JaSocial：LLM の社会的知能を評価するための日本語敬語使用フレームワーク

Muxuan Liu[1,2]　石垣達也[2]　宮尾祐介[3,2]　高村大也[2]　小林一郎[1,2]
[1] お茶の水女子大学大学院　[2] 産業技術総合研究所　[3] 東京大学
{liu.muxuan,koba}@is.ocha.ac.jp　yusuke@is.s.u-tokyo.ac.jp
{ishigaki.tatsuya, takamura.hiroya}@aist.go.jp

概要

本稿では，選択体系機能言語学の理論枠組みに基づき，大規模言語モデル（LLM）の社会的知能を評価するための新しいフレームワーク「JaSocial」を提案する．この理論を日本語敬語の使用に適用することで，敬語表現が持つ社会的背景や役割に基づき，より詳細かつ包括的に LLM の能力を分析することができるフレームワークを設計した．さらに，このフレームワークの運用を支える専用の評価データセットを新たに構築し，様々な LLM による日本語敬語生成の適切性を多角的に検証するための土台を提供する．

1　はじめに

「社会的知能」とは，人間関係を理解し管理し，さまざまな社会的状況において効果的なコミュニケーションを行う能力を指す [1, 2]．本研究では，日本語の敬語における大規模言語モデル（LLM）の「社会的知能」を評価することを目的としている．特に，日本語の敬語は社会関係や場面に強く依存しており，上司，同僚，部下などに対して異なる敬語を使用する必要がある．本研究では，社会的文脈と言語表現の役割を体系的に統合する視点を重視し，精密な評価を目指す．また，データ不足や敬語の使い方の許容度に関する主観的差異といった課題にも対応する必要がある．これらの課題を解決するために，本研究では選択体系機能言語学（SFL）の思想に基づき，「社会関係（第一層）→発話機能（第二層）→敬語表現（第三層）」という多層的な枠組みを採用し，社会的文脈（人間関係），発話機能（依頼，断り，報告など），および具体的な敬語表現の三者間の適合度に着目して LLM が異なる場面において適切な敬語を生成できるかを検証する．さらに，これらの評価を行うために新たにビジネスメールデータセットを作成し，これを研究の基盤として活用している．手法として，まず社会関係（目上，同等，目下など），発話機能（依頼，断り，報告，謝罪など），敬語レベル（尊敬語，謙譲語，丁寧語，インフォーマルなど）を多層的に注釈する．次に，言語モデルを用いた分類器でテキストを階層的に分類し，役割関係と言語機能の一貫性を分析し，実際の社会的場面に基づくメールや対話の生成結果に対して多層的な評価を行うことで，LLM の「社会関係の適合度」「発話意図の適合度」および「敬語使用の適合度」における優劣を明確にする．SFL の多層的視点を LLM 評価に導入することで，言語生成における社会的文脈の次元をより深く理解することに寄与する．また，実用的には，対話システムやメール生成などの場面で敬語の適切性を説明可能かつ定量的に評価できる手法を提供し，モデルの「社会的知能」を向上させるための指針となることが期待される．

2　関連研究

日本語生成文の自動評価手法は，近年大きな進展を遂げており，特にニューラルネットワークを用いた日本語生成およびその評価モデルが数多く提案されている．Han ら [3] は，日本語大規模言語モデルの性能を効率的に評価するため，複数の日本語評価データセットを活用した自動評価ツール「llm-jp-eval」を開発した．しかし，このツールは社会的文脈や敬語の適切性といった微妙な言語要素の評価には対応しておらず，特に敬語の評価では課題が残されている．Rahayu ら [4] は，敬語評価における文化的要因の影響を指摘している．この課題に対処するため，Feely ら [5] は，敬語の適切性を自動的に評価し，社会的文脈への適応性を測定するモデルを提案している．同様に，Sekizawa ら [6] は，事前

