

# 大規模言語モデルを使用した日本語クラス図における概念間関係推定

大野 幹人  
芝浦工業大学  
bp20084@shibaura-it.ac.jp

久住 憲嗣  
芝浦工業大学  
hisazumi@shibaura-it.ac.jp

## 要旨

この論文では、大規模言語モデル (LLM) を使用して、日本語のクラス図における概念間の関係を推定する手法を提案する。特に ChatGPT (GPT-4.0) を活用し、システム概要を考慮した推定を可能にすると共に、推定できる関係の種類(多重度・属性)を拡張する。実験では、関係線の有無と関係線の種類(関連, 集約, 継承), 多重度, 属性の推定を行い、それぞれの正解率は 95.1%, 87.2%, 78.0% となり、人間の正解率を超える結果を達成した。結果として、人間の正解率を超えたこと、既存研究の課題を解決したこと、推定できる関係の種類を拡張できたことから本手法は有効であると結論づけられた。

## 1. はじめに

ビジネス環境の急速な変化に伴い、ソフトウェア開発において高品質で迅速なリリースが求められている<sup>[1]</sup>。モデリングは品質向上の一助となるが、大規模なシステムでは工数が膨大となる。そこで要件定義書などから自動的にモデリングを行えるようにすることで工数を削減できる。自動モデリングを行う際に必要な要素としてクラス図の概念間関係推定がある。さらに、2023 年現在、OpenAI 社が開発した ChatGPT を契機に AI ブームが到来しており、急激な速度で LLM (Large Language Models) の精度が向上している。本研究ではこれらの動向を受け、日本語クラス図の概念間関係推定に ChatGPT (GPT4.0) などの最新の LLM を活用した手法を提案する。特に、システム概要を考慮した推定や、新たな種類の関係の推定にも対応することを目指す。

## 2. 関連研究

### 2.1. 日本語クラス図の概念間関係推定

久住らは日本語で作成されたクラス図におけるクラス間の関連, 継承, 集約の推定の精度向上と既存研究の課

題解決を BERT を使用して行った<sup>[2]</sup>。評価実験では精度向上が認められ、提案手法は効果があると結論づけられた。しかし、システム概要を考慮せずに推定を行なった結果、正しく推定ができない場合があるという課題が残った。また、モデルの Fine-tuning には様々なドメインの大量のデータが必要になることから、汎用性にかけるという課題も残った。

## 3. 提案手法

本研究ではシステム概要を考慮した日本語クラス図における概念間関係推定を ChatGPT (GPT-4) などの最新の LLM で行う。人間の正解率を超えることを目指し、また、これまで推定されていなかった関係の種類である多重度や属性も推定可能にする。システム概要を考慮した推定をするために、システム概要が記載された文章を学習する。さらに、様々なドメインに対応できる汎用性を獲得するために LLM の膨大な事前学習の知識を活用する。

推定の正解率向上に関しては Few-shot 学習を採用し、CoT (Chain of Thought) プロンプト戦略<sup>[3]</sup>と組み合わせ、限られた学習データでも高性能を発揮する手法を採用する。

### 3.1. 提案手法詳細

本手法では、既存研究の 2 つの問題を解決するためのアプローチを採用する。1 つ目はシステム概要が考慮されていないという問題に対し、システム概要が記載された文章を読み込むことで、システムに関する深い理解を可能にし解決する。2 つ目は様々なドメインへの汎用性が低いという問題に対し、Few-shot 学習を採用することで多様なドメインのデータを大量に用意する必要をなくすと共に、大量の知識を持っている LLM を活用することで多様なドメインに適応しやすくすることにより解決する。

これらの解決策を具体化するために、まずシステム概要が記載された要件定義書やシステムの説明文を読み込ませ、システムの構造と機能を理解する。このプロセスにより、システム概要を考慮したより正確な関係推定が可

能となる。次に、Few-shot 学習で数十のデータだけで学習すると共に、事前に大量のデータで訓練された LLM を使用することで様々なドメインにわたる複雑な関係の推定に対応できる。

推定の正解率向上に関しては Few-shot 学習と CoT プロンプト戦略を組み合わせ、推定の手順を示すと共に、例の中にも推定のプロセスを組み込むことで正解率の向上を図る。

また、既存研究ではシステムごとに推定が行われていなかったが、本手法ではシステム概要を考慮する観点からシステムごとに推定を行う。さらに、推定する関係の種類（関連、集約、継承）の拡張を行い、多重度と属性についても推定を可能にする。

