修士論文

マルチモーダル情報を利用したストーリーテリングの質の分析

足利 優多

主指導教員 岡田 将吾

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和６年３月

Abstract

Stories in human culture have established a long tradition. People tell stories for a wide variety of reasons, from simply entertaining, to transmitting knowledge across generations, to maintaining culture, to warning others of danger. This study focuses on storytelling, a communication method of telling stories. Storytelling is an interactive art that stimulates the listener's imagination and reveals the elements and images of a story through words and actions. If stories trace facts, stories are creative, deeply expressive of a person's inner life and important for understanding the human condition. The uses of storytelling are varied. For example, in product sales, telling the buyer the right story at the right time can convey that the product's value meets their needs, or interpreting the buyer's stated story can provide clues to finding a solution to a problem. According to the definition of storytelling, storytelling can also interact and actively create a reality in the mind, such as a vivid story, sensory images or events, based on the listener's own past experiences to the listener. It is then stated that a more complete story is unique and gives belief and understanding in the mind of the listener, making them a co-creator of the story. In other words, storytellers should work on improving their storytelling performance, as good or poor storytelling skills affect the images created for listeners and their understanding.

However, even if storytelling is done to someone, there needs to be someone to evaluate it, and it is not known whether the evaluation is valid. Furthermore, even after appropriate assessment, performance will not improve if people do not know how to improve their storytelling skills. Therefore, in this study, we present a framework in which the target speaker can be evaluated by a machine learning model using a storytelling dataset. By obtaining an appropriate and numerical evaluation of one's own storytelling skills, one can objectively know one's own skills. In addition, conventional research has used many modalities for estimating storytelling skills, such as prosody, gestures, and listeners' affusions, as well as the content of the speaker's utterances. It is not easy to know one's own skills, as a lot of equipment and manpower are indispensable to collect such data. Therefore, in this study, only data that is relatively easy to collect, such as the content of the speaker's speech, i.e. text data, was required, thus reducing the cost of data collection. However, some storytelling used in real life is based on personal experience. When storytelling such content, which varies greatly from person to person, it is difficult to evaluate all speakers appropriately and equally. For this reason, the content of what the speakers say is a description of a specific video, to reduce the variation between people.

It then helps speakers to improve their performance after receiving an appropriate evaluation by revealing not only the evaluation, but also the reasons that led

to the evaluation. Specifically, using text features based on speech data from the speaker and image features based on image data from a specific video, machine learning was used to learn each of the seven storytelling skill items, and the results obtained were used to elucidate what good storytelling is.

In this study, data from storytelling to illustrate a particular video were used to experiment and discuss the estimation of speakers' storytelling skills and their storytelling prowess. There were seven storytelling skill assessment items, each scored on an eight-point scale, with 1 being the lowest rating, and the score for each item was set as the objective variable in the skill estimation. The features used as explanatory variables were created by combining three types of features: linguistic features taken from before one of the final layers of BERT, linguistic features from the output vector of CLIP's text encoder and image features from the output vector of CLIP's image encoder. A Gaussian kernel SVR was used for the model and the coefficient of determination R^2 was obtained by five-part cross-validation. Experimental results showed that BERT-only features had the highest accuracy in many evaluation items, but the highest accuracy was obtained when CLIP linguistic features were used in addition to BERT for items such as 'the scene was well described in words and the content of the story (information) was accurately conveyed'. In CLIP, even the text encoder is influenced by the image during the learning phase. From this, it was considered that the ability to visualise the scene when storytelling was evaluated. In the evaluation item "confidently explained", the CLIP image features were more accurate than the BERT-only features. The fact that CLIP's image features were effective in the item "confidence", which is apparently unrelated to images, was considered to be due to the fact that when speakers speak confidently, they have a clear image of the image in their mind.

The model used in this study is SVR, and time series are not considered when creating features. By using a model that can handle time series such as LSTM as a machine learning model, when speech is input in a time series, the scene estimated images are also arranged in a time series, and by comparing the speech and scene estimated images, it is possible to quantify how coherent the speech is and improve accuracy. In the future, in order to clarify storytelling skills, it will be necessary to create another set of features using CLIP, looking at the correlation with the evaluation items as well as the correspondence of the time-series data.

目 次

図 目 次

表 目 次

第1章　はじめに

人の文化におけるストーリーは，長い伝統を確立してきた．人々がストーリーを語る理由は多種多様で，単純に楽しませるためのものから，世代を超えて知識を伝達するため，文化を維持するため，他人に危険を警告するためのものがある [1]．本研究では，ストーリーを話すコミュニケーション手法であるストーリーテリングに着目した．ストーリーテリングとは，聞き手の想像力を刺激し，言葉と行動によってストーリーの要素やイメージを明らかにするインタラクティブな芸術である [2]．物語が事実をなぞらえるものとすれば，ストーリーは創造的で，人の内面を深く表し，人の状態を知るために重要なものである [3]．ストーリーテリングの活用は様々である．例えば，製品の営業において，購入者に適切なタイミングで適切なストーリーを話すことで，製品の価値が相手のニーズを満たすことを伝えたり，購入者の述べたストーリーを解釈することで問題の解決策を得る手がかりとなる [4]．また，[2] のストーリーテリングの定義によると，ストーリーテリングは相互に作用し，聞き手に対して聞き手自身の過去の経験に基づいて，頭の中に鮮明なストーリーや感覚的なイメージ，出来事などの現実を積極的に創造することができる．そして，より完成されたストーリーは，ユニークで聞き手の心の中に信念と理解を与え，物語の共同創造者になると述べられている．つまり，ストーリーテリングスキルの良し悪しは聞き手に作られるイメージや，理解に影響を与えるため，語り手はストーリーテリングパフォーマンスの向上に取り組むべきである．しかし，誰かに対してストーリーテリングを行ったとしてもそれを評価する者が必要であり，またその評価が妥当であるかもわからない．さらに，適切な評価を受けた後も，ストーリーテリングスキルの改善方法が分からなければパフォーマンスの向上には繋がらない．そこで，本研究では，ストーリーテリングのデータセットを用いて，機械学習モデルにより対象の話者の評価を行うことができるフレームワークを提示する．自身のストーリーテリングのスキルを適切な評価で，かつ数値的に得ることで，客観的に自身のスキルを知ることができる．また，従来の研究では，ストーリーテリングスキルの推定に話者の発話内容だけでなく，韻律やジェスチャ，聞き手の相槌などの多くのモダリティが用いられてきた．そのデータをとるために多くの機材や人手が不可欠であり，自身のスキルを知ることは容易ではない．そのため，本研究においては話者が話す発話内容，つまりテキストデータといった比較的集めやすいデータのみを必要とし，データ収集コストを下げた．ただし，現実的に使用されるストーリーテリングには自己の経験を元とするものもある．そのような，個人差の大きい内容をストーリーテリ

ング行う場合，全ての話者を適切に平等に評価することは難しい．このため，話者が話す内容は特定の動画の説明とし，人によるバラつきを抑える．そして，評価だけでなく，どのような理由でその評価に至ったのかを解き明かすことで，話者が適切な評価を受けた後にパフォーマンスを改善する手助けとなる．具体的には，話者から得た発話データをもとにしたテキスト特徴量と，特定のビデオの画像データをもとにした画像特徴量を使い，7つのストーリーテリングのスキル項目ごとに機械学習を用いて学習し，得た結果を元にストーリーテリングの上手さとは何であるのか解明する．

