
一般セッション | 一般セッション | 感性に係る計測・評価技術に関する分野

[3E01] 感性に係る計測・評価技術に関する分野 General Session 7

2023年11月22日(水) 09:40 ~ 12:00 E会場 (4階 研修室)

[3E01-03] 大規模言語モデルを用いた評価グリッド法の自動化 An Automated Evaluation Grid Method Using Large-Scale Language Models

*張 帆¹、宮嶋 大輔¹、杉本 匡史¹、佐々木 香暖¹、北野 泰成¹、橋本 翔²、長田 典子¹ (1. 関西学院大学、2. 西南学院大学)

*Fan Zhang¹, Daisuke Miyajima¹, Masashi Sugimoto¹, Kanon Sasaki¹, Taisei Kitano¹, Sho Hashimoto²,
Noriko Nagata¹ (1. Kwansei Gakuin University, 2. Seinan Gakuin University)

大規模言語モデルを用いた評価グリッド法の自動化

— マルチエージェントによるインタビュー対話システム —

張 帆*, 宮嶋 大輔*, 杉本 匡史*, 佐々木 香暖*, 北野 泰成*, 橋本 翔**, 長田 典子*

* 関西学院大学理工学部/感性価値創造インスティテュート, ** 西南学院大学商学部

An Automated Evaluation Grid Method Using Large-Scale Language Models

— A Multi-agent Interview Dialogue System —

Fan Zhang*, Daisuke Miyajima*, Masashi Sugimoto*, Kanon Sasaki*, Taisei Kitano*, Sho Hashimoto**, Noriko Nagata*

E-mail: *{ zhangfan, daisukemiyajima, sugimoto.masashi, k-sasaki, kitanott99, nagata }@kwansei.ac.jp

Abstract: This study aims to improve and streamline the individual interview method involving the Evaluation Grid Method, which is traditionally laborious and time-consuming. For this, it leverages a large-scale language model to build an interview dialogue system. In particular, the study uses GPT4 to implement an interview dialogue with structured causal relationships (laddering). The process from interview dialogue to the analysis of the evaluation structure is also automated by employing BERT and the Evaluation Structure Visualization System (ESV). The results derived from the system are then compared and discussed against those obtained manually.

Keywords: Large-scale language model, GPT4, Evaluation grid method, Generative model, Multi-agent

1. はじめに

ユーザーニーズの多様化に伴い、一人一人の価値観や嗜好を的確に把握し、具体的なデザインに展開する方法論が求められている。ユーザーニーズの発掘・抽出・把握の方法は、プロダクトやサービスにおける個別最適化や感性価値（経験価値、体験価値）向上に必須なコア技術であり、心理学的手法、マーケティングリサーチ手法、行動観察手法、感性評価手法、機械学習手法など、多様な分野での研究が進められている。

ユーザーニーズを把握する手法については、従来、インタビュー調査や主観評価実験による心理学的データの収集と分析によって行われてきた。代表的なインタビュー手法として評価グリッド法[1]がある。評価グリッド法では、抽象的な価値判断・心理的な価値から客観的な判断・物理的な状態までを階層的に整理し評価構造として可視化する。ラダリングと呼ばれる上位概念・下位概念を聞き出す対話手順を行うことで、インタビューから効率よく評価構造を構築することが可能である。この手法はその有用性から多くの研究開発に用いられているものの、人的・時間的な負荷が高いという問題がある。そのため、実施を支援するツールの開発などが試みられているが、更なる効率化が期待されている[2, 3]。

一方で、自然言語処理分野では Transformer ベースの大規模言語モデル(LLM)の発展によって、対話システムの可能性が飛躍的に広がっている[4]。

そこで本研究では、大規模言語モデルを用いて評価グリッド法におけるインタビューを自動化する方法を提案する。具体的には、大規模言語モデルである GPT4 モデルをインタビューフレームワークに適用し、評価グリッド法のための因果関係（ラダリング）のインタビュー対話システムを構築する。

また BERT と評価構造可視化システム (ESV) を用いてカテゴリ化と評価構造図の作成を行うことで、インタビュー対話から評価構造分析までを自動化する。提案手法による評価構造の自動分析結果を、手作業の評価グリッド法で得られる評価構造と比較することで有用性の検証を行う。

