筑波大学 情報学群 情報メディア創成学類

卒業研究論文

大規模言語モデルに基づく
語用論的概念ラベリング

清野 駿

指導教員　若林 啓

2024 年 2 月

概要

本研究は，多数の画像から特定の一枚を選ぶ際に役立つ自然言語ラベルを生成するための，大規模言語モデルを用いた新しい手法を提案する．

近年，大量のテキストデータの学習によって高品質なテキストの生成を行う大規模言語モデルの発展が目覚ましい．大規模言語モデルは，テキストデータが用いられるすべての分野において活用される可能性を持っており，大規模言語モデルの性能を向上させる試みは，それらの分野や産業界に対する大きな貢献になるといえる．さて，大規模言語モデルについて，モデルに入力するプロンプトの方式の工夫やモデル同士の対話により，出力されるテキストの品質が向上することが知られている．また，大規模言語モデルは，入力された画像に対して画像を説明するようなラベルを生成する能力を有している．しかし，複数の画像に対して画像同士を区別可能なラベルを生成する「画像情報ラベリング」に対しては，大規模言語モデルの対話による出力品質の向上の検証は十分に行われていない．加えて，画像情報ラベリングを解決するためのフレームワークとして語用論的アプローチによる手法が研究されているが，ラベルの生成および評価を行うモデルの学習コストが大きいという課題が存在する．

そのため本研究では，大規模言語モデル同士の対話による性能向上が，画像情報ラベリングに対して有効であるかを検証する．そのたに，既存の語用論的アプローチにおけるラベルの生成モデルと評価モデルに大規模言語モデルを導入した，PCG（Pragmatic ChatGPT）を提案する．実験では，提案手法とベースライン手法によって生成したラベルを用いて被験者実験を行い，それぞれの手法によって生成されたラベルの品質を評価した．実験の結果，一定の条件下で提案手法はベースライン手法を上回る傾向がみられた．このことから，大規模言語モデル同士の対話による出力の品質情報は，画像情報ラベリングに対しても有効に働く傾向にあることが示された．

目 次

第 1 章　はじめに **1**

第 2 章　関連研究 **2**
2.1　画像に基づくテキスト生成 2
2.2　語用論的アプローチに基づく機械学習 3

第 3 章　提案手法 **4**
3.1　問題設定 4
3.2　顕著性と識別性について 4
3.3　話し手モデルと聞き手モデル 6
3.4　Pragmatic ChatGPT（PCG）手法 6

第 4 章　評価実験 **13**
4.1　実験設定 13
4.1.1　データセット 14
4.1.2　Amazon Mechanical Turk：AMT の概要 15
4.1.3　比較手法 16
4.2　実験結果と考察 18
4.2.1　実験結果 18
4.2.2　考察 19

第 5 章　おわりに **20**

謝辞 **21**

参考文献 **22**

図目次

3.1 ラベリング対象画像 5
3.2 識別対象画像 5
3.3 ラベル改善フロー 7
3.4 話し手モデル 1（初期ラベル生成フロー） 7
3.6 出力されるラベルの例 7
3.5 初期ラベル生成プロンプト．画像 10 枚が入力され，各画像を区別するようにラベルを生成する．出力は”1.[Caption for image 1] 2.[Caption for image 2] 10.[Caption for image 10]”の形式で行われる． 8
3.7 聞き手モデル（ヒント生成フロー） 8
3.9 出力されるラベル改善のヒントの例 9
3.10 話し手モデル 2（ラベル改善フロー） 9
3.8 ラベル改善のヒントを生成する処理を行うプロンプト．10 枚の画像と 10 文のラベルが入力され，語解の原因となるラベルの組み合わせを出力する．出力は”1. [caption n] and [caption n]: [reference to possible misunderstanding]. hint2. [caption n] and [caption n]: [reference to possible misunderstanding]. hint n. [caption n] and [caption n]: [reference to possible misunderstanding].”の形式で行われる． 11
3.11 ヒントに基づいて改善したラベルを生成する処理を行うプロンプト．10 枚の画像，10 文のラベル，ラベル改善のヒントが入力され，ヒントに基づいて改善を行ったラベルを 10 文出力する．出力は”1.[Caption for image 1] 2.[Caption for image 2] 10.[Caption for image 10]”の形式で行われる． 12

4.1 実験の流れ 13
4.2 同一地点で撮影した画像 18 枚を結合したパノラマ画像 15
4.3 被験者に表示される AMT タスクの例 16
4.4 ラベル長に 10 単語以内の制限を設けるプロンプト 17
4.5 出力されるラベルの例 17
4.6 ベースライン手法で使用するプロンプト 18

第1章　はじめに

近年，膨大なデータと深層学習によって構築された，事前学習済みのマルチモーダルなニューラルモデルである大規模言語モデルが注目を浴びている [1]．その活用はテキストデータを使用するすべての分野で行われ，医療分野での医学的証拠の要約など，大量にあるデータの抽出や構造化等で大きな貢献を果たしている [2]．そのため，大規模言語モデルの性能を向上させる試みは，それらの分野や産業界に対する大きな貢献になるといえる．これらのモデルは，モデルの規模（パラメータ数）と訓練データの量を増やすことで，その品質が向上し続ける傾向にある [3].