本論文では，2章でストーリーテリングスキル推定と特徴量抽出に用いたCLIPの関連研究を共有し，本研究の立ち位置を示す．3章では推定する7つのストーリーテリングスキルやデータの種類等の実験で用いるデータセットを説明する．4章でCLIPや日本語CLIPモデルの仕組みを解説し，5章で研究手法とデータセットを用いて研究する際の前処理を述べる．6章では，実験結果を示し，考察を述べる．7章で本研究をまとめるとともに，今後の展望を述べる．

本研究の貢献は以下の通りである．

1. 限られたデータを用いたストーリーテリングのスキル推定
   主に言語からのスキル推定を行うため，話者の発話内容の記録さえ録ることができれば，推定することができ，実験環境にとらわれることなく実用的である．

2. CLIP特徴量によるストーリーテリングスキルの推定精度向上
   CLIPのテキストエンコーダや画像エンコーダを用いた特徴量抽出とBERT特徴量を組み合わせることで，ストーリーテリングスキルの一部の項目においてBERT特徴量のみの場合より精度が向上した．CLIPモデルがストーリーテリングスキル推定に役立てることができると示した．

3. ストーリーテリングの上手さについての見解
   CLIPという画像と関連するモデルを用いた特徴量と推定により得た結果を元に，特定のスキルが話者が持つイメージと関連していることを示す．

第2章　関連研究

2.1　ストーリーテリングパフォーマンスに関する研究

グループ会話におけるストーリーテリングパフォーマンスのモデル作成を行った岡田ら [5] は言語的特徴と非言語的特徴といったマルチモーダルな特徴量を，話し手の対話や韻律，ハンドジェスチャやヘッドジェスチャなどから抽出した．また，聞き手の相槌時の話者の発話といった共起特徴も使われた．話し手のパフォーマンス予測の計算モデルを開発し，実験した結果，ストーリーテリングパフォーマンスの総合成績の最高精度が $R^2 = 0.299$ に達し，過去の文献よりも高精度な予測であると示された．富沢 [6] は本研究と同一のデータセットを用いて，説明や強調に付随するジェスチャ機能の推定を行った．使われた特徴量は，Kinect の姿勢データ，Openface により話者の発話ビデオからとりだされた表情特徴量，OpenSmile による音声特徴量などである．実験では SVM とニューラルネットワークが用いられ，ジェスチャ機能の推定には，姿勢データだけでなく，表情特徴量と音声特徴量を使うことで認識の精度が向上することが示唆された．

2.2　CLIP を用いた研究

CLIP は大量のデータで大規模にトレーニングされたモデルであり，幅広いアプリケーションの基盤となっている [7]．画像とテキストにより学習された CLIP は自然言語処理 [8] や画像処理 [9]，ロボティクス [10, 11, 12] など多様な研究において，目的の達成を補助する 1 つの部品として使用されたり，新たな構造の CLIP モデルが開発されている．Alex ら [13] は，画像生成タスクにおいて，意図された画像が生成されるために，入力されたテキストに合うように調整する方法として CLIP を用いた．具体的には，画像の説明文と CLIP の画像エンコーダの特徴量の内積が大きくなるように学習が行われた．Jabbar ら [8] は，ビジュアルアートワークと詩のマッチングのために，CLIP のテキスト埋め込み表現を駆使することで，詩で使われる比喩などの象徴的な表現を上手く汲み取り，より細かな詩の特徴を適切にモデル化した．Igaue ら [14] は,「不気味の谷効果」と呼ばれる人間に似たロボットや CG のアバターが人間的な振る舞いを行うことで逆に不気味な印象を与える効果について調べた．CLIP を人間の感情モデルとして使用することで，人間の顔に対する矛盾的視覚の特徴が否定的な言語表現と関連していることを示した．つ

まり，機械学習モデルにより，人間の観察者の感情と視覚的な手がかりとの関連性を提示した．

2.3 本研究の立ち位置

ストーリーテリングを用いた既存研究 [5] では，予測精度は高いが，元動画の画像との関連性を考慮した画像特徴量は使用されていない．また，ストーリーテリングを行う際の状況が本研究と異なり，1 対 1 の対話ではなく，2 人の話者による説明である．本研究と同様のデータセットを用いている先行研究 [6] では，焦点をジェスチャの機能に絞っており，用いる特徴量はジェスチャと関連のある表情特徴量等を使用している．本研究においては言語データを主に用いることで，他のモダリティを収集する必要がなく，フレームワークを使用することが容易である．CLIP を用いた研究では，CLIP のテキストか画像どちらかの埋め込み表現を獲得し，利用するものが多い．「不気味の谷効果」についての研究 [14] では，CLIP を用いた結果から人間の感情との関わりについても言及されており，本研究におけるストーリーテリングの上手さを解き明かすことと比較的近い立場であるが，用いる手法は CLIP を用いること以外全く異なるものである．

第3章　データセット

本章では，本研究で使用したデータセットについて説明する．データの概要を述べた後，データを収集した際の環境や詳細な条件を示す．収集したデータに対して2人のアノテーターにより評価が行われ，どれだけ信頼できる評価であるかをクロンバックのα係数により提示する．

3.1　データの概要

本データセットには，音声データ，テキストデータとビデオデータが含まれている．テキストデータは，音声データをもとに書き起こしを行ったものであり，ビデオデータは実験の様子を撮影したものと，実験に用いたアニメーションの2種類ある．

3.2　実験環境

1回の実験は2人の参加者によって行われる．1人は話者として参加し，視聴した動画の内容をもう1人の聞き手に対して身振り手振りを用いながら向かい合って説明する．このとき，聞き手は動画を視聴しておらず，内容についての疑問をいつでも話者に質問し聞くことができる．話者は，内容を忘れることのないよう動画を視聴しながらメモを取ることが許可されており，話す内容を失念した場合には，そのメモの内容を確認することができる．ただし，ストーリーテリングによるジェスチャや自然な会話の妨げとなることを防ぐため,可能な限りメモは見ないように指示をされる．また，使用された視聴動画は2種類あるため，1人あたり2回のストーリーテリングが行われた．

3.2.1　視聴動画

実験の話者が視聴する動画は，ストーリーテリングのジェスチャ研究でよく用いられる [15, 16]「Canary Row」と「Tweety's S.O.S」いうワーナー・ブラザースによる短編アニメーションを用いた．子供でも理解できるような分かりやすい内容で，黄色い小鳥の”Tweety Bird”と白と黒の模様の体をした猫の”Sylvester Cat”が主に登場し，猫の Sylvester が小鳥の Tweety を食べようと試行錯誤するストー

リーである．例えば,「Canary Row」では，猫禁止のマンションで飼われているTweetyを目指して，雑技団の猿のまねをして中に侵入するなどの行動を起こし,「Tweety's S.O.S」では，船で旅行に連れられているTweetyを飼い主の隙を見て捕まえに行くという大筋になっている．1つの短編アニメーションの中でもさらにいくつかのシーンに分けられており，1つのシーンで，SylvesterがTweetyを捕らえるアイデアを考え，行動し，失敗するという流れで構成されている．

3.2.2 実験参加者

参加者は合計37人（男性6人，女性31人）で，計72セッション取得した．全員が日本語話者であり，ストーリーテリングおよびストーリーに関する質問は全て日本語で行われた．実験参加者の年齢に特に制限はないが，20代〜40代が多く見受けられる．

3.3 アノテーション

話者のストーリーテリングは，2人のアノテーターによって，ストーリーテリングのスキルに関する7つの項目で評価された．ここで評価された点数から目的変数を定義し，回帰問題として機械学習を行った．

3.3.1 動画のシーン数

動画のシーンは，実際に動画を視聴して，話の区切りを定義し，それを1シーンとしている．多くのシーンは動画が暗転したところで分けられるが，中には話が切り替わっているにも関わらず暗転が起こらない場合もあるため，目視で，1つの動画をよりわかりやすい区切りでシーンを定義した.「Canary Row」は初めの状況説明をシーン0として，合計10シーンに,「Tweety's S.O.S」は合計7シーンに分けられる（表3.1）.