2. 関連研究

評価グリッド法[1]は、インタビューにより個人の持つ評価構造を階層構造として表現する手法である。元はパーソナル・コンストラクト理論と呼ばれる「人間は経験によって個々に構築された認知構造があり、認知構造に照らし合わせて自らの行動を決定し、修正する」という理論を背景に、実務向けに改良された個別面接手法である。

評価グリッド法のインタビューで行われるラダリングの手順を図1に示す。まず比較対象となる複数の刺激を用意し、参加者にそれらを嗜好度順で並べてもらう。インタビュアーは、参加者になぜそのように判断したか理由を尋ねる。そこで得られた評価項目に対し、抽象的な価値判断や心理的な価値といった上位概念を引き出すラダーアップと、具体的かつ客観的な判断や物理的な状態といった下位概念を引き出すラダーダウンの質問を繰り返し、さらに評価項目を引き出していく。最後に、得られた評価項目を階層構造として整理したものが評価構造（図1）である。

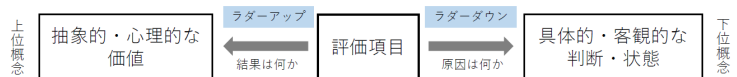


図1. 評価グリッド法におけるラダリング

3. 評価グリッドインタビュー対話システム

本研究では、評価グリッド法自動化システムを提案する。システムは図 2 のようにインタビュー対話①、評価項目クラスタリング②、代表表現抽出③、評価構造可視化④の 4 つのサブシステムから構成される。

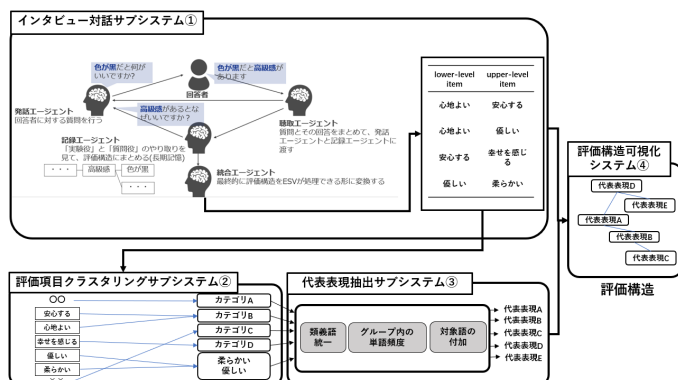


図 2. 評価グリッド法自動化システムの全体図

3.1 インタビュー対話サブシステム

インタビュー対話サブシステムのコアとなるマルチエージェントモデルを図 3 に示す。マルチエージェントシステムは複数の自律的なエージェントが協調して問題を解決するための計算モデルである。それぞれのエージェントが異なる役割やタスクを果たすことで、単一エージェントが処理する情報が制限され、LLM の一貫性欠如、いわゆるプロンプト注入やハルシネーションの問題が緩和され、問題解決が迅速に進められる。本研究の場合、単一エージェントでは参加者の発話に影響され、インタビューを早期に終了したりインタビューの任務から離脱したりする可能性が懸念される。またインタビューが長く続くことによってユーザにとって予想外の発話をする確率が増える。これらの問題がマルチエージェントの導入によって緩和できると期待される。さらに、予期しない問題が発生した際に、エージェント間の情報の引き渡し記録は問題を特定しやすい。エージェントを増やしたり修正したりすることで、システムの規模を拡大したりタスクを調整したりすることも可能である。

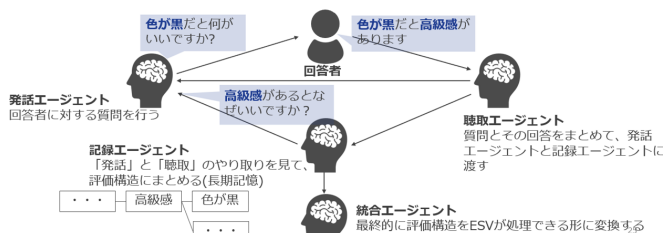


図 3. インタビュー対話のマルチエージェントモデル

今回は質問を生成する発話エージェント、評価項目を抽出する聴取エージェント、会話を記録する記録エージェント、

会話をまとめ評価項目リストを出力する統合エージェントの 4 エージェントで構成する。LLM は GPT-4 を使用する。パラメータは Top-p = 0, Temperature = 0 に設定する。