本研究では関係線の有無の推定と関係線の種類（関連、集約、継承）推定、多重度の推定、属性の推定を行う。以下にそれぞれの推定について簡単にまとめる。

### 関係線の有無の推定と関係線の種類（関連、集約、継承）推定

関係線の有無の推定とは2つのクラスの間に関係線（関連、集約、継承）があるかを推定することであり、関係線の種類（関連、集約、継承）の推定とは関係線がある場合に、その関係線の種類（関連、集約、継承）が何かを推定することである。例えば車クラスとタイヤクラスの場合、「車クラスとタイヤクラスの間には関係線があり、その関係は集約である」となる。

### 多重度の推定

関係線があり、その関係が関連、集約である場合の多重度が何であるかを推定すること。例えば、例えば車クラスとタイヤクラスの場合、「車クラスとタイヤクラスの多重度は 1:\* である」となる。

### 属性の推定

属性がどのクラスに属しているかを推定すること。例えば、「創立年、車検日という属性があり、車クラスと大学クラスがあるとき、創立年は大学クラスに属し、車検日は車クラスに属する」となる。

また、上記3つの推定実験で使用した最終的なプロンプトは Few-shot 学習、CoT プロンプト戦略を軸として全体的な推定正解率向上を図った。プロンプトは実験ごとに試行錯誤を繰り返し、正解率が向上した取り組みを組み込むことで最終的な正解率を向上させた。それぞれの工夫点を以下に示す。

### 関係線の有無の推定と関係線の種類（関連、集約、継承）推定

関連と集約の間違いが多発するため、集約の特徴である部分全体の関係がどのようなものかを説明した文章をプロンプトに組み込む。

また、クラス名を2つずつのペアのリストとして渡し推定させると、関係線がないものを有りとして誤推定してしまう事が多発する。そのため、クラス名を一意的リストとして渡し、関係線のあるクラス名のペアだけを提示するように指示する。

### 多重度の推定

多重度は多くの場合 1, \*(多) が区別できれば十分であるため、1か\*(多)のみを推定するようにした。また、推定において多重度を逆に誤推定することが多発するため、(例えば、1:\*を\*:1 と推定してしまう) 推定の形式を指示するとともに、逆に推定しないようにプロンプトで指示した。

### 属性の推定

システムの説明に加えて、クラス間の関係の情報も与えることで、それぞれの属性がどのクラスに属すかの判断材料を増やした。また、所属するクラス名を勝手に変更することが頻発するため、クラス名の変更をしないように指示した。

評価方法としては、提案手法による推定結果を F 値 (F-score) と正解率 (Accuracy) で測定し、これを人間の正解率である 72%<sup>[4]</sup> と比較する。この比較により、提案手法の有効性を定量的に評価する。実験の正解率が人間の正解率を超えれば、自動化による大きな恩恵を示すことができる。

## 4. 実験

本章で具体的な実験について述べる。第 4.1 節では実験のために用意したデータセットについて説明する。第 4.2 節では、本手法の実装に使用する最適な LLM を選定するための比較実験について説明を行う。第 4.3 節ではプロンプトで使用する言語を決定するための実験を説明する。第 4.4 節では関係線の有無と関係線の種類（関連、集約、継承）の推定実験を説明する。第 4.5 節ではニッチな分野に対する実験を説明し、第 4.6 節で多重度、第 4.7 節で属性の推定実験について説明する。

#### 4.1. データセット

今回は日本語の UML の専門書([5], [6], [7], [8], [9], [10], [11])や応用情報技術者試験の過去問のクラス図を使用して作成した2つのデータセットを用意した。データセット1には 63 個のシステムと 800 個の関係が含まれている。データセット2には 31 個のシステムと 294 個の属性が含まれている。

#### 4.2. LLM 比較

LLM の比較はデータセット 1 の内、約半分の30個のシステムで行った。以下に実験の結果をまとめ、各 LLM と F 値を比較した図 1 と表 1 を示す。(実験日: 2023 年 12 月 6 日)