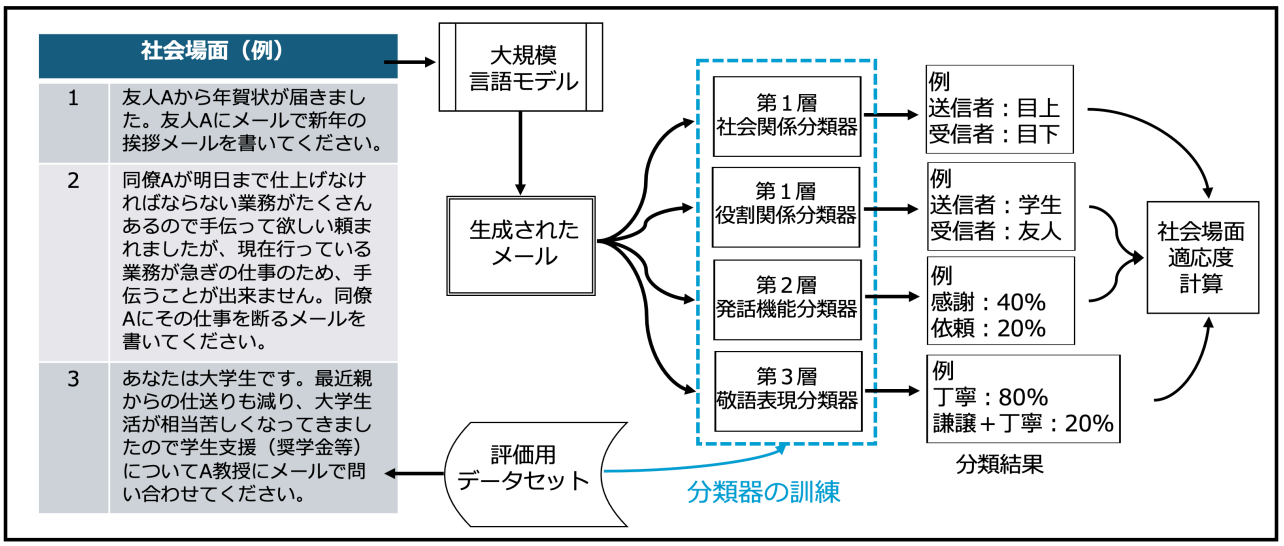

図 1　JaSocial 評価フレームワーク

学習された言語モデルを用いて社会的関係を考慮した敬語生成を実現し，その有効性を実証している．これらの研究は，敬語評価と生成において社会的文脈をモデル化する必要性を浮き彫りにしている．さらに，Ohashi [7] は，日本語の敬語使用における社会的役割の変化を明らかにするため，電子メールの分析を行った．研究では，役割が変化するにつれて，参加者が使用する敬語の頻度や形式が変化し，新たな役割や地位を反映することが示された．この結果は，敬語が単なる礼儀の表現ではなく，個人の役割認識や社会的関係性の動的変化を反映する重要な手段であることを示唆している．

ここで注目されるのが，選択体系機能言語学（SFL） [8] であり，言語を社会的文脈に基づいて分析するための強力な理論的フレームワークを提供しており，言語選択がどのように社会的関係や役割を反映するかを明らかにするのに有効である．Sato [9] は，SFL を用いた分析により，敬語が話者の社会的役割や関係性をどのように示すかを詳述している．特に，複雑な社会的要因を考慮した敬語分析が，敬語生成モデルの精度向上に寄与することを示している．また，SFL を活用した実践例として，教育や翻訳支援の分野での応用が挙げられる．例えば，Nakayama ら [10] は，特定の社会的状況に応じた敬語使用の学習を支援するプロトタイプシステムを開発し，その実用性を示した．このような実践は，敬語生成と評価が社会的文脈の理解と密接に関連していることを強調している．敬語生成および評価においては，社会的文脈や役割変化をモデルに組み込むことが重要であると考えられる．本研究では，これらの知見を基に，より汎用性の高い評価フレームワークの構築を目指す．