3.1.1 プロンプトチューニング

評価グリッド法を適切に行えるよう、プロンプトチューニングを行う。Wei ら[6]はあらかじめ関連する Question と Answer を例示し、推論の道筋を方向づける Chain of Thought を提案している。本研究では、インタビューを自動化するため、インタビュー例を各エージェントのプロンプト内に記述することで、役割の方向づけを行った。

3.1.2 発話エージェントのプロンプト

発話エージェントは実験全体のプロセスを把握し質問を生成する。テーマ、聴取エージェントがまとめたすべての評価項目、記録エージェントが記録したラダリング結果（評価項目の因果ペア）を参考にして次の質問を決定する。各エージェントの司令塔とも言える。プロンプトは以下の 6 つの要素で構成する。各要素は「<役割説明></役割説明>」のように XML タグで定義する。

- ① 役割説明
- ② 制約
- ③ 評価グリッド法の説明
- ④ インタビュー例
- ⑤ 出力例
- ⑥ インタビュー手順

①では図 4 のように発話エージェントとしての役割を明示的に指定している。山崎ら[7]の研究では、属性を与えることで属性に合った発話を生成することを可能とした。

あなたはプロの評価グリッドインタビューャーです。私はインタビューの参加者です。以下の説明に沿って参加者に対してインタビューを行ってください
—省略—

図 4. 評価グリッド法の説明のプロンプト

②では図 5 のように制約を記述した。無関係な発話や挨拶を避けることができる。

最初に話す言葉は「これからインタビューを始めます。あなたが<テーマ>だと思うものを一つ答えてください。」です。これ以外は話さないでください
・インタビュー中での評価グリッド法についての説明、参加者の回答に関する解釈についての説明は不要です。参加者の発言に対する評価や感想も不要です。下記の評価グリッド法の説明とインタビュー手順に沿ってインタビューを行ってください。
・今までの価値構造を参考してインタビューを進めてください。
—省略—

図 5. 制約のプロンプト

③では図 6 のように、評価グリッド法についての説明をおこなった。これは今から何を行うのかを抽象的に理解するために記述している。

評価グリッド法は、半構造化インタビュー用いた定性評価方法です。評価グリッド法は、デプスインタビューとしての側面を持っており、回答者が自覚していない深層心理を引き出すことができます。

—省略—

図 6. 評価グリッド法の説明のプロンプト

④では図 7 のように、インタビュー例を記述している。Question と Answer を例示し、推論の道筋を方向づけすることができる。

オリジナル評価項目：車の高級感
S (system), U (user), Upper-level item (上位項目), Lower-level item (下位項目)
S: あなたはどのような時に、車に高級感を感じますか?
U: スポーティ感がある時
S: 具体的にどのような場合に、スポーティ感があると感じますか?
U: 形が流線形
S: では具体的にどんな感じのものが流線形だと感じますか?
U: 先っぽが細く、車高が低い時
S: 他にはどういう場合に、スポーティ感があると感じますか?

—省略—

図 7. インタビュー例のプロンプト

⑤では、出力例を記述している。インタビュー終了時に、参加者から得られた評価項目の因果ペアを出力する必要がある。今回は図 8 のように CSV 形式でリスト出力するようにした。

lower-level item, upper-level item, interviewee
高級感がある, スポーティ, ID1
高級感がある, シートが本革, ID1
スポーティ, 色が赤い, ID1
スポーティ, 形が流線形, ID1
—省略—
壊れにくい, クオリティが良い, ID1
面倒くさくない, 壊れにくい, ID1

図 8. 出力例のプロンプト

⑥では、インタビュー手順を記述した。評価グリッド法におけるインタビューは、ラダリングを行うことで、参加者の評価の因果関係を明らかにする。図 1 で説明したラダリングの手順に基づき、図 9 のようにインタビュー手順のプロンプトを記述した。

