WMNZ リランキングにより LLM と機械学習分類モデルを統合したハイブリッド発話意図推定

山田　仰　小林　拓也
ＮＴＴドコモ　サービスイノベーション部
aogu.yamada.rm@nttdocomo.com　takuya.kobayashi.py@nttdocomo.com

概要

対話システムにおける発話意図推定の精度向上を目指し，キーワード特定に優れた機械学習分類モデルと大局的な文脈理解能力を持つ大規模言語モデル（LLM）による分類手法を組み合わせたハイブリッドな発話意図推定手法を提案する．両モデルは相補的な課題を持つため，提案手法ではそれぞれの分類結果をリランキング手法である WMNZ（Weighted-MNZ）により統合する．これにより機械学習モデルの過剰反応といった弱点を LLM が補完したような特性を持つ発話意図推定が可能となる．28 クラスの発話意図分類における評価実験の結果，提案手法は各単体モデルや代表的な統合手法である RRF を上回る精度を達成した．

1 はじめに

対話システムにおいて，ユーザの発話意図を正確に推定し，応答精度を向上させることは，対話システムの実用性を左右する重要な課題である[1]．この発話意図推定は，ユーザの入力テキストを事前に定義された複数の意図カテゴリのいずれかへと分類するクラス分類のタスクとして扱われる．

従来，この分類タスクのアプローチには主に二つの手法が存在する．

- **機械学習モデルによる分類[2]：**特定の対話ドメインにおける発話を想定したテキストと意図カテゴリのラベルのペアからなるデータセットを用いて分類モデルを構築し，これを用いて発話を分類する手法
- **大規模言語モデル（LLM）による分類[3]：**大規模言語モデルが持つ強力な文脈理解能力を活用し，プロンプトに記述された指示や少数の例に基づいて発話を分類する手法

これらの従来手法の特徴として，機械学習モデルによる手法が特定のキーワードを捉える能力に長ける一方，LLM による手法が文脈全体を包括的に捉える能力に優れるという，互いに相補的な特性を持つ．

そこで本研究では，両者の長所を組み合わせたハイブリッドな発話意図推定手法を提案する．具体的には，機械学習モデルと LLM がそれぞれ独立して算出した分類結果を両者の特性を考慮して重み付けを行う WMNZ (Weighted-MNZ) [4]というリランキング手法を用いて統合する．これにより各モデルの弱点を補い合い，精度を向上させることをめざす．

本研究では，28 クラス分類の発話意図推定タスクを対象とし，提案手法の有効性を検証する．実験の結果，提案手法が従来の意図推定手法や，他のリランキング手法と比較して，最も高い精度を示した．

2 関連技術

発話意図推定に関連する技術動向について述べる．具体的には，機械学習モデルによる分類手法と，大規模言語モデル（LLM）による分類手法，そして複数の分類手法の結果を統合するハイブリッドな分類手法について述べる．

2.1 機械学習モデルによる分類手法

従来，対話システムにおける発話意図推定では，大規模なテキストコーパスから分類モデルを構築する手法が主流であった．SVM（Support Vector Machine）[5]や BERT (Bidirectional Encoder Representations from Transformers) [6]に代表されるモデルによりドメイン固有のテキストと意図カテゴリのラベルから成るデータセットを学習することで，高い分類モデルを構築することが期待される．また大規模なテキストコーパスで事前に言語の一般知識を学習させたモデルに対し，タスク固有のラベル付きデータを少量用いて追加学習を行うことで準備に必要なコストを軽減する手法もある．

しかしこれらの機械学習モデルによる分類は学習データ中の特定のキーワードに過度に依存する傾向があり，文脈全体を踏まえた判断ができないという課題が指摘されている[7]．例えば，「ドコモの新プランである XXX を契約したが，アカウント情報を忘れた」という入力に対し，本来ならば「アカウント設定」というカテゴリに分類すべきところ，ドコモの新プラン名というキーワードを過度に重視し「料金プラン問い合わせ」というカテゴリに誤分類するというような問題が発生してしまう．

2.2 LLM による分類手法

近年の GPT-3[8]以降の大規模言語モデル（LLM）の登場により，文脈内学習（In-Context Learning）と呼ばれる新たなアプローチが可能となった．これはモデルに追加の学習を施すことなく，プロンプトにタスクの指示や少数の分類例（Few-shot）を提示するだけで，モデルがその意図を汲み取りタスクを遂行する能力である．分類例を全く与えないゼロショット学習も可能であり，タスク固有の学習データが不要なため，迅速なシステム構築に適している[9]．LLM は膨大な知識を内包し，文章全体の意味やニュアンスを包括的に理解する能力に長けている．一方で，その性能はプロンプトの設計に大きく左右され，また，特定のドメインに深く特化した細かな知識や分類境界の判断においては，機械学習モデルに及ばない場合がある[10]．