3　フレームワーク設計

JaSocial の評価フレームワークの全体像を図 1 に示す．

3.1　評価基準データセットの構築

表 1　各社会関係とメール通数

社会的関係	送信者→受信者	合計
目上から目下	教員→学生	200
目上から目下	従業員→部下	200
目下から目上	学生→教授	200
目下から目上	従業員→上司	200
同輩から同輩	学生→友人	200
同輩から同輩	従業員→同僚	200
合計	-	1200

Liu ら [11] は，社会的文脈を考慮した日本語ビジネスメールのコーパスを構築した．このデータセットは，文法構造のみにとどまらず，社会的文脈の中で意味を構築し，交換するための社会記号体系（social semiotic system）と捉える選択体系機能言語学の理論を基盤としており，送信者と受信者の社会的状況に基づく 8 種類の「発話機能」（感謝，謝罪，依頼，提案，お知らせ，挨拶，問い合わせ，催促）の大分類タグでメール全体が注釈されている．しかし，元のデータセットに付与されていたタグは，メール全体に適用された包括的なものであり，詳細な分析を行うには十分ではなかった．そこで，本研究では，元のデータセットから可能な限り同一の社

会的役割を持つ送信者が，異なる立場に対してメールを送るデータを選定した．具体的には，表 1 に示す社会関係から 1200 通のメールを抽出し，既存のタグを文単位に分割・再注釈することで，SFL に基づく詳細な分析が可能な形式へ再加工した．

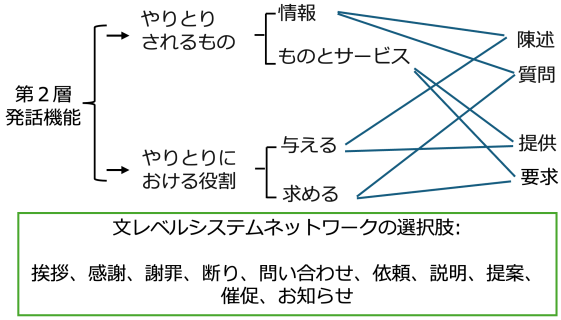

図 2　文レベル視点出発の発話機能および具体的な意味選択の選択肢

図 2 で示されている文レベル視点出発の発話機能および具体的なシステムネットワークの選択肢の中から，各文が主に表している選択肢をタグとして付与した．文レベルシステムネットワークの選択肢は，発話機能体系の横断性および文脈依存性により，自由に組み合わせ可能である．第 2 層発話機能は「やりとりされるもの」と「やりとりにおける役割」の 2 軸から構成され，それぞれが具体的な動き（陳述，質問，提供，要求）へと展開する．この階層的構造により，複数の選択肢が同一発話内で並列的または重複的に使用される．

表 2　発話例とタグ付け

例 1: A 殿からの引継ぎ資料を読んでいたのですが、いくつかファイルが欠けているようです.
第二層: やり取りされるもの: 情報 第二層: やり取りにおける役割: 与える 第二層: 送信者の動き: 陳述: 説明 第三層: 丁寧語
例 2: 大変申し訳無いのですがご理解の程よろしくお願いします.
第二層: やり取りされるもの: 情報 第二層: やり取りにおける役割: 与える 第二層: 送信者の動き: 陳述: 謝罪＋挨拶 第三層: 謙譲語＋丁寧語
例 3: 申し訳ないけどなるべく早くお願い！
第二層: やり取りされるもの: 情報＋ものとサービス 第二層: やり取りにおける役割: 与える＋求める 第二層: 送信者の動き: 陳述+要求: 謝罪＋催促 第三層: インフォーマル

例えば，表 2 の例 1 では「陳述」と「説明」が「情報を与える」行為として結合し，丁寧語を用いて不足ファイルの存在を受信者に伝える意図を表現している．また，例 2 では「謝罪」と「挨拶」が「情報を与える」行為として同時に現れ，発話者の謝罪と礼儀を示す意図が読み取れる．一方，例 3 では「謝罪」と「催促」が「情報」と「ものとサービス」の両方を含む行為として結合し，インフォーマルな表現を通じて発話者の親密さや緊急性を示している．これらの例は，選択肢が発話の目的や文脈に応じて柔軟に組み合わされることを示しており，選択肢間の固定的な対応を超えた多様性を反映している．以上より，文レベルのタグ付けのデータセットは，発話者の意図や文脈に応じた発話の多様性を実現する基盤を提供でき，言語使用の文脈的特徴や敬語の運用に関するさらなる研究を支える重要な基盤リソースとして位置づけられる．