3.3.2 評価内容

評価項目は表3.2に示す通り7つの項目がある．各項目に対して，全くその通りであるときの点数を8点，全くそうでないときの点数を1点として8段階で評価を行った．評価をする際には，話者がストーリーを説明する説明動画とその説明動画に該当する視聴動画をシーン単位で見比べながらスコアを付与した．説明者間・シーン間で出来る限りスコアの順序関係は保持するようにし，各項目は互いに影響をさせることなく，それぞれ独立して点数をつけることとした．

表 3.1: 各シーン内容

題 / シーン	Canary Row	Tweety's S.O.S
0	状況説明	状況説明
1	マンションに入り追い出される	傘で撃退、眼鏡を落とす
2	排水管の外から家に入る	出航
3	排水管の中に入るが，ボウリングで撃退	眼鏡に落書き、綱渡り
4	猿真似して侵入	薬入れ替え，爆発
5	ドアボーイの真似	花火
6	シーソーを使う	エンディング
7	ターザンのように	-
8	電線伝うが，電気を浴びる	-
9	エンディング	-

表 3.2: 評価項目

評価番号	内容
Q1	自信をもって説明していた（自信）
Q2	言いよどみなく円滑に説明していた（流暢さ）
Q3	説明に用いる語彙が多様かつ的確であり，説明内容は機知に富んでいた（語彙）
Q4	説明の中にユーモアがあった（ユーモア）
Q5	情景を上手く言葉で言い表せており，正確に物語の内容（情報）を伝えていた（的確さ）
Q6	説明中のジェスチャ・アイコンタクトは適切だった（ジェスチャ）
Q7	説明には冗長な部分がなく，簡潔に要点を説明した（簡潔さ）

3.3.3　クロンバックのα係数

クロンバックのα係数とは，1951年にリー・クロンバックによってつくられた評価係数であり，主にアンケートの信頼性の評価に使われることが多い．式3.1で表され，0〜1の値を取る．信頼性の1つの側面として**内的一貫性**という，似た内容の質問を複数回したとしても，同程度の結果が得られるかという指標があり，クロンバックのα係数を利用することでその数値化が行える．得られた数値は，1に近いほど信頼性が高いため，一般的に最低でも0.7程度あると良く，往々にして0.8以上あることが望まれる．

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{i=1}^{m} {\sigma_i}^2}{{\sigma_x}^2}\right) \tag{3.1}$$

m 項目数

σ_i 各項目の分散

σ_x 各項目を合計した得点の分散

2人アノテーターの各評価項目ごとにクロンバックのα係数を適用すると表3.3のようになった．Q7の説明の冗長性に関するアノテーションは0.635と比較的低い値であるが，その他の項目に関しては十分信頼性があるといえる．

表 3.3: 各評価項目に対するα係数

Q1: 自信	Q2：流暢さ	Q3：語彙	Q4：ユーモア	Q5：的確さ	Q6：ジェスチャ	Q7：簡潔さ
0.823	0.861	0.816	0.871	0.796	0.887	0.635

第4章　CLIP

既存研究では，CLIP のエンコーダによる埋め込み表現を用いた研究や，CLIP のモデル構造を修正して解きたいタスクに適応させている．本研究では，CLIP に 2 つあるエンコーダのどちらも使用し，特徴量抽出の 1 手法として CLIP モデルを使用するため，本章にて仕組みと実験で使用したモデルの詳細を説明する．

4.1　CLIP の仕組み

CLIP（Contrastive Language–Image Pre-training）[17] は OpneAI によって 2021 年に公開されたマルチモーダルモデルである．モデルの学習には対照学習という，ある空間において似たデータを近くに，異なるデータを遠くに置くように学習する自己教師あり学習の 1 つである手法が使われた [18]．モデルの構造は図 4.1 のようにテキストエンコーダと画像エンコーダで構成されている．テキストエンコーダは Transformer をベースとし，画像エンコーダは ResNet または ViT（Vision Transformer）をベースとしている．学習データは，Wikipedia のサイトで 100 回以上現れる単語と，その単語をネット上から検索して収集した画像のペアで，4 億組のペアを持つ教師データを用いてテキストがどの画像とペアであるかを予測するタスクを解く．具体的には，特定の画像と全てのテキストの cos 類似度を計算し，正解ペアの cos 類似度が最大になるようにする学習を全ての画像に対して繰り返す．CLIP は Zero-Shot の機能を持つため，事前学習なしに他のタスクに対応することができ，例えば，画像のランキングをするタスク [19] で使われることがある．CLIP の活用において，画像の分類器を作成することが考えられる．特定の画像が入力されたときにテキストのペアを探し，cos 類似度が最大となるテキストが画像を説明しているものである．たとえば，動物の画像を分類するタスクを考える．犬の画像を正しく犬のカテゴリに分けたい場合，あらかじめ用意しておいた動物のテキスト集合から得られた埋め込み表現全てと犬の画像の埋め込み表現との cos 類似度を計算し，最も高い値となったテキストに分類を行う．しかし，そのとき入力されるテキストはプロンプトを工夫することにより精度が向上することがわかっている [17]．今回の例では，”dog”と単語のみを入力するよりも，”a photo of a dog”とすることがプロンプトエンジニアニングといわれる技術である．また，さらにテキストに情報を与えることが精度に良い影響を及ぼす場合がある．例えば，衛星データを使用した場合は”a satellite photo of a label”とする．実際，図 4.2

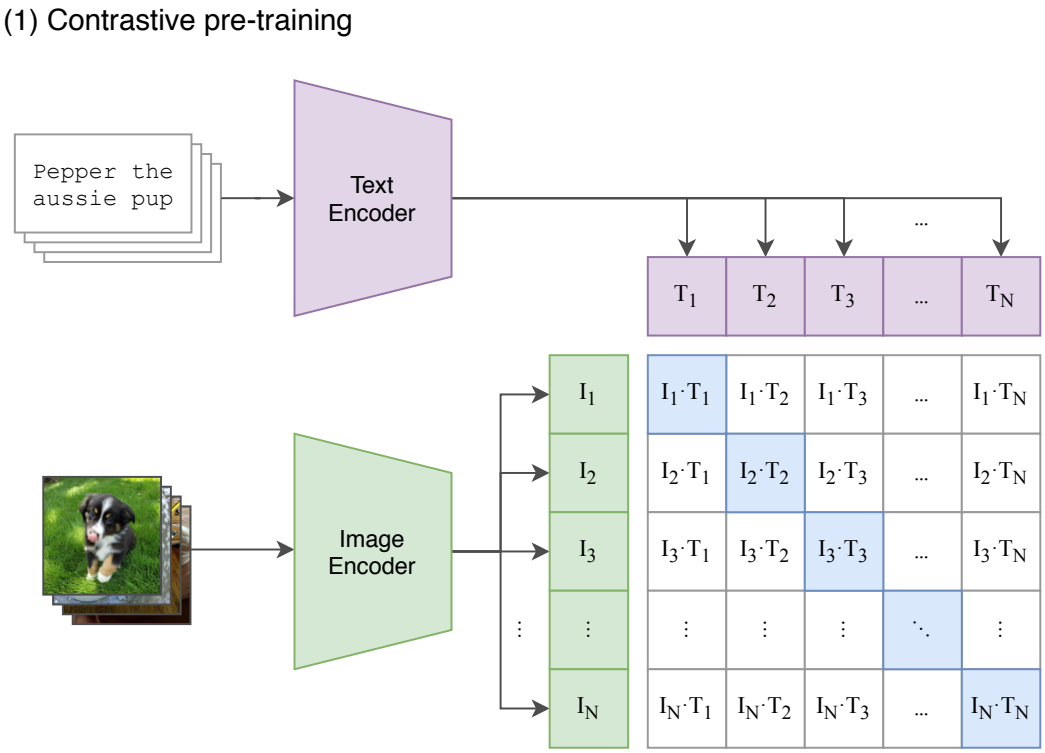

図 4.1: CLIP モデル構造[1]