評価グリッド法のインタビューの手順は、オリジナル評価項目の抽出とラダリングの繰り返しです。

オリジナル評価項目とは、インタビューの起点となる評価項目です。

ラダリングは評価項目からより抽象的な評価項目（上位概念）より具体的な評価項目（下位概念）を引き出すための手順です。

ラダリングでは、評価項目からより抽象的な上位概念を引き出すラダーアップと、より具体的な下位概念を引き出すラダーダウンを行います。具体的な手順は：

1. オリジナル評価項目に対してラダーダウンを行い、これ以上下位概念が引き出せなくなるまで繰り返す
2. オリジナル評価項目に対してラダーアップを行い、これ以上全て下位概念が引き出せなくなるまで繰り返す
3. 次のオリジナル評価項目に移行し、1と2を実行
4. すべてのオリジナル評価項目が終了したら、インタビューを終了します。

—省略—

図 9. インタビュー手順のプロンプト

3.1.3 聴取エージェントのプロンプト

聴取エージェントはインタビュー参加者に質問し、答えを評価項目にまとめて発話エージェントと記録エージェントに渡す。参加者の発話をフィルターリングし、評価項目だけ発話エージェントに渡し、参加者の発話に影響させない役割が担っている。プロンプトは図 10 のように構成した。

① 役割説明

② 制約

③ 会話例

<役割説明>

あなたはプロの評価グリッドインタビュワーです。私はインタビューの参加者です。以下の説明に沿って参加者に対してインタビューを行って、答えをまとめてください。

</役割説明>

<制約>

- ・ルール以外は話さないでください。会話例の中の参加者会話を出力しないでください。
- ・私は質問に対して答えます。その時、以下のルールに従って答えをまとめて名詞だけを出力してください。

—省略—

- ・答えに逆の概念が存在する場合、参加者が言いたい概念を推測して出力してください。

</制約>

<会話例>

<例 1>

S: 車に高級感を感じるためには何がどうなっていることが必要ですか?

U: スポーティ感がある時

S: スポーティ感がある

</例 1>

—省略—

</会話例>

図 10. 聴取エージェントのプロンプト構造

また、出力を安定させるため、聴取エージェントは前回の会話記録は持たない。

3.1.4. 記録エージェントのプロンプト

記録エージェントは発話エージェント・聴取エージェントの内容を記録する。記録エージェントは発話エージェントの「記憶」の役割を担っている。プロンプトは以下のように構成した。

- ① 役割説明
- ② 評価グリッド法の説明
- ③ 出力例

3.1.5. 統合エージェントのプロンプト

統合エージェントは全ての会話記録と記録エージェントのまとめ結果を参考にして、最終的に得られた評価項目の因果ペアリストを出力する。プロンプトは前項と同様に以下のよう構成した。

- ① 役割説明
- ② 評価グリッド法の説明
- ③ 評価項目の因果ペアリストの例

3.1.6. プロンプトの切り替え

以上の発話・聴取エージェントのプロンプトは、ポジティブなテーマ（〇〇の良いところ）だけに対応できる。ネガティブなテーマでは「〇〇だと何故悪いのですか」などの対応方法が必要だが、プロンプトに両方の対応方法を入れるとプロンプトが膨大しすぎて、コストが増えるだけでなく、出力も不安定になる。そのため、本研究では発話と聴取エージェントのプロンプトをポジティブとネガティブの2セットを用意する。

3.2. 評価項目クラスタリングサブシステム

収集した評価項目の表記ゆれの統一を行った上で、類似項目のクラスタリングを行う。

3.2.1. 前処理

収集した評価項目は最初に前処理を行う。前処理は、形態素解析で単語単位に分割、必要な単語の抽出、単語の活用形を原型に統一の順で行う。形態素解析は MeCab の Neologd 辞書を用いる。評価項目の分類に必要な単語の抽出のため、助詞などの意味を持たない単語を除外する。方法としては、名詞・動詞・形容詞・副詞のみを抽出し、その中で、よく使われるが関連性のない単語である「良い」「思う」「する」を削除する。