さて，大規模言語モデルについて，モデルに入力するプロンプトの方式の工夫やモデル同士の対話により，出力されるテキストの品質が向上することが報告されている [4, 5, 6, 7, 8]．また，大規模言語モデルは，入力された画像に対して画像を説明するようなラベルを生成する能力を有している．しかし，複数の画像に対して画像同士を区別可能なラベルを生成する「画像情報ラベリング」に対しては，大規模言語モデルの対話による出力品質の向上の検証は十分に行われていない．加えて，似た画像の区別を目的とした画像キャプション生成の手法の研究は行われている [9, 10, 11, 12] が，これらの研究では画像キャプション生成モデルの訓練を行っていることから，大規模言語モデルを活用したキャプション生成を行うことが出来ないという課題がある．

本研究の目的は，画像情報ラベリングにおいて大規模言語モデルの潜在能力を活用し，加えて語用論的アプローチを組み合わせることにより，ラベリングの精度を向上させることにある．語用論的アプローチとは，聞き手の解釈を考慮して誤解を避ける言語を生成する手法であり，複数のテキスト生成タスクにおいて品質の向上をもたらすことが報告されている．提案される手法は，大規模言語モデルによって生成されたラベルの中から，誤解を引き起こしやすいものを特定し，これを再び大規模言語モデルで修正するというアプローチに基づいている．

実験においては，複数の部屋の画像を用い，それぞれに対してラベルを生成した．このラベルの評価には Amazon Mechanical Turk を利用し，人間による誤解の程度を検証した．実験結果より，テキスト長が短いラベルにおいて，提案手法によってラベルの品質が向上する傾向にあることが示された．これは，大規模言語モデルの対話による性能向上が，複数の画像に対する区別可能なラベルの生成において有効である可能性を示唆している．

第2章　関連研究

LLM のプロンプト工夫と会議による性能向上 LLM による画像キャプショニング

本章では，本研究で扱う画像情報ラベリングタスクに関連する研究として，画像に基づくテキスト生成技術について 2.1 節で述べる．また，提案手法の着想に関連する研究として，語用論的アプローチに基づく機械学習手法について 2.2 節で述べる．

2.1　画像に基づくテキスト生成

画像内容に基づくテキスト生成技術の発展は，画像キャプション生成の研究によって牽引されてきた [13]．特に，深層学習を用いた画像キャプション生成では，ニューラルネットワークに画像の特徴を抽出し，得られた特徴に基づいて再帰型ニューラルネットワークでテキストを生成するモデルの枠組みが用いられる [14]．この枠組みにおいて Xu ら [15] は，注意機構を用いて画像内の注目すべき箇所の情報を抽出することで，生成されるキャプションの品質が向上することを示した．

一方近年，タスクによらない画像やテキストの表現学習によって，深層学習の性能が大幅に向上することが明らかになってきた [3]．この知見を発展させて，画像とテキストのマルチモーダルモデルを構成し，これを大規模な画像とテキストのデータセットを用いて特徴量表現を学習する手法が研究されている [1, 16]．事前学習済みの大規模マルチモーダルモデルを用いることで，画像キャプション生成の品質が向上する [17] だけでなく，画像のゼロショット分類 [18] や，画像質問応答 [19] などのタスクの精度も向上することも示されている．

さらに近年では，モデル規模と訓練データをスケールアップすることで，テキストでの指示（プロンプト）に応じた画像に基づくテキスト生成ができることが明らかになった．Li ら [20] は，事前訓練済みの画像特徴量抽出器とテキスト生成器の間を橋渡しする Querying Transformer を訓練する手法によって，画像に関する自然な会話や，柔軟な指示に応じた説明の生成ができることを示した．GPT-4[21] は，入力した画像を描画するプログラムを出力したり，グラフの解釈もできることを示している．例えば,「この画像の特徴を 10 個挙げてください」というプロンプトと共に画像を入力することで，自然言語テキストで画像の特徴を得ることができる．

これらの深層学習に基づく手法はブラックボックスであり，どこまでのタスクを実現できるかは未知数であるが，基本的には 1 枚あるいは少数の画像とテキストの間を結びつける処理を想定している．このため，多数の画像に対して，それらを区別する自然言語ラベルを生成する問題設定における性能は十分に検証されていない．本研究では，画像情報ラベリングタスクにおいて，これらのテキスト生成手法の性能を評価すると共に，画像情報ラベリング