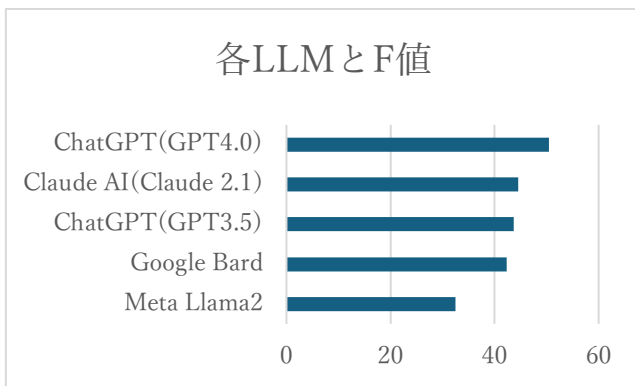


図 1 各 LLM と F 値

表 1 各 LLM と F 値

LLM	F値
Meta Llama2	32.5
Google Bard	42.3
ChatGPT(GPT3.5)	43.7
Claude AI(Claude 2.1)	44.5
ChatGPT(GPT4.0)	50.4

実験の結果、図 1、表 1 から F 値や正答率において ChatGPT (GPT4.0)が一番高かったため、今回の手法では ChatGPT(GPT4.0)を使用する。

#### 4.3. 言語(日本語と英語)による影響

日本語と英語のプロンプトによって F 値がどの程度変

化するかを調査した。本実験では関係線の種類(関連, 集約, 継承)の推定だけを行った。また, クラス名やシステムの説明の翻訳には ChatGPT(GPT4)を使用し, システム概要を考慮してクラス名の翻訳を行った。一部のプロンプトは手動で翻訳した。以下に実験結果をまとめた表 2 を示す。(実験日: 2023 年 12 月 6 日 ~ 21 日)

表 2 言語(日本語と英語)による影響の比較結果

言語 (日本語と英語) による影響				
言語	Precision	Recall	F-score	Accuracy
日本語	0.585	0.524	0.541	0.813
英語	0.560	0.539	0.547	0.819

OpenAI から発表された論文[12]では, GPT 4を英語で使用の方が日本語で使用するよりも 5.6%正解率が良くなるということがわかっているため, 今回の実験の結果は英語の方が高い精度となると予測していたが, 表 2 に示されている通り, 実際には大差がないという結果になった。

以上より, 日本語プロンプトと英語プロンプトを使用した際の正解率において劇的な差があるわけではないため, 次の実験では両方のプロンプトを使用して実験する。

#### 4.4. 関係線の有無と関係線の種類(関連, 集約, 継承)推定

本節では, 関係線の有無と関係線の種類(関連, 集約, 継承)推定実験について詳述する。(実験日: 2023 年 12 月 21 日)

##### 4.4.1. プロンプトの工夫点

工夫した点を以下に示す。

- CoT プロンプト戦略の使用  
注意 5 で推定の手順を指示することで, 正解率を向上させた。また, Few-shot 学習の各例にも思考の手順を記載した。
- Few-shot 学習  
それぞれの関係の定義と特徴, 例を学習させた。
- 反例を学習  
Few-shot 学習の時に, 例だけでなく反例も学習させることで正解率を向上させた。
- 関連と集約の誤推定を軽減  
関連と集約の区別が難しく, 誤推定が多発した。そのため, 集約の特徴である部分全体の

関係について詳細に記述して理解させることで誤推定を削減し、正解率を向上させた。

- クラス名をリスト形式で記述  
 クラス名をペアとして記述するのではなく、リストとして記述した。例えば自動車、トラック、軽自動車という3つのクラスがあるとき、ペアで記述すると

```
自動車 トラック
自動車 軽自動車
自動車 自動車
トラック 軽自動車
トラック トラック
軽自動車 軽自動車
```

となり、クラス名のリストで記述すると

```
自動車
トラック
軽自動車
```

となる。

ペアで記述すると関係のないクラスも関係ありと誤推定することが多々あったが、クラス名のリストで記述することで誤推定を減少させ、正解率を向上した。

- 無駄な関係の推定を削減  
 単語間に意味的な関係があるために、クラス図では関係がないにも関わらず関係があると誤推定されてしまうことが多発した。そのため、システム開発を意識したクラス図の作成をさせるために注意6で指示し、余計な関係の推定をしないように注意7で指示している。これによって関係なしをありと誤推定する確率を減らすことに成功し、正解率が向上した。

以上がプロンプトの説明となる。英語版のプロンプトは日本語のプロンプトをDeepLで英訳したものを使用した。

#### 4.4.2. 推定結果

日本語プロンプトと英語プロンプトを使用した推定をそれぞれ3回ずつ行い、その平均を算出した。日本語プロンプトを使用した方はF値の平均と正解率はそれぞれ79.8%、95.1%となり、英語プロンプトを使用した方はF値の平均と正解率はそれぞれ75.5%、93.4%となった。また、関係線の有無と関係線の種類(関連、集約、継承)推定の内、関係線の種類の推定に対するF値の平均はそれぞれ74.3%、69.1%となった。

