考慮し，付加的情報を含むコメントの時間帯別の使用率を分析する．図 4 に示す通り，使用率は時間帯によって大きく異なることが分かる．特に，序盤（試合開始から 5 分まで）では約 21%のコメントが付加的情報を含んでおり，これは中盤や終盤の約 2 倍に相当する．この傾向は，実際のサッカー実況で一般的に見られる特性と一致している．通常，実況者は試合の展開や大会の全体構造に関する説明から始め，試合序盤には視聴者に有益な情報を多く提供する傾向がある．したがって，試合序盤の高い使用率は，このような実況者の行動傾向を反映していると考えられる．

タイミング　SoccerNet-Caption [5] に含まれるキャプションデータには，主要イベントに関するラベル[6]が付与されており，これを元に，実況コメントへのラベル付けを擬似的に実施した．具体的には，主要イベントの時間を示すタイムスタンプの前後 5 秒間に発話開始時間が含まれる実況コメントを，その主要イベントに関連する事例として擬似的に対応付けた．その後，主要イベントに関連する付加的情報を提供するコメントの使用率を算出した結果，主要イベント周辺では，全体における使用率 13%に比べて高いことが判明した（詳細は付録 C.3 に示す）．特に，イエローカードやレッドカードの場面では 23%であり，この傾向が顕著であった．

5　議論・今後の展望

本稿では，大規模言語モデルによる few-shot learning を活用した 428,834 件のコメントのラベリングにより，現実の実況において付加的情報を含むコメントが 13%存在することがわかった．つまり，我々は，現実のサッカー実況には映像の説明としての側面だけでなく，付加的情報を提供するという重要な側面もあると定量的に示した．また，分析により，付加的情報の使用率や提供タイミングに関する，構築したラベル付きデータセットの特徴を捉えた．なお，このデータセットは，付加的情報の提供を含む実況生成システム構築に利用できる．

本研究の展望としては，付加的情報に関する詳細な分析と，この情報を活用した実況生成システムの構築が挙げられる．今後，我々は付加的情報を含むコメントを細かく分類[7]することで，付加的情報の内容を詳しく分析する．つぎに，実況コメントから付加的情報を抽出し，各試合ごとの実況準備資料（付加的情報の知識ベース）を構築する．そして，この準備資料を活用しつつ付加的情報を含む実況文を適切なタイミングで生成する，新たな実況生成タスクに取り組む．これは，映像の理解と効果的な情報提供を必要とする挑戦的な Vision and Language タスクであり，より現実に近いサッカー実況生成への一歩であると考えられる．

間で異なる事例は，2 人で議論しラベルを決定した．

4）　Precision, Recall, F_1 は，Yes ラベルに対して算出した．

5）　「その他」は，入力可能なプロンプトの上限超過や非英語のコメントの影響など，様々な理由でラベル付けができなかったコメントを指す（詳細は B に示す）．

6）　コーナーキック，イエローカードの提示，ゴール，選手交代などを指す．

7）　付加的情報の分類のラベルセットは，(1) スタッツか経歴やエピソードかといった言及内容，(2) 選手やチームなどの言及対象，(3) 過去か現在の試合かという時間軸を基準に作成する．

参考文献

[1] Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Generating live sports updates from twitter by finding good reporters. In **2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)**, Vol. 1, pp. 527–534. IEEE, 2013.

[2] Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. Generating live soccer-match commentary from play data. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 33, pp. 7096–7103, 2019.

[3] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In **Proceedings of the IEEE international conference on computer vision**, pp. 706–715, 2017.

[4] Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. Generating racing game commentary from vision, language, and structured data. In Anya Belz, Angela Fan, Ehud Reiter, and Yaji Sripada, editors, **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 103–113, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics.

[5] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 5073–5084, 2023.

[6] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops**, pp. 4508–4519, June 2021.

[7] Masashi Oshika, Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. Transformer-based live update generation for soccer matches from microblog posts. **arXiv preprint arXiv:2310.16368**, 2023.

[8] Byeong Jo Kim and Yong Suk Choi. Automatic baseball commentary generation using deep learning. In **Proceedings of the 35th Annual ACM Symposium on Applied Computing**, pp. 1056–1065, 2020.

[9] Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, and Jie Tang. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In **Proceedings of the 32nd ACM International Conference on Information and Knowledge Management**, CIKM '23, p. 5391–5395, New York, NY, USA, 2023. Association for Computing Machinery.

[10] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. **Computational linguistics**, Vol. 26, No. 3, pp. 339–373, 2000.

[11] Tatsuro Oya and Giuseppe Carenini. Extractive summarization and dialogue act modeling on email threads: An integrated probabilistic approach. In Kallirroi Georgila, Matthew Stone, Helen Hastie, and Ani Nenkova, editors, **Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)**, pp. 133–140, Philadelphia, PA, U.S.A., June 2014. Association for Computational Linguistics.

[12] Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. In Trevor Cohn, Yulan He, and Yang Liu, editors, **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 2636–2648, Online, November 2020. Association for Computational Linguistics.