タスクを分割して再構成することによるタスク性能の向上を検討する．

2.2　語用論的アプローチに基づく機械学習

言語の生成にあたって，聞き手の解釈を考慮して，誤解を避けるような語を選ぶことは，コミュニケーションにおいて重要な要素である．このような，話し手による聞き手の解釈を考慮したアプローチは，計算語用論と呼ばれる [22]．Rational Speech Acts (RSA) フレームワークは，計算語用論の代表的な計算モデルである [23, 24]．RSA フレームワークでは，話し手と聞き手のモデルが，それぞれ相手の目的や解釈について再帰的に推論する過程を計算する．例えば，RSA フレームワークにおける話し手モデルは，内部的に持つ聞き手モデルに発話を解釈させたときに，伝達したい意図通りの解釈が得られるような発話を選択する．

この知見を反映して，RSA フレームワークに基づいてテキスト生成タスクの品質を向上させる手法が研究されている [25]．William ら [26] は，室内のナビゲーションタスクにおいて，計算語用論に基づく手法を提案した．話し手モデルによって生成された複数の指示を聞き手モデルに入力し，聞き手モデルが意図通りに解釈する指示を出力する．SAIL データセットを用いた評価により，語用論的アプローチによってナビゲーションの成功率が向上することが示された．

画像キャプション生成においても，語用論的アプローチによって，聞き手が似た画像と区別できるようなキャプションを生成する手法が提案されている [9, 10, 11]．Andreas ら [12] は，画像の参照ゲームにおいて，聞き手の誤解を低減するようなキャプションを生成する手法を提案した．これらの手法は，ニューラルネットワークでモデル化された話し手と聞き手の学習を伴うことから，近年極めて高い品質を達成している大規模事前学習済みモデルを活用することが難しい．本研究では，大規模事前学習済みモデルを再学習することなく用いる語用論的アプローチによって，誤解可能性を考慮した画像情報ラベリングを実現する手法を検討する．

第3章　提案手法

3.1　問題設定

本研究では，入力された画像の集合に対して，各画像を誤解なく一意に識別可能なラベルを生成する問題に取り組む．

n 枚の画像の集合を $X = \{x_1, x_2, \ldots, x_n\}$ とする．X に対して付与されるラベルの集合を $Y = \{y_1, y_2, \ldots, y_n\}$ とする．ラベル $y_1, y_2, \ldots, y_n$ は画像 $x_1, x_2, \ldots, x_n$ にそれぞれ対応しており，各ラベルは対応する画像の特徴を説明するものとする．

ここで，ラベル y_r（$r \in [1..n]$）と X を被験者に与えたとき，被験者が y_r に対応する画像 x_r を選択しようとして，x_r 以外を選択してしまう確率を誤解可能性 p とする．

提案手法では，p が最小となるラベルの組み合わせ Y を生成することを目的とする．

3.2　顕著性と識別性について

ラベルの誤解可能性を決定づける要素として,「顕著性」と「識別性」を導入する．

顕著性は，ラベルが対応する画像の特徴をどれだけ正確に捉えて表現できているかを指すパラメータである．一方，識別性は，ラベルに対応する画像がそれ以外の画像からどれだけ一意に識別可能かを指すパラメータである．

顕著性と識別性は互いに独立したパラメータである．そのため,「顕著性は高いが識別性は低いラベル」等も存在する．

顕著性と識別性について具体的な例を示す．図 3.1，3.2 が与えられた際に，図 3.1 を誤解なく一意に特定できるラベルを生成する，という問題を考える．ここではラベルによって説明を行う画像をラベリング対象画像とする．また，ラベリング画像とは別の画像，即ち誤解が生じぬように区別すべき画像を識別対象画像とする．

図 3.1: ラベリング対象画像

図 3.2: 識別対象画像

この問題における，顕著性と識別性の有無を反映したラベルを，それぞれ以下に示す．

顕著性と識別性どちらも低いラベル　「広い部屋」．ラベリング対象画像の特徴を十分に説明できておらず，かつ識別対象画像との区別もできていない．

顕著性が高く，識別性が低いラベル　「天井に張り巡らされた木製の梁と，天井から吊るされたシャンデリアが特徴的な，白い壁の寝室」．ラベリング対象画像の説明は十分できているが，識別対象画像にも同様の特徴が含まれており，2 画像の区別ができていない．

顕著性が低く，識別性が高いラベル　「暖炉のある部屋」．ラベリング対象画像のみに存在する特徴「暖炉」を表現することで，識別対象画像との区別に関してはできている．しかし，ラベリング対象画像に含まれる特徴の中で特徴「暖炉」が画像を占める面積は小さく，画像の特徴を十分に説明できているとは言えない．

顕著性と識別性どちらも高いラベル　「木製の梁とシャンデリアが特徴的な，真っ白なベッドのある寝室」．部屋の特長を十分に説明しており，かつ識別対象画像との区別ができる特徴「真っ白なベッド」を含んでいる．