以下にそれぞれの混同行列と実験結果を示す。図2, 3は一回目の推定の混同行列である。表3, 4は日本語プロンプト、英語プロンプトそれぞれの実験結果をまとめた表である。表5は表3, 4の平均をまとめたものであり、表6は関係線の種類(関連、集約、継承)に焦点を当てて算出したPrecision, Recall, F-score(F値), Accuracy(正解率)の平均をまとめた表である。

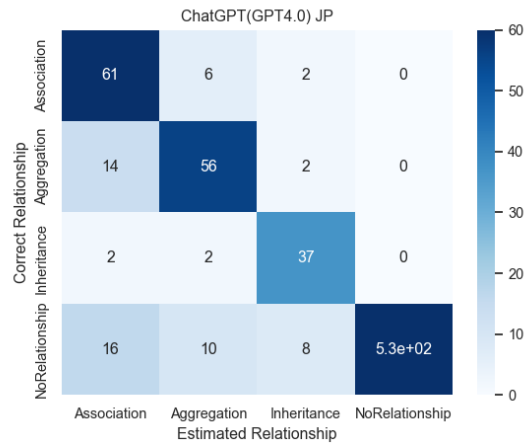


図2 日本語プロンプトを使用したChatGPT(GPT-4)による関係線の有無と関係線の種類(関連、集約、継承)推定(一回目)の混同行列

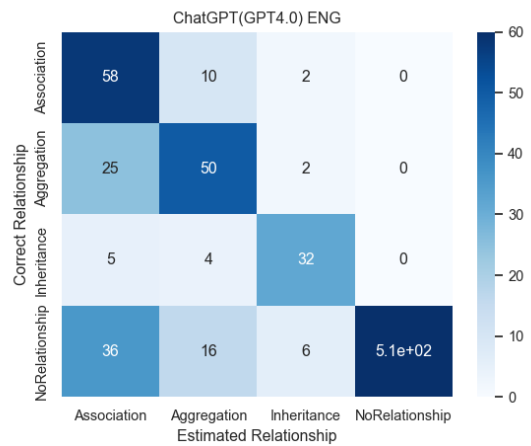


図3 英語プロンプトを使用したChatGPT(GPT-4)による関係線の有無と関係線の種類(関連、集約、継承)推定(一回目)の混同行列

表3 関係線の有無と関係線の種類(関連、集約、

## 継承)推定(日本語プロンプト)

関係線の有無と関係線の種類推定(日本語プロンプト)				
	Precision	Recall	F-score	Accuracy
一回目	0.792	0.876	0.828	0.958
二回目	0.738	0.832	0.778	0.947
三回目	0.753	0.839	0.788	0.946
平均	0.761	0.849	0.798	0.951
分散	0.000513	0.000372	0.000463	0.0000311
不偏分散	0.000769	0.000558	0.000694	0.0000467

表 4 関係線の有無と関係線の種類(関連, 集約, 継承)推定(英語プロンプト)

関係線の有無と関係線の種類推定(英語プロンプト)				
	Precision	Recall	F-score	Accuracy
一回目	0.714	0.789	0.738	0.930
二回目	0.715	0.791	0.742	0.930
三回目	0.745	0.849	0.786	0.942
平均	0.724	0.810	0.755	0.934
分散	0.000206	0.000769	0.000465	0.0000339
不偏分散	0.000309	0.001153	0.000698	0.0000508

表 5 関係線の有無と関係線の種類(関連, 集約, 継承)推定の結果

関係線の有無と関係線の種類(関連・集約・継承)推定				
言語	Precision	Recall	F-score	Accuracy
日本語	0.761	0.849	0.798	0.951
英語	0.724	0.810	0.755	0.934

表 6 関係線の有無と関係線の種類(関連, 集約, 継承)推定のうち種類推定の結果だけを抽出

種類推定の結果だけを抽出				
言語	Precision	Recall	F-score	Accuracy
日本語	0.681	0.822	0.743	0.952
英語	0.632	0.780	0.691	0.937

図 2, 図 3 より, 英語プロンプトの方が日本語プロンプトと比べて関係線の有無の誤推定の数が多いことがわかる。また, 関連, 集約, 継承の全てにおいて正解の推定数が日本語プロンプトの方が多く結果となった。また, 関係線の種類(関連, 集約, 継承)推定に焦点を当てた表 6 より, 関係線の種類(関連, 集約, 継承)も高い正解率