のようにプロンプトのアンサンブル学習によって 5 ポイントもの改善を示した．

本研究で使用される動画はアニメーションであり，そこから取り出された画像も現実を写した写真ではなく，イラストである．このときにもプロンプトエンジニアニングを行うことによる効果があるのか確かめた．まず，排水管のイラストが描かれた画像を分類することを考える．排水管に加え，アニメーションに出てくる 2 つのアイテム（猿とシーソー）との cos 類似度を測定し，画像が正しく分類されているかを確認する．単語のみを入力した場合の結果が表 4.1 である．排水管（drainpipe）は 0.475 と他のアイテムより高いため正しく振り分けることが成功したといえるが，誤りである猿（monkey）の値は 0.433 で排水管の数値とは僅かしか変わらなかった．そこで，プロンプトエンジニアニングにより，単語の前に”a photo of”をつけた結果が表 4.2 である．排水管（drainpipe）の値は単語のみを入力とした場合の 0.475 よりも大きく上がり，0.648 で他の要素とは 0.3 近くの差をつけた．つまり，プロンプトエンジニアニングはイラストに対しても正しく機能し，その効果により，さらに確信をもって画像を分類することが可能となった．

[1]Alec Radford and Jong Wook Kim and and et al.（2021）. Learning Transferable Visual Models From Natural Language Supervision , p2，図 1 より引用

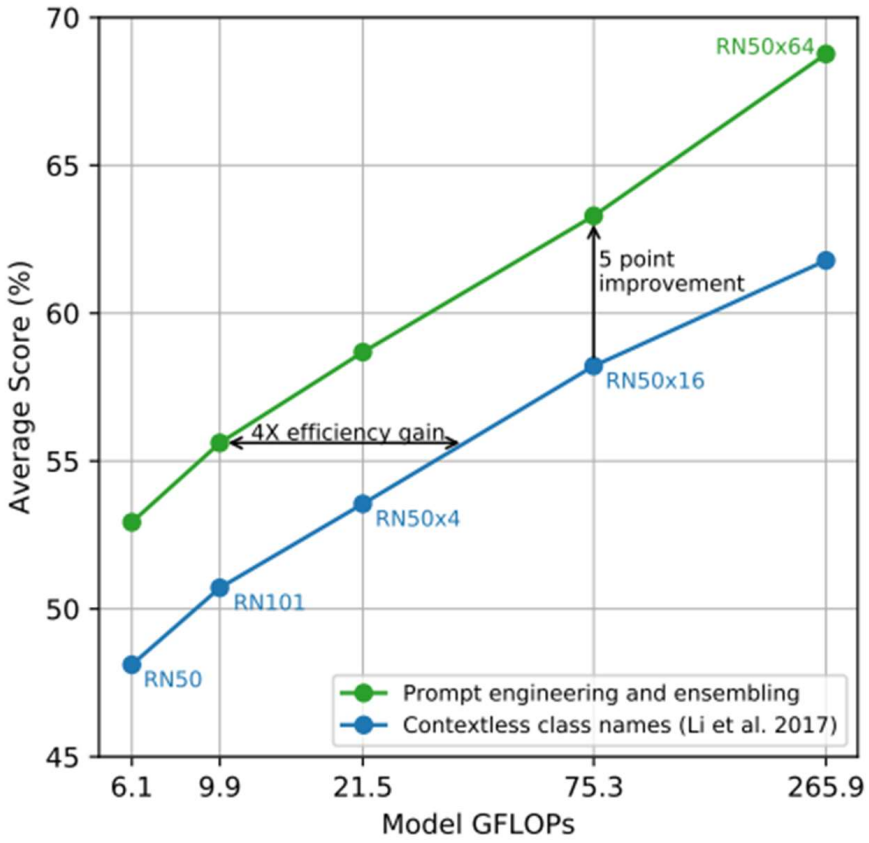

図 4.2: プロンプトエンジニアニングによる効果[2]

表 4.1: プロンプトエンジニアニングのイラストへの効果 1：単語

入力内容	cos 類似度
a drainpipe	0.475
a monkey	0.433
a seesaw	0.092

表 4.2: プロンプトエンジニアニングのイラストへの効果 2：プロンプトエンジニアニング

入力内容	cos 類似度
a photo of a drainpipe	0.648
a photo of a monkey	0.321
a photo of a seesaw	0.030

4.2 CLIP 日本語モデル

上記 CLIP モデルのテキストエンコーダは英語で学習されており，日本語には対応していない．本研究で使用されたデータセットの発話データは全て日本語である

ため，日本語に対応したCLIPモデルを使用する必要がある．そこで，sonoisa(園部 勲)の日本語版CLIPモデル（clip-vit-b-32-japanese-v1）[3]を本研究のCLIPモデルとして利用した．日本語版CLIPモデルは，OpenAIのCLIPモデルの画像エンコーダをそのまま使い，テキストエンコーダを蒸留により作成された．具体的には，英語の画像キャプションデータを日本語に機械翻訳し，英語と日本語の対訳ペアを作る．対訳ペアの英語文全てをOpenAIのCLIPのテキストエンコーダに入力し，512次元のベクトルを複数得て，教師データとする．データセットの日本語文を入力したときに，教師データのベクトルを出力するように平均二乗誤差を使い転移学習を行う．日本語版CLIPモデルにより埋め込まれた日本語の類似度行列とOpenAIのCLIPモデルにより埋め込まれた英語の類似度行列を比較すると，日本語版CLIPはOpenAIのCLIPと比べて，平均で0.042変化し，プラスの方向に最大0.138，マイナスの方向に0.035ほどであり，若干文同士を似ていると判断する傾向が見られるものの，全体として許容できる範囲である[20].

[2]Alec Radford and Jong Wook Kim and et al.（2021）. Learning Transferable Visual Models From Natural Language Supervision , p7，図4より引用

[3]https://huggingface.co/sonoisa/clip-vit-b-32-japanese-v1

第5章　提案手法

本章では，ストーリーテリングの上手さを推定するための機械学習をするにあたり重要となる特徴量について，画像とテキストの関連を示すCLIPを利用した手法を提案する．本手法を用いることで話者のストーリーテリングスキルの推定だけでなく，ストーリーテリングの上手さとはなにかを解き明かすことに繋がると考える．

5.1　シーン推定

CLIPでは，入力された画像とテキストそれぞれで512次元のベクトルを得ることができる．ベクトルは同じ空間上に映されるためcos類似度によりどれだけ画像とテキストが近いかを計算することができる．計算結果を本研究では，類似度として定義する．類似度を役立てることで，話者が話している箇所が動画のどのシーンに該当するかを推定することができる．ここで必要なデータは動画と動画から切り取った画像がどのシーンを示すかのデータ，テキスト化された発話内容である．画像群をCLIPの画像エンコーダへの入力，発話テキストの1文をCLIPのテキストエンコーダへの入力とすることで，最終的に出力として，シーン番号を得る．シーン番号は，入力されたテキストとの類似度が最も高いとされた画像が紐づいているシーン番号であり，本研究では，動画から得た画像全てに順番に番号を振り，一定範囲の番号のかたまりに対して順にシーン番号を振り分ける．