3.3　話し手モデルと聞き手モデル

本研究では，語用論に基づくアプローチを用いて，ラベルの品質を改善する．語用論は，言語交渉の過程において話し手と聞き手の意図の相互理解を探求する学問分野であり，特に皮肉表現や比喩表現などの複雑な言語現象を研究対象としている．本研究における語用論的アプローチは，話し手モデルと聞き手モデルの 2 つの対話モデルに基づいて構築される．話し手モデルと聞き手モデルの双方向の対話を通じて，誤解のないラベルの生成を目指す．

話し手モデルの目的は，画像を正確に表現したラベルの生成である．このモデルでは，画像の主要な特徴を十分に説明し，ラベルの曖昧性を排除することを重視する．

一方，聞き手モデルは，話し手モデルによって生成されたラベルが誤解を招く可能性があるかを評価する．このモデルはラベルが提供する情報に関して，他のラベルとの間で起こり得る誤解の有無を検証し，その結果を話し手モデルに対してフィードバックする．

3.4　Pragmatic ChatGPT（PCG）手法

本研究では，前節における話し手モデルと聞き手モデルの両方に ChatGPT を導入した，Pragmatic ChatGPT（PCG）を提案する．

PCG の目的は，ラベルの識別性を高めることである．PCG では，話し手モデルによるラベルの生成と，聞き手モデルによるラベルの評価を行い，この 2 つのモデルの処理の反復により，ラベルの誤解可能性を段階的に減少させる．

PCG におけるラベル改善処理のプロセスを図 3.3 に示す．ラベル改善処理では，初めに，話し手モデル 1 が対象画像に対する初期ラベルを生成する．続いて，この初期ラベルに対し，聞き手モデルが改善に向けた具体的なヒントを提供する．これらのヒントを基に，話し手モデル 2 は改良されたラベルを作成し，それを再び聞き手モデルへと提出する．このプロセスの繰り返しにより，ラベルの誤解可能性を徐々に低減させ，ラベルの識別性を向上させることを目指す．各サイクルは，ラベルの精度と明確性の向上に寄与し，最終的にはより効果的な画像認識と解釈を可能にすることが期待される．

図 3.3: ラベル改善フロー

図 3.3 における，話し手モデル 1（初期ラベル生成フロー）の処理の流れを図 3.4 に示す．話し手モデル 1 は，入力された画像の集合に対して，各画像の特徴を説明するようなキャプション生成を行う．このとき，話し手モデル 2 にあるような，ラベル改善のヒントを踏まえた生成などは行わず，純粋なキャプショニングによってラベルを生成する．

図 3.4: 話し手モデル 1（初期ラベル生成フロー）

話し手モデル 1 で行っている，初期ラベルを生成する処理を行う ChatGPT プロンプトを図 3.5 に示す．

出力されるラベル例を図 3.6 に示す．

1. A bedroom with large windows offering a view, a cozy sitting area, and a neutral color palette.

図 3.6: 出力されるラベルの例

次に図 3.3 における聞き手モデル（ヒント生成フロー）の処理の流れを図 3.7 に示す．聞き

```
Prompt: Now input 10 images. For each group of images you have
entered, output a sentence that captures three characteristics of
 the images, one sentence for each image, so that each image can
be distinguished. Never output in any format other than the
specified format. The output format should be as follows. 1.[
Caption for image 1] 2.[Caption for image 2] ~~ 10.[Caption for
image 10] Example: 1. A living room, white walls, wallpaper with
pink and black.

Responce: Please upload the 10 images you'd like me to analyze.

Prompt: {Image URLs}
```

図 3.5: 初期ラベル生成プロンプト．画像 10 枚が入力され，各画像を区別するようにラベルを生成する．出力は"1.[Caption for image 1] 2.[Caption for image 2]　10.[Caption for image 10]"の形式で行われる．

手モデルは，入力された画像の集合と，話し手モデルによって生成されたラベルから，各ラベルの誤解可能性について評価を行い，ラベルの改善に役立つヒントを出力する．

図 3.7: 聞き手モデル（ヒント生成フロー）

聞き手モデルが生成するヒントは，誤解する可能性の高い 2 つのラベルの組み合わせを示す形式をとる．例えば，ラベル 1 とラベル 3 に同じ単語が含まれており，それぞれのラベルに対応する画像にも同様の特徴が含まれていた場合，そのラベル組の識別性は互いに低くなると考えられる．そのような場合，そのラベルの組の誤解可能性が高いとして，改善すべきラベル組はラベル 1 とラベル 3 である，といったようにラベル改善のヒントを出力する．

聞き手モデルで行っている，ラベル改善のヒントを生成する処理を行う ChatGPT プロンプトを図 3.8 に示す．

出力されるヒントの例を図 3.9 に示す．

hint 1. Caption 3 and Caption 10: Both mention contemporary elements with large windows and a spacious layout, which could refer to modern architecture with a minimalist theme.