で分類できていることがわかる。日本語プロンプトを使用した方の F 値と正解率が良くなった原因は主に 2 つ考えられる。1 つ目は翻訳した文章が LLM にとって理解しづらい文章になっていた可能性である。今回は日本語プロンプトを作成してから DeepL を使用して翻訳し英語プロンプトを作成したため, 翻訳の些細な誤差が LLM の理解を低下させた可能性がある。2 つ目は GPT4 の各言語のトレーニングデータへの対応力である。GPT-4 は多言語でトレーニングされているが, 使用される言語によってトレーニングデータの量や質が異なる。今回使用したデータセットは全て日本語で記述されたクラス図とそのシステムの説明から作成しているため, 日本語で記述されたクラス図に対する対応力が日本語の GPT4 の方が高い可能性がある。

分散について表 3, 4 より, F 値の分散の値は日本語プロンプト, 英語プロンプトのそれぞれが低く, 各回の F 値が F 値の平均から約±3%の範囲に収まっており, 正解率についても約±1%の範囲に収まっている。

まとめると表 5 より, 日本語プロンプトを使用した方の正解率が高く, 実験の正解率は人間の正解率である 72% [5]を大きく超える 95.1%となった。

## 4.4.3. 推定内容

実験の推定内容について詳しく説明し, 考察する。以下に応用情報技術者試験で出題されたショッピングシステムの推定例を示す。使用したプロンプトは上記で説明したプロンプトに以下を加えたものである。

## [クラス名のリスト]

ショッピングカート  
 予約販売商品  
 通常商品在庫管理  
 商品  
 商品在庫管理  
 予約販売商品在庫管理  
 通常商品

## [システムの説明]

P 社では, 新たな事業展開として, インターネットを用いた通信販売を開始することにした。通信販売のための販売用 Web サイトは, 新規に開発する。販売用 Web サイト及び販売用 Web サイト内で用いるショッピングカートに関する説明を次に示す。

## [販売用 Web サイト]

インターネットに公開し一般の顧客が買物に利用す

る。顧客は、P 社から付与される顧客 ID でログインしてから買物をする。顧客は、商品カタログを画面に表示し、ショッピングカートに商品を追加したり、ショッピングカートから商品を削除したりして、購入する商品を選ぶ。顧客は、商品を選び終わったら、ショッピングカート内の商品の購入手続を行う。商品には、通常商品と予約販売商品の 2 種類がある。通常商品を購入した場合の配送手続では、即座に商品の配送処理が行われる。予約販売商品を購入した場合の配送手続では、配送のための情報がデータベースに保存され、実際の配送処理は商品の発売開始日以降に行われる。商品の配送処理は、既存の配送処理システムと連携することによって行う。販売用 Web サイトは、購入された商品の情報を配送処理システムに通知する。配送処理システムは、通知された商品の情報をとりまとめて、配送業者に集配依頼の情報を送る。

[ショッピングカート]

顧客がショッピングカートに商品を追加すると、追加された商品の在庫数を、追加された数量分だけ減らす。ただし、商品の在庫数が不足している場合は、ショッピングカートに商品を追加せず、在庫数も減らさない。顧客がショッピングカートから商品を削除すると、削除された商品の在庫数を、削除された数量分だけ増やす。販売用 Web サイトの開発を行うに当たり、データベース及びショッピングカートの設計を次のように行った。

[データベースの設計]

販売用 Web サイトで使用するデータベースには、商品在庫情報テーブル、ショッピングカート情報テーブル及び販売明細テーブルを用意する。商品在庫情報テーブルには、商品名や単価などの商品に関する情報と、その在庫数を格納する。商品は、商品 ID で一意に識別する。ショッピングカート情報テーブルには、ショッピングカートに入っている商品の商品 ID と数量を格納する。ショッピングカートは、顧客 ID で一意に識別する。販売明細テーブルには、顧客が購入した商品の情報を格納する。販売明細は、注文 ID と商品 ID の複合キーで一意に識別する。注文 ID は、購入手続を行ったときに発行される ID である。なお、販売用 Web サイトに用いるデータベースでは、トランザクション内でテーブルに対する更新アクセスが発生するとテーブル単位のロックがかかり、トランザクション終了時に、すべてのロックが解除される仕組みになっている。