具体的には，図5.1のようなステップを踏む．まず，動画からシーンを十分に説明できるキーフレームを抜き出し，抜き出した画像と該当するシーン番号とを結びつける．次に，1人の話者が話した文章の中でシーンを推定したい箇所を1文取り出す．取り出した文をCLIPでベクトル化したものと動画から抜き出した画像群のCLIPでベクトル化したものとのcos類似度を測定し，類似度が最も高いペアの画像が示すシーンを，取り出した文のシーンとする．このシーン推定の方法では，画像との類似度が低い場合でも，必ずどれか1つの画像が選ばれていまい，動画に関係のない話をしている文を入力したとしても，動画のシーンのいずれかに当てはまってしまう．この課題を解決するため，画像とテキストとの類似度に閾値を設け，閾値を下回った場合は該当シーンなしとする．

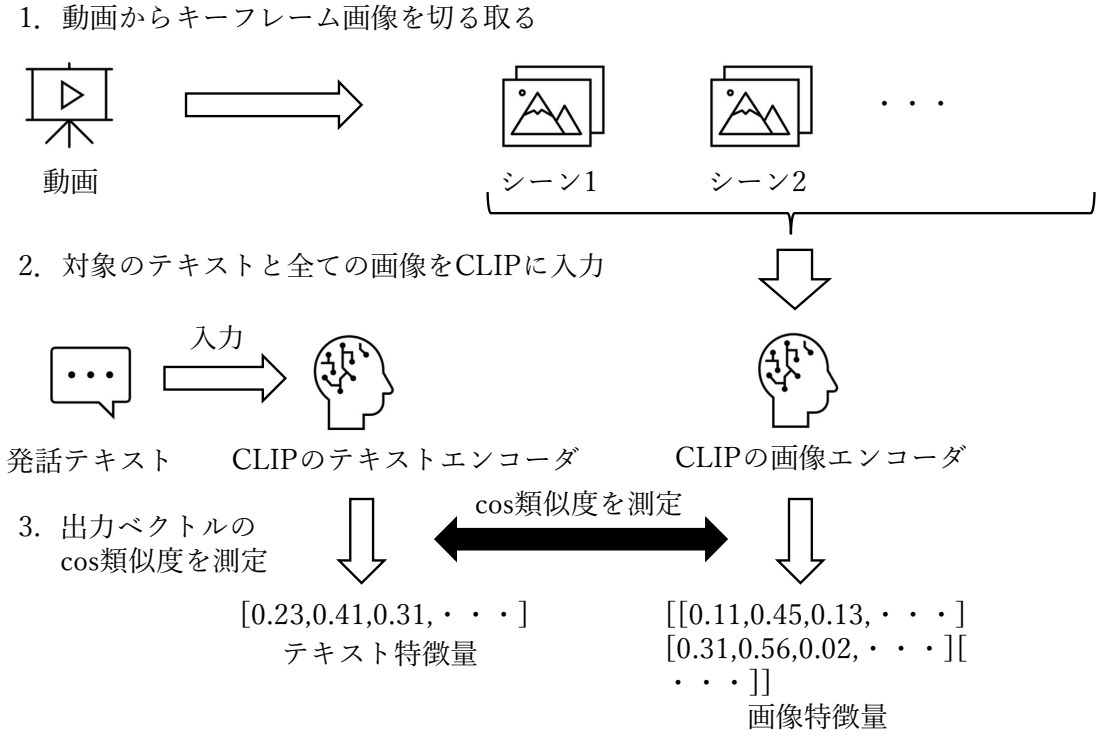

図 5.1: シーン推定の概要

5.2 特徴量抽出

ストーリーテリングスキルの推定に回帰モデルを利用するにあたり，入力となる説明変数に注目する．本研究では，皆が同じ動画を視聴し，ストーリーを説明するため，視聴した動画と関連がある特徴量を入力することが精度の向上に繋がると考えられる．そこで，本手法では，従来から使用される BERT の言語特徴量だけでなく，画像との関連も学習している CLIP を用いて，言語特徴量と画像特徴量を準備する．

5.2.1 言語特徴量

BERT による特徴量抽出

BERT（Bidirectional Encoder Representations from Transformers）[21] とは Google が 2018 年に発表した大規模言語モデルである．Transformer を基礎とし，双方向のタスクを学習に組み込んだことで，単方向のタスクで学習していた従来の言語モデルと比べて強力なモデルとなり，NLP の分野で多くの貢献 [22] をしている．BERT の特徴量は，何段ものエンコーダを通して変わっていくため，最も学習が進んだ最終層の 1 つ手前から 768 次元のベクトルを取得する．

本研究では，日本語の発話に対しての特徴量が必要となるため，東北大学の日本語 BERT モデル[4]により特徴量を取り出した．また，シーン単位でのスキル推定時には，話者やシーンごとに文章数が異なるため，平均と分散を取り結合した 1536 次元の特徴量として取り扱う．

CLIP による言語特徴量抽出

CLIP はテキストのエンコーダと画像のエンコーダを併せ持ち，2 つのエンコーダから出力される 512 次元のベクトルが，画像とペアであるテキスト同士で cos 類似度が大きくなるように学習されている．学習済みのテキストエンコーダに対象のテキストを入力することで 512 次元のベクトルを得ることができる．例によって，話者やシーンごとに文章数が異なるため，出力された 512 次元ベクトルの平均と分散を取り結合した 1024 次元の特徴量として取り扱う．

5.2.2 画像特徴量

CLIP による画像特徴量抽出

5.2.1 の CLIP による言語特徴量抽出の項でも述べた通り，画像の特徴量は画像エンコーダに入力することで 512 次元のベクトルを得ることができる．本研究では，シーン推定により，選ばれた画像のみを画像エンコーダに入力する．話者の発話はシーンごとに分割され入力されるため，文章数が少なく，閾値があることでシーン推定をしても画像が 1 枚も選ばれない可能性がある．そこで，閾値はそのままで．シーン推定時にテキストと画像の cos 類似度の TOP3 までを選ばれた画像とする．結果として，選ばれる枚数はそれぞれの発話ごとに異なるため，画像枚数分得たベクトルの平均と分散を取り結合した 1024 次元のベクトルを特徴量として取り扱う．

5.3 データ前処理

第 3 章に記載したデータセットで本実験を行うための前処理は大きく分けて 2 つある．1 つはテキストに関する前処理である．データセットでは，ストーリーテリング行う動画が保存されており音声を取り出すことができるが，話し手の発話には「あの..」や「ええと」などのフィラーと呼ばれる言葉と言葉の間を繋げる，意味を持たない単語が入ることが頻繁にある．しかし，一部のフィラーは他の役割を持つことがあり，単純に該当の単語を切り取ることはできない．例えば,「あの」という単語はフィラーだけでなく，指示語の役割も持つ．フィラーの除去のため，