図 3.9: 出力されるラベル改善のヒントの例

最後に図 3.3 における話し手モデル 2（ラベル改善フロー）の処理の流れを図 3.10 に示す．話し手モデル 2 では，画像の集合，既に生成されているラベル，聞き手モデルによって生成されたラベル改善のヒントをもとに，新たなラベルを生成する．このとき，ラベル改善のヒントをもとにして既に生成されているラベルの改善を行うことで，よりラベルの誤解可能性を小さくすることを狙う．

図 3.10: 話し手モデル 2（ラベル改善フロー）

話し手モデル 2 で行っている，ヒントに基づいて改善したラベルを生成する処理を行う ChatGPT プロンプトを図 3.11 に示す．

画像の集合を X，n 回目の改善サイクルにおけるラベルを Yn，n 回目の改善サイクルにおけるヒントを H，改善サイクルを回す回数を m として，ここまでのラベル改善処理をまとめた流れは以下の通りである．

1. ChatGPT に対して画像の集合 X を入力する．X に含まれる各画像に対して，それぞれ誤解なく画像を区別できるラベルを出力するように，プロンプトを入力する．ここで出力されたラベルの集合を Y1 とする．

2. ラベルの集合 Y1 と画像の集合 X を ChatGPT に入力する．Y1 に含まれる各ラベルについて誤解可能性を分析し，ラベル改善のヒントを出力するように，プロンプトを入力する．出力されたラベル改善のヒントを H1 とする．

3. ChatGPT に対して画像の集合 X，ラベル改善のヒント H1，ラベルの集合 Y1 を入力する．H1 を参考にして，Y1 を改善した新たなラベルを出力するように，プロンプトを入力する．ここで出力されたラベルの集合を Y2 とする．

4. 2 と 3 を繰り返して，最終的に得られたラベルの集合 Ym（m は任意の自然数）を，誤解可能性の低いラベルとして提出する．

```
Prompt: Now enter 10 images and 10 captions corresponding to
each image with 10 sentences.For each input image and caption
group, verify that each caption can distinguish each image
without misunderstanding. Then, output the combination of
captions that you think is most likely to cause misunderstanding
in the form of [Caption 1 and Caption 2]. The input format is as
follows. 10 image files: to be entered later. Caption 10
sentences: [caption number]. [Caption text]. As an example, the
following will be input. 1. red bed room, white curtains, yellow
modern lamps 2. ~~ omitted ~~ 10. red bed room, blue chairs, red
wall.  Output format to be described later

Responce: Please enter 10 pairs of image files and their
corresponding captions. For each image, we will review the
potentially misleading captions and identify the most confusing
caption combinations. Once you have provided us with the images
and captions, we will begin our analysis.

Prompt: {Image URLs}

Responce: To proceed with the analysis, please provide the
captions for each image. Once I have the captions, I will be able
 to review them in the context of the images and identify any
potential for confusion.

Prompt:
{labels}

These are the captions for each image. Never output in any
format other than the specified format. Output should be in
accordance with the following output format. hint 1. [caption n]
and [caption n]: [reference to possible misunderstanding]. hint2.
 [caption n] and [caption n]: [reference to possible
misunderstanding]. ~~ hint n. [caption n] and [caption n]: [
reference to possible misunderstanding]. Example: hint 1. Caption
 1 and Caption 10: Both highlight colors (pink and red) and may
refer to the art piece.
```

図 3.8: ラベル改善のヒントを生成する処理を行うプロンプト．10 枚の画像と 10 文のラベルが入力され，語解の原因となるラベルの組み合わせを出力する．出力は”1. [caption n] and [caption n]: [reference to possible misunderstanding]. hint2. [caption n] and [caption n]: [reference to possible misunderstanding]. hint n. [caption n] and [caption n]: [reference to possible misunderstanding].” の形式で行われる．

```
Prompt:
Now enter 10 images and 10 captions corresponding to each image.
In addition, enter a hint referring to possible misunderstandings
 about each caption.For each group of images you have entered,
improve the captions based on the hints and output one sentence
for each image, capturing three characteristics of each image
that distinguish each image.Input 10 images, 10 caption sentences
, and the hints in that order.


Responce:
Understood. Please enter the 10 images, a caption for each, and
a hint to prevent misunderstandings about each caption in order.
We will then provide improved captions that capture the
characteristics of each image based on the hints.


Prompt: {Image URLs}

Responce:
Continue by providing a caption for each image and tips to
prevent misunderstandings about that caption. Based on this
information, create an improved caption that captures three
features of the image.


Prompt: First, the captions corresponding to each image are as
follows.

{labels}

Never output in any format other than the specified format.
Output should be in accordance with the following output format.
hint 1. [caption n] and [caption n]: [reference to possible
misunderstanding]. hint2. [caption n] and [caption n]: [reference
 to possible misunderstanding]. ~~ hint n. [caption n] and [
caption n]: [reference to possible misunderstanding]. Example:
hint 1. Caption 1 and Caption 10: Both highlight colors (pink and
 red) and may refer to the art piece. Below are some tips on
misconceptions to avoid." +

{hints}

Based on these tips, improve the original caption so that the
caption captures three features of each image that distinguish
each image. Never output in any format other than the specified
format. The output format should be as follows. 1.[Caption for
image 1] 2.[Caption for image 2] ~~ 10.[Caption for image 10]
Example: 1. Example: 1. A living room with wood floors, white
walls and ceiling, and wallpaper with pink and black patterns.
```