推定結果は以下のようになった。

このケースでは、与えられたクラス名間の関係を UML クラス図の文脈で推定します。以下のクラス名があります：

1. ショッピングカート 2. 予約販売商品 3. 通常商品在庫管理 4. 商品 5. 商品在庫管理 6. 予約販売商品在庫管理 7. 通常商品

これらのクラス名とシステムの説明を元に、推定される関係は次のとおりです：

1. **\*\*ショッピングカートと商品\*\***: これは関連の関係です。ショッピングカートは、商品を含んでいるので、ショッピングカートと商品の間には明確な関連があります。

2. **\*\*商品と商品在庫管理\*\***: これは集約の関係です。商品在庫管理は、商品の在庫情報を管理しますが、商品が存在しなくても商品在庫管理は機能します。

3. **\*\*予約販売商品と予約販売商品在庫管理\*\***: これも集約の関係です。予約販売商品在庫管理は予約販売商品の在庫を管理するが、予約販売商品が存在しなくても、予約販売商品在庫管理は存在し続けることができます。

4. **\*\*通常商品と通常商品在庫管理\*\***: これも集約の関係です。通常商品在庫管理は通常商品の在庫を管理しますが、通常商品が存在しなくても、通常商品在庫管理は存在し続けることができます。

5. **\*\*商品、予約販売商品、通常商品\*\***: これは継承の関係です。予約販売商品と通常商品は、商品の特殊な形態です。したがって、予約販売商品と通常商品は商品クラスから属性やメソッドを継承します。

これらの関係を JSON 形式で表すと以下のようになります：

```
``json
[
  ["ショッピングカート", "商品", "関連"],
  ["商品", "商品在庫管理", "集約"],
  ["予約販売商品", "予約販売商品在庫管理", "集約"],
  ["通常商品", "通常商品在庫管理", "集約"],
```

```
["商品", "予約販売商品", "継承"],
["商品", "通常商品", "継承"]
]
...
```

答え

```
ショッピングカート 商品在庫管理 関連
予約販売商品 商品 継承
通常商品 商品 継承
商品在庫管理 予約販売商品在庫管理 継承
商品在庫管理 通常商品在庫管理 継承
```

推定内容について考察すると、商品と予約販売商品、通常商品についてはシステムの説明の部分で「商品には、通常商品と予約販売商品の2種類がある。」という詳細な説明があるため、システムの概要とシステムの説明を理解してうまく推定することができている。このようにシステムの概要を考慮した推定が可能となっている。一方で、今回誤推定した["通常商品", "通常商品在庫管理", "集約"]について、通常商品在庫管理が通常商品を束ねるというのはシステム的には正解だが、今回のクラス図的には正しくなかった。このようにシステム概要を考慮できても必ず正しいクラス図を作れるわけではないことがわかる。

#### 4.5. ニッチなクラス図の推定

銀行システムやショッピングサイトのような前例やデータが豊富ではない「ニッチ」な分野に対する推定能力を調査することで、LLM のクラス図における概念間関係推定に対する推定の汎用性を確かめる。しかし、日本語で記述されたニッチな分野のクラス図を見つけることが難しかったため、英語のクラス図を翻訳して使用した。実験の結果を表7に示す。(実験日: 2024年1月1日)

表 7 ニッチなクラス図の推定結果

システム名	正答率
農業用地管理システム	0.900
患者の病歴管理システム	0.873
スマートオートドア制御システム	0.956

実験の結果、表7に示されている農業系のようなニッチな分野におけるクラス図を高い正答率で推定できるこ

とが明らかになった。特に、関係線の有無の推定についてはどのケースでも正答率が高かった。ただし、医療系クラス図における関係線の種類推定では一部成功しないケースもあり、この点において改善の余地がある。しかし、全体的には関係線の有無と関係線の種類推定を合わせてみると、高い水準の推定精度が達成されている。この結果から、ニッチな分野に対応できる能力があることが確認されたが、さらなる精度向上を目指すとともに、より多様な分野についても調査をする必要がある。

#### 4.6. 多重度の推定

多重度の推定において、多くの場合には多重度が1または\*(多)であるかが分かれば十分であることと、すべての多重度(例えば、0.1や0..\*, 1.3など)を推定することは難しいと判断したため、今回は1と\*(多)の多重度に絞って推定を行った。また、プロンプトについては関係線の有無と関係線の種類(関連, 集約, 継承)推定で最も高いF値と正解率を示した日本語プロンプトを使用する。以下に実験の詳細を記述する。(実験日: 2023年12月26日)