3.2.2. ベクトル化

前処理を行った文に対し、Tf-Idf を用いてベクトル化を行う。Tf-Idf は単語の出現回数により文章をベクトル化する手法の1つで、ある単語のある文書における出現頻度(tf)と、ある単語がいくつの文書で使用されているかを表す逆文書頻度(idf)を乗じたものである。クラスタリングにおいて文を

特徴づける単語は重要であり、idf は特定の文にしか登場しない単語であればあるほど値が高くなるため、Tf-Idf を用いてベクトル化を行う。

3.2.3. クラスタリング

ベクトル化した評価項目に対し、スペクトラルクラスタリングを用いて分類する。スペクトラルクラスタリングはグラフのノードをクラスタに分類するアルゴリズムである。ベクトル化した文から類似度グラフ・類似度行列を作成し、固有値問題を解くことで分類する。ここでは類似度グラフ作成の際の計算方法に最近傍法を、固有ベクトルのクラスタリングに K-means を使用する。

3.3. 代表表現抽出サブシステム

クラスタごとに評価項目を統一する代表的な言葉(代表表現)を抽出する。評価構造は人が解釈しやすいようできる限り簡潔化する必要があり、類似した評価項目の代表化処理が不可欠である。このため、クラスタ内の単語頻度や概念抽象化・係り受け解析に基づいて代表表現を選出する。

はじめに、3.2.1 と同様にクラスタ内の評価項目に前処理を施し、単語に分割する。

次に、我妻ら[8]の手法に倣い、類義語判定の閾値を自動的に決定することで単語のクラスタの中から類義語を統一する。類義語判定の閾値決定に際して、意味の類似した単語を概念単位でまとめ概念間の上下関係を記述した EDR 概念体系辞書と、単語や文章を分散表現として表すことができる自然言語処理モデル(BERT, word2vec)を活用する。

その後、類義語統一を施した単語のクラスタから重要度が最も高い単語を決定する。重要度の算出には Tf-Idf を用いる。Tf-Idf により求めた単語がクラスタ全体に占める割合を算出し、それが閾値を下回っていた場合、EDR 概念体系辞書を用いてクラスタ内の単語の概念を一段階上のものに置き換える。再度クラスタ内の単語の重要度を算出し、値が最大のものを代表表現とする。

また、概念の抽象化が施されなかったクラスタについては、代表表現と係り受けの関係にある単語をクラスタ内の評価項目から探索し、代表表現に付加する。係り受け関係の抽出には形態素解析器 Ginza を活用する。

3.4. 評価構造可視化サブシステム

評価構造可視化サブシステム(ESV)は、評価構造図を構築することができる。代表表現抽出サブシステムが出力した CSV 形式のデータから評価構造図を作成した。

4. 実験

4.1. 実験概要

提案手法の有効性検証のため、手作業によるインタビューで構築した評価構造と、本対話システムによるものを比較する。比較対象は杉本ら[9]が行った三宮(神戸市の繁華街)の

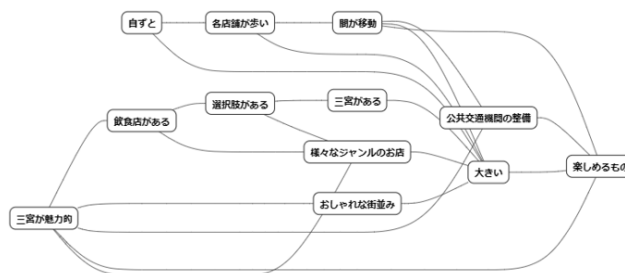
実験参加者は三宮に訪問経験のある学生 10 名とした。

まず図 11 にインタビュー対話サブシステムにおける対話の一例を示す.