[4]https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

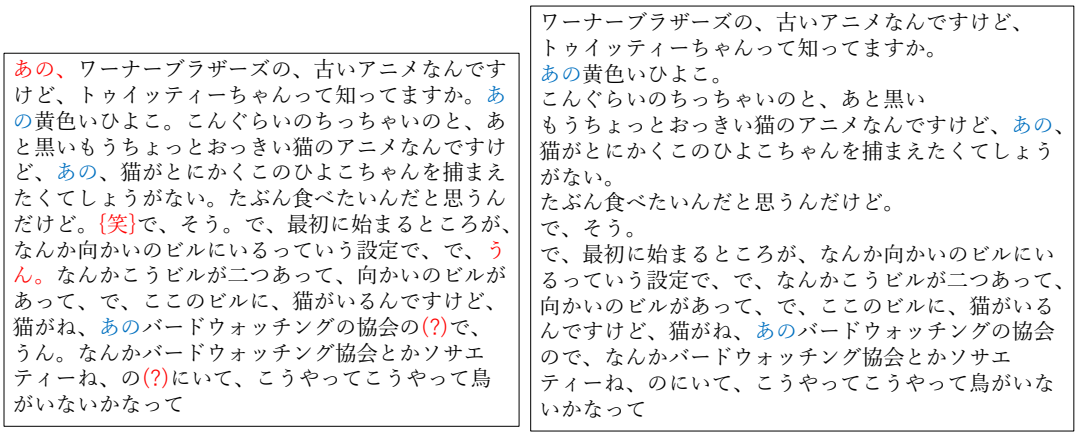

図 5.2: テキスト処理前と後の比較

本研究では GiNZA[5] と呼ばれるオープンソースの日本語 NLP ライブラリを用いた．GiNZA とは，2019 年に公開された日本語の形態素解析や依存構造解析を高速で行うフレームワークで，spaCy[6] という高度な自然言語ライブラリを基盤に置くことで，高精度の解析を行うことができる．GiNZA を使い発話内容を形態素解析することで，単語の役割を判別することが可能となり，フィラーを取り除くことができた．また，アノテーションの際に付けられた {笑} や（?）といった特殊な記号もノイズとなるためテキスト処理を行い，削除した．これらの処理によってフィラーやその他特殊な記号を取り除いたようすが図 5.2 である．赤い文字で表されたフィラーは削除されるが，青い文字で表した指示語は削除されずに残っていることがわかる．その後，文章をアノテーションされたシーンごとに切り分け，シーン単位のテキスト集合とする．実験において使用されるテキストは，全て本手順によりシーンごとに分割されたテキストである．

2 つ目の前処理は，動画からのキーフレーム選択である．本研究で使用した視聴動画は 29.97 のフレームレートであり，2 つの動画のどちらも 6 分を超えるため，全てのフレームを取り出す場合の画像枚数は軽く 1 万枚を超え，データ容量や多くの計算時間がかかってしまう．しかし，例えば，1 秒ごとに切り出した場合，動画の全ての行動を説明した画像が存在しない場合がある．もし，全ての行動を示す画像がない状態で，シーン分類を行ったときには，発話内容に対する画像がないことで，シーン分類の精度に悪影響を与えてしまう．そのため，アニメーション内の速い動きを含め全ての行動を捕捉しつつ，画像枚数が多くなりすぎないよう，人手で確認した結果，本データセットにおいては，5 フレームごとの切り出しが最適だと判断した．

[5]https://megagonlabs.github.io/ginza/
[6]https://spacy.io/

第6章　実験・評価

本章では，提案手法をもとにストーリーテリングデータセットに対して実施した実験の方法と結果を提示し，考察を述べる．

6.1　機械学習モデル

本実験で，使用した機械学習モデルは回帰モデルのSVR（Support Vector Regression：サポートベクタ回帰）であり，カーネル関数にはガウシアンカーネルを選択した．SVRのハイパーパラメータはグリッドサーチにより調整したものを定め，K分割交差検証（K-Fold Cross-Validation）のK=5により，データを5分割し，4つを訓練データ，残り1つをテストデータとしてスコアを求め，同様に求めた計5回のスコアの平均を全体のスコアとした．スコアは決定係数R^2とRMSEを求めた．

6.2　結果

目的変数はスキル評価のQ1〜Q7であり，単位はシーンごととする．7項目あるため，項目ごとに別々で学習モデルを作成した．SVRに入れる特徴量セットを4種類用意して，比較した．1つ目はBERT特徴量768次元の平均と分散をとったもので1536次元（図6.3のBERT），2つ目はBERTの特徴量に加え，CLIPのtextエンコードモデルの特徴量512次元の平均と分散をとったもので計2560次元（図6.3のBERT+CLIP(T)）で，3つ目はBERTの特徴量とCLIPのtextエンコードモデルの特徴量に加え，選ばれた画像をCLIPの画像エンコードモデルに入力したCLIPの画像特徴量の平均と分散をとったもので計3584次元（図6.3のBERT+CLIP(T)+CLIP(I)），4つ目はBERT特徴量にCLIPの画像特徴量を加えたもので計2560次元（図6.3のBERT+CLIP(I)）である．CLIPの画像特徴量の抽出にあたり，シーン推定を行う必要があるがそのときの画像ベクトルとテキストベクトルのcos類似度には0.28の閾値を設けた．閾値により，0.28未満のcos類似度の場合は該当するシーンがないとされ，画像特徴量を得る際のCLIPに入力する画像としてカウントされることはない．閾値の0.28という数値は図6.1で示すように閾値を0.25〜0.35まで0.01ずつ変化させた際に十分の画像枚数を得ることができ，かつ発話文章数との相関が低いものとした．発話文章数との相関が低

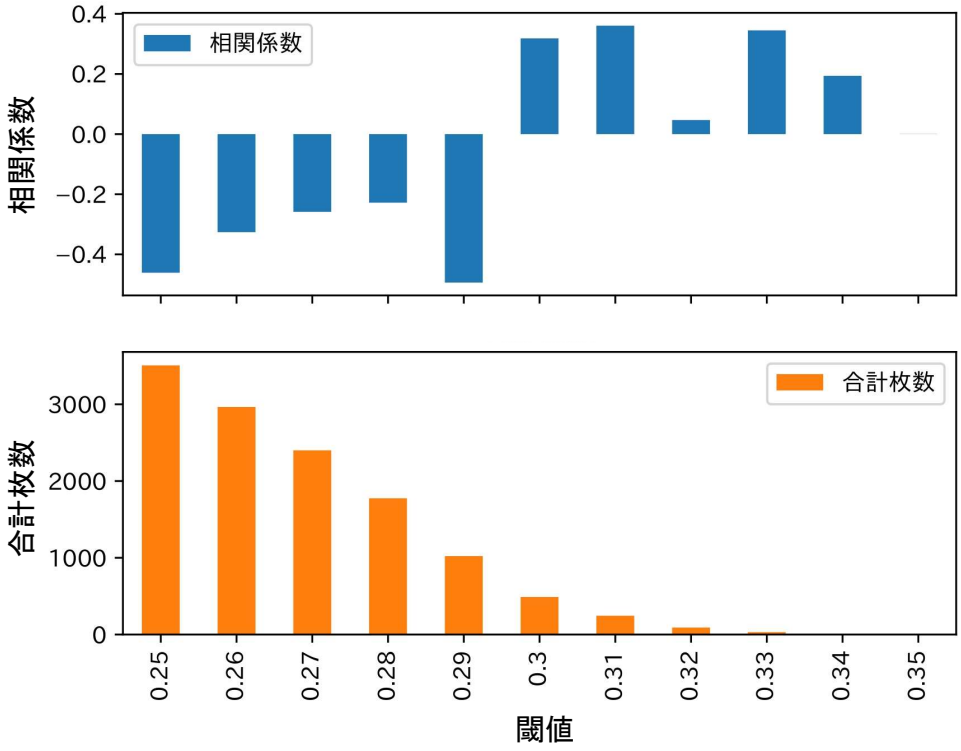

図 6.1: 閾値の変化：発話量との相関と選択された画像枚数

い閾値を選択することで，話者が大量に発話したことにより多くの画像が選ばれてしまうことを避けており，CLIP の画像特徴量と発話文章数とが異なる特徴量であることを示し，本手法が CLIP を使用した独自的な手法であることを示唆している．RMSE は図 6.2 で示す通り，各特徴量で最大でも 1.4 を上回る項目がなく，平均は 1.1 を下回る結果となった．質問項目の点数が 1 点から 7 点 8 段階評価であることを踏まえると比較的良い数値であるといえる．