##### 4.6.1. プロンプトの工夫点

多重度の推定において使用したプロンプトの工夫点を以下に示す。

- Few-shot 学習  
多重度の例を与えることで、推定の思考プロセスを学習させた。
- CoT プロンプト戦略の使用  
片方ずつ考えさせることや推定の仕方を教えることで、思考の順番を示し、正解率を向上させた。
- 多重度の推定が逆にならないように指示  
多重度の推定を逆に行ってしまうことが多々あった。例えば、["車", "1", "タイヤ", "多"]を["車", "多", "タイヤ", "1"]と推定してしまうことが多発した。そのため、注意3, 4, 5で指示することで多重度を逆に考えてしまう回数を減少させた。しかし、依然として逆に推定してしまうことが発生したため、推定の形式通りに推定するように指示した。これによって多重度を逆に推定することが激減し、正解率

が大幅に向上した。

#### 4.6.2. 推定結果

日本語プロンプトを使用した実験を 3 回行った。実験の結果は F 値と正解率の平均が 69.0%, 87.2%となり、混同行列は以下の図 4 のようになり、推定結果は表 8 のようになった。

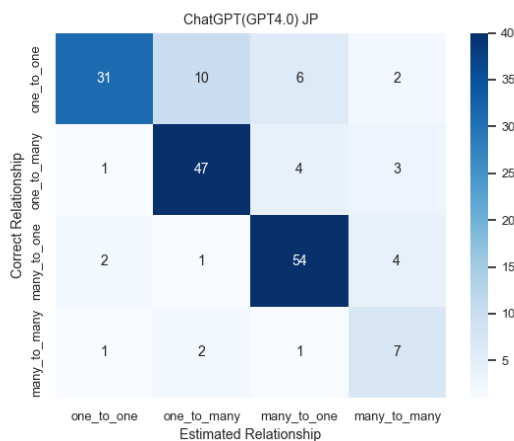


図 4 多重度推定(一回目)の混同行列

表 8 多重度推定結果

多重度推定				
	Precision	Recall	F-score	Accuracy
一回目	0.734	0.752	0.733	0.895
二回目	0.672	0.632	0.639	0.843
三回目	0.711	0.716	0.698	0.878
平均	0.706	0.700	0.690	0.872
分散	0.000653	0.002523	0.001504	0.000466
不偏分散	0.000980	0.003784	0.002257	0.000699

図 4 に示された多重度推定の混同行列から、多対多の推定が他のカテゴリに比べて若干低い正解率であることがわかる。しかし、他の多重度の推定については良く推定できていることがわかる。

表 8 より、F 値の分散については他の推定よりも大きくなっており、各回の F 値が F 値の平均から約 ±5% の範囲にあることから、結果に散らばりがあることがわかる。よって回答に一貫性がないことに留意しなければならない。

全体として表 8 より、実験結果の正解率が 87.2%であり、

これは人間の正解率である 72%を超える結果になったことから、関係(多重度)を拡張できたと言える。

#### 4.7. 属性の推定

属性の推定実験について紹介していく。また、プロンプトについては関係線の有無と関係線の種類(関連, 集約, 継承)推定で最も高い F 値と正解率を示した日本語プロンプトを使用する。(実験日: 2023 年 12 月 24 日)

##### 4.7.1. プロンプトの工夫点

属性の推定において使用したプロンプトの工夫点を以下に示す。

- Few-shot 学習  
クラスと属性の例を与えることで、推定の思考プロセスを学習させた。
- システムの説明を参照するように注意  
単純なプロンプトを使用した際に、候補が複数提示されることがあった。部分点として評価しても良かったが、システムの説明を参照させることで 1 つに絞り正解率を向上させた。
- クラス名を変更しないように注意  
単純なプロンプトを使用した際に、クラス名を英語にしたり、別の単語を使用したりとクラス名が変更されてしまうことがあったため、指示を追加し、正解率を向上させた。これにより、クラス名の一貫性を保つことで、より正確な属性の割り当てが可能になった。

##### 4.7.2. 推定結果

日本語プロンプトを使用した実験を 3 回行った。結果は F 値と正解率の平均が 73.6%, 78.0%となり、混同行列は図 5 のようになり、推定結果は表 9 のようになった。