3.2　分類器の評価プロセス

評価プロセスでは，メール全体と文単位の特徴を分類器で分析し，社会的文脈，発話機能，敬語使用を統合した詳細なレポートを生成し，そのレポートから社会場面適応度を算出する．提案する「JaSocial」フレームワークの分類器は以下の三層構造から成る：

- **第一層分類器：社会関係，役割関係**
  社会関係分類器は，送信者と受信者の上下関係（目上→目下，目下→目上，同輩→同輩）を判定するために 3 つの二値分類器を使用する．それぞれの分類器は，メール中の文脈的手がかりに基づき，「1（該当）」または「0（非該当）」を出力する．3 つの分類器の中で出力が「1」となった結果をレポートに反映する．
  役割関係分類器は，送信者と受信者の具体的な社会的役割（例：学生→教師，上司→部下，同僚→同僚など）を識別するために 6 つの二値分類器を使用する．各分類器の出力が「1」の場合，その関係性がレポートに記載される．
- **第二層分類器: 発話機能**
  発話機能分類器は，文単位で複数の発話機能（例：感謝，謝罪，依頼など）を同時に分類し，メール全体の発話機能分布を集計する．各発話機能の占有率（メール中にその機能が占める割合）は，以下の計算式に基づいて算出される：

$$\text{占有率} = \frac{\text{該当する発話機能を持つ文の数}}{\text{メール全体の文数}}$$

表 3 メールの内容とタグの表

項目	内容
社会場面	あなたは今日学校で出題された宿題の内容を忘れてしまいました．その上，その宿題の提出は明日となっています．そこで，どのようなメールを送れば，友人から早く宿題の内容を返信してくれるか考えなさい．
第一層: 共通_社会関係	送信者の社会的立場: 同輩 受信者の社会的立場: 同輩 送信者の役割: 学生 受信者の役割: 友人 内外関係: 内 送信者数: 個人 受信者数: 個人
件名	宿題，ヘルプ！！！
受信者呼び名	A さん
本文	文 1: ごめん！ 文 2: 今日の宿題の内容，申し訳ないけど，もう一回教えて！ 文 3: ホントにごめん！！！ 文 4: 助けて欲しい．
タグ	文 1 タグ: [第二層: やり取りされるもの: 情報, 第二層: やり取りにおける役割: 与える, 第二層: 送信者の動き: 陳述: 謝罪, 第三層: インフォーマル] 文 2 タグ: [第二層: やり取りされるもの: 情報, 第二層: やり取りにおける役割: 与える＋求める, 第二層: 送信者の動き: 陳述＋要求: 謝罪＋依頼, 第三層: インフォーマル] 文 3 タグ: [第二層: やり取りされるもの: 情報, 第二層: やり取りにおける役割: 与える, 第二層: 送信者の動き: 陳述: 謝罪, 第三層: インフォーマル] 文 4 タグ: [第二層: やり取りされるもの: ものとサービス, 第二層: やり取りにおける役割: 求める, 第二層: 送信者の動き: 要求: 依頼, 第三層: インフォーマル]
送信者	<sender>

例えば，メール中に 10 文が存在し，そのうち「感謝」と分類された文が 4 文,「依頼」と分類された文が 2 文の場合，それぞれの占有率は以下のように計算される：

– 感謝の占有率：4/10 = 40%

– 依頼の占有率：2/10 = 20%

- **敬語表現の割合**
  敬語表現分類器は，各文における敬語の使用形式を判定する．この分類器は文単位で敬語使用のいずれかを出力し，メール全体の敬語使用割合を計算する．