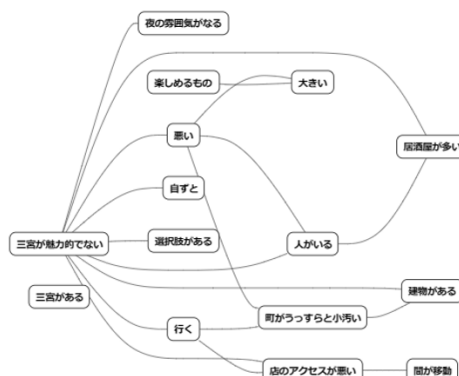
システムが参加者の答えの中からテーマを的確に特定し、次の質問を生成していることがわかる。また終了時の判断も的確に行っていることが確かめられた。

すべての実験でラダーアップ・ラダーダウンの両方を行っていることを確認した。また、参加者9名全員について、すべてのラダーリングが終わった後に対話を正しく終了した。

次にシステムが出力した評価構造図の一部を図 12 に示す。本図はインタビュー対話で収集された評価項目に対し、クラスタリング・概念抽出を行い、ESV で可視化したものである。



(a)三宮の魅力的なところ



(b)三宮の魅力的でないところ

図 12 システムで構築した評価構造図の例

＜インタビュー開始＞

system:これからインタビューを始めます。あなたが三宮の魅力的なところだと思うものを一つ答えてください。

user:街の雰囲気

system:街の雰囲気を感じるためには、何がどうなっていることが必要ですか？

user:自然と都市が融合していること

system:自然と都市が融合していると感じるためには、何がどうなっていることが必要ですか？

user:自然豊かである

system:自然豊かであると感じるためには、何がどうなっていることが必要ですか？

user:山や海が身近にあること

system:山が身近にあると感じるためには、何がどうなっていることが必要ですか？

user:建物が多すぎないこと

—省略—

system: それでは、自然と都市が融合しているとなぜ三宮が魅力的だと感じるのですか？

user: 気分が落ち着くため

system: 自然と都市が融合しているとなぜ三宮が魅力的だと感じるのですか？

user: いつでもリフレッシュできるため

system: 自然と都市が融合しているとなぜ三宮が魅力的だと感じるのですか？

user: 人の雰囲気も落ち着いているため

system: 落ち着いていると何故魅力的だと感じるのですか？

user: リラックスできるため

system: 落ち着いていると何故魅力的だと感じるのですか？

user: のんびりした時間を過ごせるため

—省略—

system: 観光名所を楽しむとなぜ三宮が魅力的だと感じるのですか？

user: 愛着が湧いたため

system: それ以外で、観光名所を楽しむと何故魅力的だと感じるのですか？

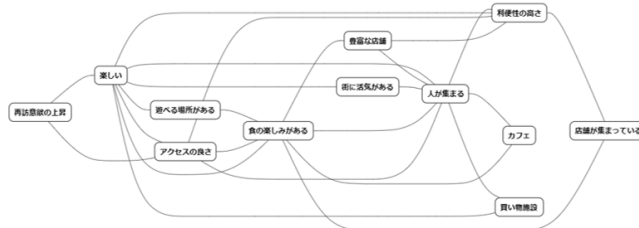
user: 他にはありません

system: それでは、他に三宮の魅力的なところだと思うものはありますか？

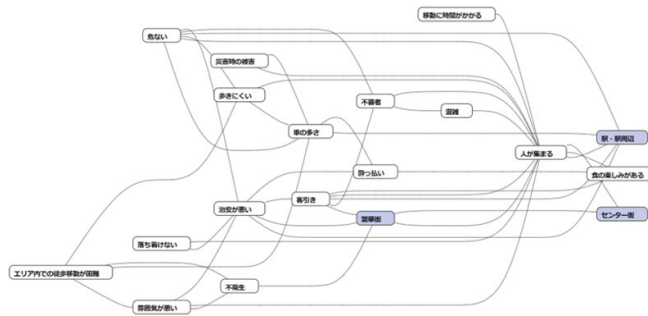
user: 他にはありません

system: ありがとうございます。では、インタビューは以上です。ありがとうございました。

〈終了〉



(a)三宮の魅力的なところ



(b)三宮の魅力的でないところ

図 13 手作業で構築した評価構造図の例

評価グリッド法自動化システムを評価するため、システムの結果(図12)を人間による評価構造図(図13)と比較する。

魅力的なところについては、どちらの図にも最上位概念に「再訪意欲の上昇」・「行く」が、最下位概念に「店舗が集まっている」・「楽しめるもの」が現れており、概ね一致していると言える。魅力的でないところについては、最上位概念では共通要素が見られなかったものの、最下位概念では「再訪意欲の上昇」・「行く」が一致していた。