12

図 3.11: ヒントに基づいて改善したラベルを生成する処理を行うプロンプト．10 枚の画像，10 文のラベル，ラベル改善のヒントが入力され，ヒントに基づいて改善を行ったラベルを 10 文出力する．出力は”1.[Caption for image 1] 2.[Caption for image 2]　10.[Caption for image 10]”の形式で行われる．

第4章　評価実験

4.1　実験設定

本実験では，各手法の性能を比較するため，生成されたラベルの誤解可能性を検証する．被験者に複数の画像と 1 つのラベルを与え，対応する画像を選択できるかを調査する．

実験全体の流れをまとめた図を図 4.1 に示す．図の左部は手法によるラベルの評価部分で，右部は被験者実験によるラベルの評価部分である．

図 4.1: 実験の流れ

ラベルの生成には以下の 2 つの手法を用いる．

1. ChatGPT を用いたベースライン手法
2. Pragmatic ChatGPT

ChatGPT を用いたベースライン手法は，ChatGPT を用いて画像に対して 1 対 1 でラベルを生成する単純な手法である．ChatGPT を用いたベースライン手法を評価基準とし，これとの比較により提案手法の評価を行う．ベースライン手法の詳細な処理内容は 4.2 節で述べる．

また，実用性と汎用性の評価のために，それぞれの手法でラベル長を制限することによる以下の 2 種類の出力方法でラベル生成を行い，被験者実験による手法の評価に用いる．

1. 通常出力（normal output）：ラベル長に制限を設けずに出力
2. 短出力（minimal output）：ラベル長を制限（例：10 単語以内）して出力

ラベルの生成元となる画像には，Matterport3D データセットを使用する．データセット内に含まれる，同一地点から撮影した部屋の画像を結合することによってパノラマ画像を得る．パノラマ画像を 1 つの部屋を表す概念として扱い，いくつかの部屋のパノラマ画像に対してラベリングを行う．データセットの詳細は 4.1.1 節で述べる．

被験者実験は Amazon Mechanical Turk（AMT）を使用して行う．AMT はマイクロタスク型のクラウドソーシングサービスである．ワーカーに対してタスクをリクエストし，ワーカーがアクションを行い，その結果を回収する．

本実験では，ワーカーに複数の画像とラベルを与え，ラベルに紐づく画像を選択してもらう．AMT タスクの詳細は 4.1.2 節で述べる．

4.1.1 データセット

実験では，Matterport3D データセット[1]を使用する．Matterport3D データセットは，90 の建物から成る，194,400 枚の RGB-D 画像による大規模な RGB-D データセットである．

実験では，データセットに含まれる画像に対してパノラマ画像化処理を行った後，寝室を撮影した画像 10 枚を抽出して実験に使用した．パノラマ画像化処理について，詳細を述べる．データセットは，上下と水平方向の合わせて 3 回，それを 60 度ずつ横方向にずらして 6 回，計 18 回の撮影を同一の地点で行った画像群で構成されている．実験では，同一地点で撮影した 18 枚の画像を適切な位置で結合する処理を，すべての画像に対して行うことで，各撮影地点ごとにパノラマ画像を作成した．

生成したパノラマ画像の例を図 4.2 に示す．

[1] https://niessner.github.io/Matterport/

図 4.2: 同一地点で撮影した画像 18 枚を結合したパノラマ画像

実験では，パノラマ画像の中から寝室を撮影した画像 10 枚を抽出して使用した．

パノラマ画像を実験で使用する理由　ラベリング対象とする画像をパノラマ画像とする理由は，画像に含まれる特徴量を均一化するためである．データセットに含まれる画像のうち，家具や照明など部屋全体の特徴を捉えた画像はラベリングの対象として適切だが，壁のみが写っている等の部屋の特徴を含まない画像はラベリングの対象として適切ではない．このような，画像に含まれる特徴量の不均一によって引き起こされる意図的ではないラベリング精度の低下を回避するために，事前にパノラマ画像化処理を行い画像の特徴量を均一化する．

寝室の画像のみを抽出する理由　ラベリング対象とする画像を寝室だけに制限する理由は，実験の難易度を適切に設定するためである．実験で使用するデータセットに含まれる画像の部屋の種類はキッチンやバスルーム，リビングなど複数存在している．これらの異なる種類の部屋の画像に対してラベリングを行った場合,「調理器具」「ベッド」などの一つ単語で各画像を容易に区別できる可能性がある．このような，画像同士の識別性を不用意に向上させかねない要因を排除するために，データセットから寝室の画像のみを抽出して実験に使用した．