表 6.1: SVR 決定係数数値

評価項目	BERT	BERT+CLIP(text)	BERT+CLIP(text)+CLIP(image)	BERT+CLIP(image)
Q1: 自信	0.051	0.038	0.052	0.067
Q2：流暢さ	0.064	0.074	0.041	0.032
Q3：語彙	0.138	0.142	0.131	0.112
Q4：ユーモア	0.362	0.331	0.305	0.297
Q5：的確さ	0.114	0.128	0.093	0.083
Q6：ジェスチャ	0.210	0.196	0.147	0.117
Q7：簡潔さ	0.093	0.082	0.079	0.079

表 6.2: RMSE 数値

評価項目	BERT	BERT+CLIP(text)	BERT+CLIP(text)+CLIP(image)	BERT+CLIP(image)
Q1: 自信	0.931	0.939	0.935	0.925
Q2：流暢さ	1.252	1.245	1.265	1.272
Q3：語彙	0.982	0.980	0.988	0.999
Q4：ユーモア	1.222	1.252	1.276	1.283
Q5：的確さ	1.061	1.053	1.073	1.079
Q6：ジェスチャ	1.254	1.268	1.304	1.327
Q7：簡潔さ	0.871	0.877	0.879	0.878

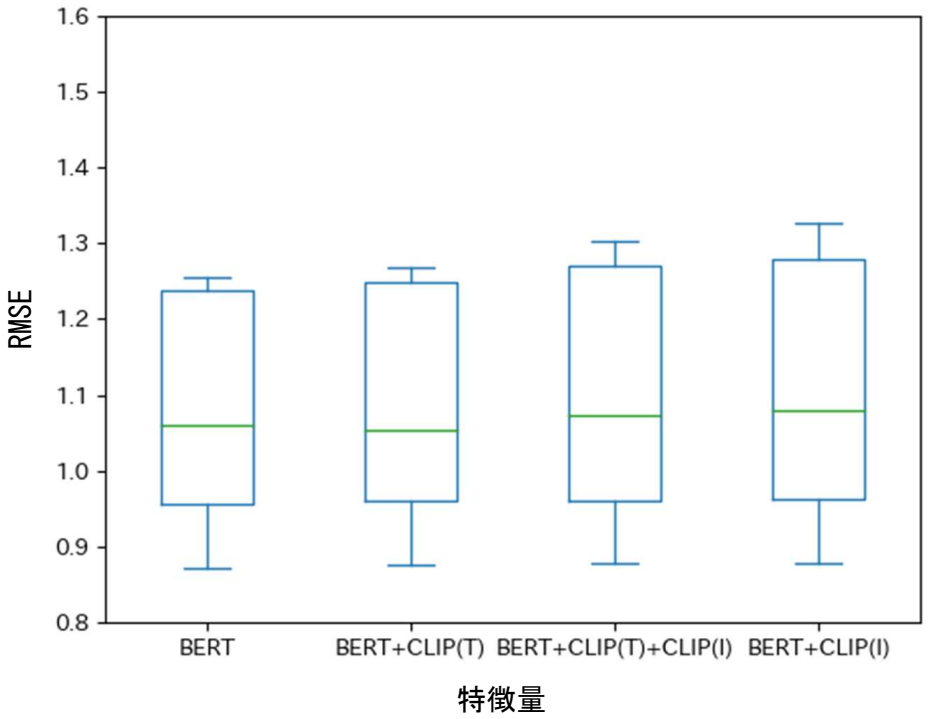

図 6.2: RMSE 箱ひげ図

6.3 考察

最も高い精度で予測することができた Q4 は「説明の中にユーモアがあった」という評価項目である．BERT 特徴量だけで作られた特徴量の BERT が高い数値を示したことから，Q4 のユーモアの有無は言語的な要素が大きく関連し，言葉のボキャブラリや文の構成が影響していると考えられる．BERT+CLIP(I) が最も高い精度となった Q1「自信をもって説明していた」では，BERT+CLIP(T)+CLIP(I) よりも BERT+CLIP(T)+CLIP(I) から CLIP の言語特徴量を抜き，BERT 特徴量と CLIP の画像特徴量を持つ BERT+CLIP(I) の方が精度が高くなっていることから話者が話した内容といった言語的な特徴だけでは，推定には不十分であり，画像特徴量といった別の特徴により情報を与える必要があるといえる．Q5「情景を上手く言葉で言い表せており，正確に物語の内容（情報）を伝えていた」において

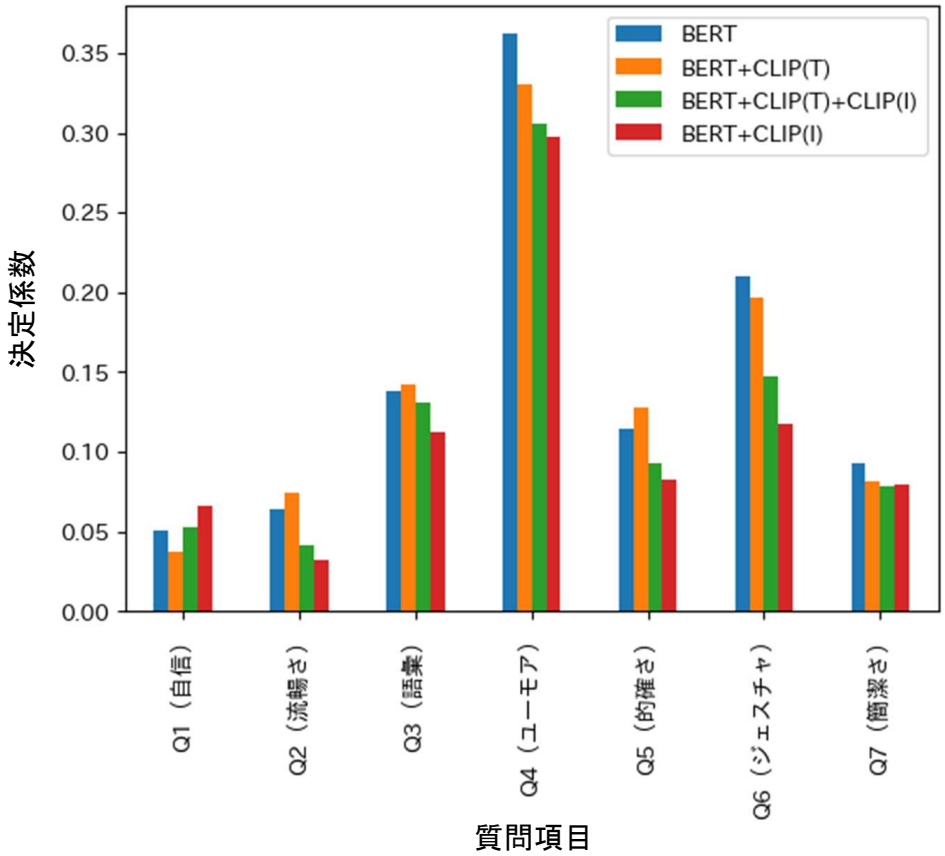

図 6.3: SVR 決定係数グラフ

は，BERT 特徴量に CLIP の言語特徴量を加えた BERT+CLIP(T) が唯一 BERT のみの決定係数を超え，最も高い精度を出した．理由として Q5 が他の評価項目とは異なり，情景といった画像の情報に関わる内容を言語化するスキルが必要であり，CLIP が画像との関わりを持ちながら学習を進めるモデルであることが影響したからではないかと考える．