また対話サブシステムで得られた評価項目の平均ノード数と平均エッジ数（因果ペア数）は表1に示されている通り、ともに手作業によるものよりも少なくなっている。これは、AIとの機械的な対話に参加者のモチベーションを上げず、話を広げにくくし、評価構造の抽出を難しくした可能性があると考えられる。今回は出力の安定性を確保するために、Top-pもTemperatureも共に0に設定したが、今後はパラメータを調整し質問の表現を豊かにするための「質問編集エージェント」の設置により、改善が可能であると考えられる。

表 1. 平均ノード数、平均エッジ数比較
(a)三宮の魅力的なところ (b)三宮の魅力的でないところ

	手作業	対話システム		手作業	対話システム
平均ノード	47.9	25.9	平均ノード	30.1	14.3
平均エッジ	46.1	24.4	平均エッジ	28.4	13.0

代表表現抽出サブシステムの評価を行うために、代表的な表現の「ふさわしさ」を評価グリッド法の経験者である大学院生2名に4段階で評価してもらった。合計点数が8点満点のうち6点以上だった割合を正解と定義し、正解率を計算した。その結果、表2に示す通り、BERTを使った方法では57.8%、Word2vecでは60.9%と、十分に高い正解率が観察された。これらの結果は、提案した手法の有効性を示すものである。

表 2. ノード正解率

	正解率(%)
BERT	57.8
Word2vec	60.9

5. まとめ

本研究では、評価グリッド法におけるインタビューを自動化するため、大規模言語モデルを用いたインタビュー対話システムの構築を行った。システムの出力結果を手作業による評価構造図と比較したところ、十分なノード数と正解率が得られ、手法の有用性が示された。

参 考 文 献

[1] 讃井純一郎, 乾正雄: レポートリー グリッド発展手法による住環境評価構造の抽出ー認知心理学に基づく住環境評価に

関する研究 (1)ー, 日本建築学会計画系論文, No.367, pp.15-22, 1986.

- [2] 土田義郎, 小酒祐貴: 評価グリッド法支援ツールの開発と応用, 日本建築学会技術報告集, 14.27, pp.205-208, 2008.
- [3] 大谷俊太, 橋本翔, 杉本匡史, 長田典子: 単語の係り受け関係に基づく印象評価構造の自動構築, 第17回日本感性工学会春季大会, 2D2-04, 2022.
- [4] 坂本有紀, 内田貴久, 石黒浩: 大規模言語モデルを用いたユーザモデル推定機能を持つ対話システムの検討, 人工知能学会全国大会論文集, vol. 37, p.301OS2c01, 2023.
- [5] 山崎天, 坂田亘, 川本稔己, 小林滉河, Nguyen Tung, 上村卓史, 中町礼文, 李聖哲, 佐藤敏紀: ペルソナー貫性の考慮と知識ベースを統合した HyperCLOVA を用いた雑談対話システム, 人工知能学会研究会資料 言語・音声理解と対話処理研究会, vol. 93, pp. 113-118, 2021.
- [6] C. Qian, X. Cong, C. Yang, W. Chen, Y. Su, J. Xu, Z. Liu, M. Sun: Communicative agents for software development, arXiv preprint arXiv:2307.07924, 2023.
- [7] J. Wei, et al: Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems, vol. 35, no.3, pp. 24824-24837, 2022.
- [8] 我妻正太郎, 照屋絵理, 竹内理: 自動作成された類義語抽出ルールによる類義語抽出精度向上手法, 情報科学技術フォーラム講演論文集(FIT), vol.20, no.2, pp.287-288, 2021.
- [9] M. Sugimoto, Y. Yagi, N. Nagata: How different tourist sites evoke different emotions: investigation focusing on the urban and rural sites in Japan, In: Kurosu, M., Hashizume, A. (eds) Human-Computer Interaction. HCII 2023. Lecture Notes in Computer Science, vol 14012, Springer, Cham, 2023.