4.1.2　Amazon Mechanical Turk：AMT の概要

Amazon Mechanical Turk（AMT）はマイクロタスク型に特化したクラウドソーシングサービスである．ワーカに対して何らかのタスク（アンケート，データラベリングなど）をリクエストし，ワーカがタスクに対して何らかのアクション（アンケートに回答する，データにラベルを付与するなど）を行い，ワーカが入力したデータを回収することで，問題を解決する．

実験では，4 枚の画像からラベルに対応する画像を選ぶタスクを 100 回行い，その正答率を手法ごとに比較する．

実験で行う AMT のタスクの設定を以下に示す．

- ワーカに対して 4 枚の画像とラベルを与える．
- ワーカは，ラベルに紐づくと考えられる画像を一枚選択する．

AMT タスクの例を図 4.3 に示す．

図 4.3: 被験者に表示される AMT タスクの例

各手法ごとに，正解画像，ラベル，被りなくランダムに抽出した 3 枚の不正解画像の組み合わせによって作成したタスクを 100 回行う．
タスクの作成手順の詳細は以下の通りである．

1. 10 枚の画像と，各画像に対応する 10 個のラベルが手法によって生成される．
2. ある画像が正解となるタスクを 10 タスク作成する．このとき，不正解となる 3 つの画像は，ランダムに被りなく抽出して組み合わせる．
3. 10 枚の画像それぞれに対して 2 の処理を行い，100 通りの組み合わせのタスクを作成する．

4.1.3 比較手法

実験では，4 種類の方法で出力したラベルを使用して比較実験を行う．
ラベルの生成には以下の 2 つの手法を用いる．ベースライン手法との比較により，提案手法の評価を行う．

1. ChatGPT(ベースライン手法)

2. Pragmatic ChatGPT(PCG，提案手法)

さらに，手法の一般性と実用性の評価のために，各手法で以下の 2 種類の出力方法でラベルを生成する．

1. 通常出力（normal output）：ラベル長に制限をかけずに出力

2. 短出力（minimal output）：ラベル長を 10 単語以内に制限し出力

短出力では，図 4.4 に示すプロンプトをラベル出力処理部分に追加することで，ラベル長を 10 単語以内に収めるように制御する．

Sentences must be output within 10 words.

図 4.4: ラベル長に 10 単語以内の制限を設けるプロンプト

ラベル長の制限は，手法の実用性を検証するために行う．無制限に長いラベルを生成することにより，理論上誤解を排除することは可能であるが，これは実用的ではない．従って，実験では，短いラベルによる誤解の可能性を低減できるかどうかについても検証を行う．これは，ラベルの長さが誤解可能性に与える影響を評価するために不可欠である．実際の応用においては，簡潔でありながら効果的なラベルが求められるため，短いラベルにおける性能評価は手法の実用性と妥当性を確認する上で重要である．

2 手法それぞれで通常出力と短出力のラベルを生成し，合計 4 種類のラベル群のセットで被験者実験を行う．実験結果では，各手法で生成したラベル群の平均単語数もあわせて示す．

ベースライン手法によって生成されたラベルの例を図 4.5 に示す．

Bright bedroom with large windows, modern decor, scenic view.

図 4.5: 出力されるラベルの例

このようなラベルの集合を各手法で生成し実験に使用する．

実験で比較する各手法の名前と内容を箇条書きで説明する．

1. ChatGPT

   - ChatGPT を使用して，1 枚の画像を入力して 1 つのラベルを出力する手法．以下のプロンプトを与えて，画像に対するラベルを出力する．

Generate a one-sentence description that captures only three features of the room you have provided.

図 4.6: ベースライン手法で使用するプロンプト

2. Pragmatic ChatGPT（PCG）
    - ChatGPT に「話し手モデル」「聞き手モデル」を導入し，ラベル改善サイクルを 10 回行う手法.
3. ChatGPT minimal
    - ChatGPT に単語長 10 単語までの制約を加えた手法．4.4 のプロンプトを，ラベル出力部分に追加する.,
4. Pragmatic ChatGPT minimal
    - Pragmatic ChatGPT に単語長 10 単語までの制約を加えた手法．4.4 のプロンプトを，ラベル出力部分に追加する.

4.2 実験結果と考察

4.2.1 実験結果

表 4.1 は，手法ごとの AMT タスクの正答率と，タスクに使用したラベルの平均単語数を示している．ChatGPT および ChatGPT minimal がベースライン手法，Pragmatic ChatGPT および Pragmatic ChatGPT minimal が提案手法である.

手法	正答率	平均単語数
ChatGPT	0.43	26.9
Pragmatic ChatGPT	0.40	20.7
ChatGPT minimal	0.31	7.5
Pragmatic ChatGPT minimal	0.33	8.1

表 4.1: 手法ごとの AMT タスクの正答率と平均単語数

表 4.1 より，ラベル長を制限した場合（ChatGPT minimal と Pragmatic ChatGPT minimal），提案手法はベースライン手法の正答率を上回った.