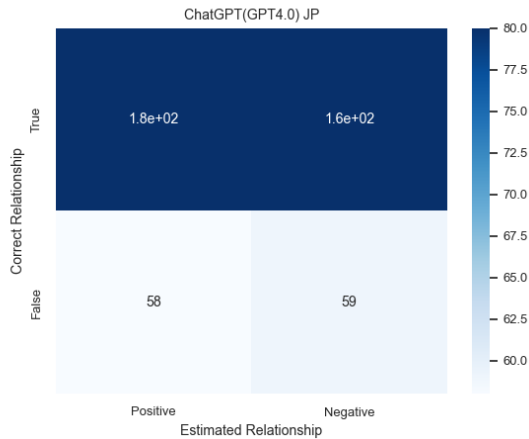


図 5 属性推定(一回目)の混同行列

表 9 属性推定結果

属性推定				
	Precision	Recall	F-score	Accuracy
一回目	0.823	0.775	0.792	0.785
二回目	0.806	0.775	0.779	0.769
三回目	0.823	0.775	0.792	0.785
平均	0.817	0.775	0.788	0.780
分散	0.0000651	0.0000000	0.0000389	0.0000565
不偏分散	0.0000977	0.0000000	0.0000584	0.0000847

表 8 から実験の正解率の平均が 78.0%であり、これは人間の正解率である 72%を上回っていることから、属性推定が比較的正確であることが示されている。また、F 値の分散も小さく、各回の F 値が平均の±2%の範囲に収まっている事がわかる。正解率の分布も同様に小さく約±1%の範囲に収まった。しかし、関係線の有無と関係線の種類の推定、多重度の推定の 2 つと比べると正解率が低いため、改善の余地があることも確かである。

全体として、実験の正解率が人間の正解率を超えることができたため、関係(属性)の拡張に成功したと言える。

## 5. まとめ

すべての実験結果を表 10 にまとめる。

表 10 実験全体の結果まとめ

推定種類	F-scoreの平均	Accuracyの平均
関係線の有無と関係線の種類(関連, 集約, 継承)推定	0.798	0.951
多重度推定	0.690	0.872
属性推定	0.788	0.780

実験の結果から主に3つのことが明らかになった。1つ目は関係線の有無と関係線の種類(関連, 集約, 継承)推定において、人間の正解率と既存研究の正解率を大きく超えることができたこと(正解率平均: 95.1%)であり、2つ目は既存研究の課題であった「システム概要が考慮できていない点」と「汎用性が低い点」を解決することができたことであり、3つ目は多重度と属性の推定を可能にしたことで推定できる関係の種類に成功したこと(それぞれの正解率平均: 87.2%, 78.0%)である。

以上の実験結果から、本手法は有効であることがわかった。

また、主に3つのことが今後の改善のための課題として挙げられる。1つ目はデータセットに不足があること。2つ目は汎用性が広がったが完璧ではないこと。3つ目はシステム概要を考慮する事が必ずしも正しいクラス図の概念間関係を推定できるわけではないことである。

今後の展望としては、課題を改善するとともに本研究をさらに拡張し、要件定義書からクラス図作成までの一連の流れの自動化を達成することである。これによって、作業工数の大幅な削減という目標を達成し、システム製作の効率化に貢献できると考える。

## 参考文献

- [1] 独立行政法人情報処理推進機構, “DX 白書 2023,” p.13, 2023.
- [2] 久住憲嗣, 山内陽太: 機械学習を用いた自然言語文書からのドメインモデル推定手法, 組込みシステム技術に関するサマワーショップ, 2023.
- [3] OpenAI, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” arXiv, 2023.
- [4] 井原輝人, 内田真司, 飯田元, “自然言語処理的アプローチによるクラス図関連線の予測,” ソフトウェア・シンポジウム 2021 in 大分, 2021.
- [5] 竹政昭利, 徹底攻略 UML モデリング技能認定試験問題集 L1(T 1/T2)対応, 株式会社ソキウス・ジャパン, 2007.
- [6] 竹政昭利, 合格 Expert UML モデリング技能認定試験入門レベル(L 1)対応問題集, 株式会社技術評論社, 2004.

- [7] 株式会社テクノロジーアート, UML モデリング教科書 UML モデリング L 1 第二版, 株式会社翔泳社, 2008.
- [8] 児玉公信, UML モデリング入門, 日経 BP 社, 2008.
- [9] 荒井玲子, オブジェクト指向モデリングセルフレビューノート, 株式会社ディー・アート, 2005.
- [10] カサレアル/川添真智子, UML モデリング技能認定試験スキルアップ教本[レベル1], 株式会社秀和システム, 2005.
- [11] クレーグ・ラーマン, 実践 UML 第二版 パターンによる統一プロセスガイド, 株式会社ピアソン・エデュケーション, 2003.
- [12] OpenAI, “GPT-4 Technical Report,” arXiv, p.8, 2023.