6.3.1 ストーリーテリングの上手さ

本実験を通して，CLIP の特徴量を加えたことで BERT 特徴量のみの学習よりも精度が向上した項目は，Q1「自信をもって説明していた」と Q5「情景を上手く言葉で言い表せており，正確に物語の内容（情報）を伝えていた」であった．CLIP は学習時に言語と画像の両方を利用して学習するモデルであり，相互に関わりあっている．そのため，テキストエンコーダであっても単純にテキストのみに影響される訳ではなく，画像エンコーダにおいても同様である．Q5 の精度が向上した理由として CLIP のテキストエンコーダが画像との関わりも持っているからではないかと述べたが,「情景を上手く言葉で言い表せたか」と書いてあるように話者は

ストーリーテリングを行う際にどれだけその情景をイメージで来ていたのかが評価されたものと考えられる．そこから考えるに，Q1 において CLIP の画像特徴量を入れた BERT+CLIP(I) が最も高い精度を示したことは自信の有無に元動画から抽出された画像が関わっている．つまり，自信をもって動画の内容を説明するということは，話者が話す内容を頭の中に画像として正しくイメージできていることが影響していると考えられる．ストーリーテリングの上手さとして，発話という言語的表現は重要であるが，それだけでなく，適当な画像を思い浮かべることも上手さの 1 つの要素になるとわかった．

第7章　おわりに

7.1　まとめ

本研究では，特定の動画を説明するストーリーテリングのデータを使って，話者のストーリーテリングスキルの推定とストーリーテリングの上手さについて実験・考察した．ストーリーテリングスキルの評価項目は7つあり，それぞれ1を最も低い評価として，8段階で点数をつけ，各項目ごとの点数をスキル推定の際の目的変数に定めた．説明変数に用いる特徴量は，BERTの最終層の1つ前から取り出した言語特徴量とCLIPのテキストエンコーダの出力ベクトルの言語特徴量とCLIPの画像エンコーダの出力ベクトルの画像特徴量の3種類を組み合わせて作成した．モデルには，ガウシアンカーネルのSVRを使用し，5分割交差検証により決定係数 R^2 とRMSEを求めた．実験結果では，BERTのみの特徴量が多くの評価項目で最も高い精度となったが,「情景を上手く言葉で言い表せており，正確に物語の内容（情報）を伝えていた」といった項目では，BERTに加えてCLIPの言語特徴量を使用したときに最も高い精度を示した．CLIPでは，学習の段階でテキストエンコーダであっても画像の影響を受ける．このことから，ストーリーテリングを行う際に情景をイメージできているかが評価されていると考えた．また,「自信をもって説明していた」という評価項目でも，BERTのみの特徴量よりもCLIPの画像特徴量を加えた特徴量が最も高い精度となった．自信の有無という一見画像とは関連のない項目において，CLIPの画像特徴量が効いたことは，話者が自信をもって話す際には，頭の中に画像を明確にイメージしているためではないかと考察した．

7.2　今後の展望

今回使用したモデルはSVRであり，特徴量を作成する際にも時系列は考慮していない．機械学習モデルにLSTMといった時系列を扱えるモデルを使用することで，時系列に発話を入力した際にシーン推定画像も時系列に並べ，発話とシーン推定画像を比較することで，どれだけ話にまとまりをもって話すことができているかが数値化でき，精度をより向上できると考える．また，CLIPの画像特徴量を使用したSVRでは7つの質問項目の内1つの項目でしか精度の向上が見られなかった．理由としては，CLIPの精度がまだ低いことで，発話内容と画像との誤った結

び付けが行われていることが考えられる．今後は，ストーリーテリングスキルの解明のために，時系列データの対応と共に，評価項目との相関をみながらCLIPを使用した別の特徴量を作成していくことが必要である．

謝辞

本研究を行うにあたり，実験の発話内容書き起こしを協力いただいた株式会社インターグループの方々に深く感謝申し上げます．また，主指導教員の岡田将吾准教授から多大なるご指導があったことを感謝いたします．そして，本研究について様々な面でご協力頂いた皆様には深く御礼申し上げます．

参考文献

[1] Sutinen E. Suhonen J. et al. Lugmayr, A. *Serious storytelling – a first definition and review. Multimed Tools.* 2016.

[2] NATIONAL SROTYTELLING NETWORK. What is storytelling? `https://storynet.org/what-is-storytelling/`. (Accessed on 01/31/2024).

[3] Carol Haigh and Pip Hardy. Tell me a story — a conceptual exploration of storytelling in healthcare education. *Nurse Education Today*, Vol. 31, No. 4, pp. 408–411, 2011.

[4] David A. Gilliam and Karen E. Flaherty. Storytelling by the sales force and its effect on buyer–seller exchange. *Industrial Marketing Management*, Vol. 46, pp. 132–142, 2015.

[5] Shogo OKADA, Mi HANG, and Katsumi NITTA. Predicting performance of collaborative storytelling using multimodal analysis. *IEICE TRANSACTIONS on Information and Systems*, Vol. E99-D, No. 6, pp. 1462–1473, 2016.

[6] 富澤駿. マルチモーダル情報を用いたコミュニケーション中のジェスチャ機能ラベルの推定. 北陸先端科学技術大学院大学, 修士論文, 2020.

[7] Adam Kolides, Alyna Nawaz, Anshu Rathor, Denzel Beeman, Muzammil Hashmi, Sana Fatima, David Berdik, Mahmoud Al-Ayyoub, and Yaser Jararweh. Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts. *Simulation Modelling Practice and Theory*, Vol. 126, p. 102754, 2023.

[8] Muhammad Shahid Jabbar, Jitae Shin, and Jun-Dong Cho. Ai ekphrasis: Multi-modal learning with foundation models for fine-grained poetry retrieval. *Electronics*, Vol. 11, No. 8, 2022.

[9] Honggang Zhao, Guozhu Jin, Xiaolong Jiang, and Mingyong Li. Sde-rae:clip-based realistic image reconstruction and editing network using stochastic differential diffusion. *Image and Vision Computing*, Vol. 139, p. 104836, 2023.

[10] Tatsuya Matsushima, Yuki Noguchi, Jumpei Arima, Toshiki Aoki, Yuki Okita, Yuya Ikeda, Koki Ishimoto, Shohei Taniguchi, Yuki Yamashita, Shoichi Seto, Shixiang Shane Gu, Yusuke Iwasawa, and Yutaka Matsuo. World robot challenge 2020 – partner robot: A data-driven approach for room tidying with mobile manipulator, 2022.

[11] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory, 2023.

[12] Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action, 2022.

[13] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

[14] Takuya Igaue and Ryusuke Hayashi. Signatures of the uncanny valley effect in an artificial neural network. *Computers in Human Behavior*, Vol. 146, p. 107811, 2023.

[15] 細馬宏通. 非言語コミュニケーション研究のための分析単位 : ジェスチャー単位 (連載チュートリアル; 多人数インタラクションの分析手法〔第 5 回〕). 人工知能, Vol. 23, No. 3, pp. 390–396, 2008.

[16] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. The University of Chicago Press, 1992.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[18] 小林一郎折口希実. 敵対的サンプルを用いた対照学習の性能向上への取り組み, 2022.

[19] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[20] sonoisa(園部 勲). 【日本語モデル付き】2022年にマルチモーダル処理をする人にお勧めしたい事前学習済みモデル. https://qiita.com/sonoisa/items/00e8e2861147842f0237. (Accessed on 01/29/2024).

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[22] M. V. Koroteev. Bert: A review of applications in natural language processing and understanding, 2021.