[13] 上田佳祐, 石垣達也, 小林一郎, 宮尾祐介, 高村大也ほか. 実況における発話ラベル予測. 研究報告音声言語情報処理 (SLP), Vol. 2021, No. 1, pp. 1–6, 2021.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In **International Conference on Machine Learning**, pp. 28492–28518. PMLR, 2023.

[16] OpenAI. Introducing chatgpt. https://openai.com/blog/new-models-and-developer-products-announced-at-devday, 2023.

[17] Jacob Cohen. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, Vol. 20, No. 1, pp. 37–46, 1960.

[18] Steven Bird, Edward Loper, and Ewan Klein. **Natural Language Processing with Python**. O'Reilly Media Inc., 2009.

[19] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. **Computational linguistics**, Vol. 32, No. 4, pp. 485–525, 2006.

A　データ構築の詳細

本節では，データ構築時に Whisper large-v2 を使用したときに発生した問題と対応方法について述べる．

ノイズ　実況音声は，観客の歓声などのノイズが含まれるが，Whisper large-v2 の出力はおおむね正確である．しかし，実況者が沈黙しているときに，実況者の最後の発話テキストが繰り返し出力されることがある．この繰り返しによるノイズは，現実の実況を収集する上で不要であり，さらにはコメントのラベルの集計結果に影響を与えうるため，分析において問題となる．我々は，繰り返しの初めのテキストのみを残し，それ以降の書き起こしは削除した．

書き起こし単位　Whisper large-v2 による音声の書き起こしの際，テキストが途中で断片化（セグメント化）されることがある．これは，一貫性のないラベリング単位を作ることになり，加えて文章として成立していないものを含んでしまうため，大規模言語モデルラベリング性能に悪影響を与える可能性がある．我々はこの問題を避けるため，nltk [18] の Punkt Tokenizer [19] を使って Whisper large-v2 の書き起こしを適切な文単位に分割する[8)]．

B　制限

本研究にはデータ構築における課題が残されている．それは，自動ラベリングの不具合と信頼性である．

まず，自動ラベリングの不具合であるが，これにはいくつかの型がある．例えば，大規模言語モデルがロシア語などの実況コメントの文脈情報を汲み取れず，-1 や 2 のような指示外のラベルをつける場合があった．また，Punkt Tokenizer の後処理にも関わらず，実況コメントが 10,000 以上の単語で構成される事例があり，その影響でプロンプトが OpenAI のトークン制限を超過する場合があった．

つぎに，自動ラベリングの信頼性である．4 節で自動ラベルと人手ラベルを比較した際，偽陽性はなかったが，偽陰性は存在した．偽陰性が発生した原因としては，(1) データ構築処理に起因するものと (2) 大規模言語モデルに起因するものがある．(1) の例として取り上げるのは，「In 2010.」である．これは，Punkt Tokenizer による処理後にもかかわらず，「In 2010.」が 1 文として認識されてしまい，意味のある情報だと判断されなかったことが原因として考えられる．(2) の例として取り上げるのは，「He's been taken under the wing by Victor Fernandez…（省略）」である．本来ならば，この発話は映像外の知識を必要とする．しかし，大規模言語モデルはうまくこの文脈を捉えられず，付加的情報を含まないと判断してしまった．

8)　分割後の文にも時間情報を付与するため，元の断片的なテキストに付属する時間情報を用いた．その結果，実況コメントの発話区間どうしに重なりが発生した．しかし，この設定は DVC と同様であるため，特に問題はないと考える．

表 2　データセットの統計値．

実況コメント数	428,832
1 試合あたりの平均実況コメント数	932.24
1 発話あたりの平均文字数	10.17
付加的情報の 1 発話あたりの平均文字数	21.44
1 発話あたりの平均発話時間	8.50 秒
付加的情報の 1 発話あたりの平均発話時間	13.74 秒

表 3　サッカーの各主要イベントに関連するコメントの付加的情報の使用率．

イベント	使用率 (%)
コーナーキック	11.74
ゴール	16.41
怪我	14.09
選手交代	16.98
ペナルティ	**19.13**
ペナルティ失敗	13.13
イエロー/レッドカード	**23.29**

C　分析の詳細

C.1　自動ラベリングの詳細

本節では，Few-shot 事例に含むラベルの割合を変化させたときの自動ラベリングの性能の変化について調査する．ラベルの割合を同数（付加的情報を含むもの 3 件，含まないもの 3 件）にした場合は，Accuracy が 0.74，Precision が 0.27，Recall が 0.67，F_1 が 0.38 となった．これは，自動ラベリングの際に用いた割合 (付加的情報を含む 2 件と，含まない 5 件) の時と比較して，全ての評価指標で下がっている．

C.2　データセットの統計

構築したデータセットの統計を表 2 に示す．表から，付加的情報を含む実況コメントは，文字数・発話時間ともに，平均より長いことが分かる．

C.3　タイミング分析の詳細

各主要イベントにおける付加的情報の使用率を表 3 に示す．