これにより，LLM が社会的文脈に応じた適切な敬語表現や発話機能を生成しているかを，直感的かつ定量的に評価することが可能となる．

3.3 社会場面適応度の計算

評価基準データセットの一部を教師データとして活用して，各層の分類器に対して「タグマッチングモデル」を構築し，社会場面における典型的な発話機能の割合や敬語使用の分布（尊敬語・謙譲語・丁寧語など）を統計的に把握し，内部的に「理想タグ」として保持する．

メール生成の評価プロセスでは，タグマッチングモデルが推定する「理想的な社会関係・役割関係」「理想的な発話機能の分布」「理想的な敬語使用率」などの基準を，生成されたテキストの特徴と比較する．一致度は以下の指標によって測定される：(1) 社会関係および役割関係の一致度，(2) 発話機能占有率の差異，(3) 敬語使用率の期待値との差分．これらのスコアを 0〜1 に正規化した上で加重平均を行い，メール全体の社会場面適応度を算出する．例えば，学生→教師の問い合わせ場面では，理想タグとして「目下→目上」の社会関係,「依頼・感謝を中心とした発話機能の分布」，および「謙譲語優位の敬語使用」が設定される．この理想タグに対する生成メールの一致度を定量的に評価することで，該当社会場面への適応度を総合的に測定することが可能である．

4 おわりに

本研究では，日本語敬語を用いた LLM の社会的知能を評価するための新しいフレームワーク「JaSocial」を提案した．本稿では，フレームワークの設計とデータセット構築を中心に述べたが，今後の課題として，各分類器の実装と性能評価，実際の LLM 生成結果を用いたフレームワークの検証は今後の課題として残している．

謝辞

この成果は, 国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の助成事業 (JPNP20006) による支援の結果得られたものである.

参考文献

[1] Edward L Thorndike. Intelligence and its uses. **Harper's Magazine**, Vol. 140, pp. 227–235, 1920.

[2] Daniel Goleman. **Emotional Intelligence: Why It Can Matter More Than IQ**. Bantam Books, 1995.

[3] Namgi Han・植田暢大・大嶽匡俊・勝又智・鎌田啓輔・清丸寛一・児玉貴志・菅原朔・BowenChen・松田寛・宮尾祐介・村脇有吾・劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会年次大会発表論文集, No. 30, 2024.

[4] Desi Rahayu. Japanese honorific language in various domains. **Proceedings of the Fourth Prasasti International Seminar on Linguistics (Prasasti 2018)**, Vol. 134, pp. 1–13, 2018.

[5] Weston Feely, Eva Hasler, and Adrià de Gispert. Controlling Japanese honorifics in English-to-Japanese neural machine translation. In Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Nobushige Doi, Yusuke Oda, Ondˇrej Bojar, Shantipriya Parida, Isao Goto, and Hidaya Mino, editors, **Proceedings of the 6th Workshop on Asian Translation**, pp. 45–53, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Takashi Sekizawa and Hitomi Yanaka. Keigo transformation using pre-trained language models. **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 1234–1245, 2023.

[7] Jun Ohashi. An emerging role-identity and honorifics: A longitudinal study of email exchanges in a japanese community. **Journal of Pragmatics**, Vol. 127, pp. 36–55, 2018.

[8] M. A. K. Halliday and Christian M. I. M. Matthiessen. **Halliday's introduction to functional grammar /**. Routledge,, Abingdon, Oxon :, 4th ed. edition, 2014.

[9] Yoko Sato. Analyzing keigo usage through systemic functional linguistics. **Japanese Linguistics**, Vol. 40, No. 2, pp. 45–67, 2024.

[10] Tetsuo Nakayama and Yusuke Morita. Development of a prototype system for learning keigo usage in specific social contexts. **Educational Technology Research**, Vol. 32, No. 1, pp. 89–102, 2009.

[11] Muxuan Liu, Tatsuya Ishigaki, Yusuke Miyao, Hiroya Takamura, and Ichiro Kobayashi. Constructing a japanese business email corpus based on social situations. In **Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation**, pp. 499–509, 2023.