ラベル長に制限が無い場合，提案手法はベースラインの正答率を下回った.

また，すべての比較手法において，平均単語数と正答率に若干の相関がみられた.

4.2.2　考察

本研究における提案手法は，ラベル長に制限がある場合に特に効果を発揮することが明らかになった．ラベル長が限定される状況では，画像の多様な特徴を一つのラベルで表現することには限界がある．この点において，提案手法はラベルの識別性を高める方向でのアプローチを取り，結果としてベースライン手法よりも高い正答率を達成した．このことから，短いラベルでの効果的な識別性の確保が提案手法の大きな強みであると言える．

一方で，ラベル長に制限がない場合の提案手法の性能については，いくつかの問題点が浮き彫りになった．無制限のラベル長では，各画像を包括的に表現することが可能であり，その場合，ラベルの顕著性を高めることがより重要となる．しかし，提案手法が依然として識別性の向上に重点を置いているため，この状況には適切に対応できていない．この結果，ベースライン手法と比較して，提案手法の正答率は低下することが観察された．

これらの考察に基づき，提案手法の改善策として，ラベルの顕著性と識別性のバランスを調整する機構の導入が提案される．具体的には，ラベルが長い場合は顕著性を，短い場合は識別性を優先させることで，ラベル長に応じた適切なラベル生成を行う．この改善により，ラベル長に制限がない状況においても，誤解可能性が低く，より効果的なラベル生成が期待できる．最終的には，この改善によって提案手法がベースライン手法と同等以上の効果を発揮する可能性がある．

第5章　おわりに

本研究では，画像に対するラベル生成の問題を語用論的な観点から解決する試みを行った．特に，話し手モデルと聞き手モデルを導入し，その対話を通じてラベルの改善を図る新たなアプローチを提案した．

提案手法は，話し手モデルが作成したラベルを聞き手モデルが評価し，そのフィードバックに基づいてラベルを再生成するプロセスを繰り返すことで，より適切なラベルの生成を目指す．これにより，生成されたラベルが画像を一意に特定できるものであることが期待される．

また，本研究では生成したラベルの評価に Amazon Mechanical Turk を利用し，生成されたラベルが人間にとってどの程度理解しやすいかを検証した．その結果，提案手法によって生成されたラベルは，特にラベル長に制約がある場合において，ベースライン手法よりも優れた結果を示した．これは，提案手法がラベルの識別性を向上させることに重点を置いて設計されているためである．

今後の課題としては，より多様な画像やシーンに対するラベル生成の実施と，提案手法の汎用性の評価が挙げられる．また，聞き手モデルの評価基準やフィードバックの精度を向上させることも重要である．これらの課題を解決することで，本手法はさらに広範な応用が可能となると考えられる．

謝辞

本研究の一部は，JST CREST (#JPMJCR22M2) の助成によって行われた．

参考文献

[1] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv [cs.CV] 2306.13549*, 2023.

[2] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, Chunhua Weng, and Yifan Peng. Evaluating large language models on medical evidence summarization. *NPJ Digit Med*, Vol. 6, No. 4, p. 158, 2023.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[4] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. pp. 24824–24837, 2022.

[5] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. 2022.

[6] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. 2023.

[7] Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. Progressive-hint prompting improves reasoning in large language models. 2023.

[8] Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations. 2023.

[9] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. 2017.

[10] Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. Pragmatically informative image captioning with Character-Level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 439–443, 2018.

[11] Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. Pragmatic Issue-Sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1924–1938, 2020.

[12] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1173–1182, 2016.

[13] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: generating sentences from images. In *Proceedings of the 11th European conference on Computer vision: Part IV*, pp. 15–29, 2010.

[14] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 39, No. 4, pp. 652–663, 2017.

[15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of Machine Learning Research*, Vol. 37, pp. 2048–2057, 2015.

[16] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-Language pre-training: Basics, recent advances, and future trends. *arXiv [cs.CV] 2210.09263*, 2022.

[17] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in Vision-Language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2136–2148, 2023.

[18] 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. *Proceedings of Machine Learning Research*, Vol. 139, pp. 8748–8763, 2021.

[19] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are Few-Shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6088–6100, 2022.

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image pre-training with frozen image encoders and large language models. 2023.

[21] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv [cs.CL] 2303.12712*, 2023.

[22] Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12619–12640, 2023.

[23] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, Vol. 336, No. 6084, p. 998, 2012.

[24] Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.*, Vol. 20, No. 11, pp. 818–829, 2016.

[25] Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. In *North American Chapter of the Association for Computational Linguistics*, pp. 1951–1963, 2017.

[26] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. Going beyond literal command-based instructions: extending robotic natural language interaction capabilities. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1387–1393, 2015.