2.3 複数手法の分類結果を統合するハイブリッド手法

複数の分類手法を組み合わせるハイブリッドな研究も進められている．複数の異なる分類器の結果を統合するアンサンブル学習は，単一のモデルよりも頑健で高精度な予測を実現するための一般的な手法である[11]．統合手法には，各モデルの出力を単純に多数決で決定する方法や，予測スコアを平均化する方法などがある．

より高度な統合手法として，各分類器が出力する候補のスコアリスト（ランキング）を統合して最終的な順位を決定するリランキングが存在する．情報検索の分野で提案された RRF (Reciprocal Rank Fusion)[12] は，複数の検索エンジンの結果を順位に基づいて統合する代表的な手法である．

本研究で利用する MNZ (Min-Normalized Zero) 及び WMNZ (Weighted-MNZ) もリランキング手法の一つである．MNZ は，複数のモデルがそれぞれ出力するスコアをモデルごとに最小値・最大値を用いて正規化（Min-Max Normalization）し，それらを足し合わせることで最終的なスコアを算出する．これにより，スコアのスケールが異なるモデル間でも公平な統合が可能となる．さらに WMNZ は MNZ を拡張し，各モデルの信頼度や特性に応じて重みを付けてスコアを統合する．これにより例えば文脈理解に優れた LLM とキーワード特定に優れた機械学習モデルといった，異なる特性を持つモデルの長所を選択的に活用した，より精緻な結果統合が期待できる．本研究では，この WMNZ の特性が発話意図推定の精度向上に有効であると考え採用する．

3 提案手法

本研究では機械学習モデルと大規模言語モデル（LLM）の長所を活用し，発話意図推定の精度向上をめざすハイブリッド手法を提案する．本手法の特徴は両モデルが独立して算出した分類結果をそれぞれの予測信頼度と両者の予測の一致度を考慮して動的に重み付けし統合する点にある．この統合処理には複数のランキング結果を効果的に統合する手法である WMNZ を適用する．

3.1 処理フロー

提案手法の処理フローは以下の通りである．

1. ユーザからの入力発話を受け取る．
2. 入力発話を機械学習モデルによる分類器と LLM による分類器のそれぞれに並列に入力し，全意図カテゴリに対するスコアリストを個別に算出する．
3. 得られた 2 つのスコアリストを WMNZ アルゴリズムに入力し，スコアの再計算を行う．
4. 再計算された最終スコアが最も高いカテゴリをユーザの発話意図として決定する．

3.2 各モデルによるスコアリング

3.2.1 機械学習モデルによるスコア算出

本研究では，機械学習モデルとして SVM に基づく分類器を利用する．まず特定のドメインの対話ログからなる学習データを用い，入力発話を TF-IDF などの手法によってベクトル化する．このベクトル表現を特徴量として SVM を学習させタスク特化の分類モデルを構築する．予測時には入力発話を同様に

ベクトル化し，学習済み SVM に入力することでスコアリストを得る．

3.2.2 LLM によるスコア算出

本研究では LLM には gpt4.1 のモデルを用いる．ゼロショット学習のアプローチを採用し, LLM へのプロンプトを工夫することで，各カテゴリに対するスコアを直接出力させる．具体的には，プロンプト内でタスク定義，分類対象の全カテゴリリスト，そして入力発話を提示するとともに，「各カテゴリに合致する確率スコアを 0 から 100 の整数値で，指定の JSON 形式で出力してください」といった指示を与える．LLM は指示に従い，発話がどのカテゴリにどの程度合致するかを判断し，全カテゴリに対するスコアを構造化データとして出力する．この出力値を正規化することで, LLM によるスコアリストを得る．

3.3 WMNZ によるスコア統合

SVM によるスコアリストと LLM よるスコアリストを WMNZ により統合する（式 1）．

$$Score(c) = \sum_i score_i(c) * \sum_i w_i \qquad (1)$$

c は順位付けしたい分類カテゴリを表す．i は分類モデルのインデックスを表し，本研究では SVM と LLM がある．$score_i(c)$，w_i は，それぞれ分類モデル i がカテゴリ c に与えた正規化済みスコア，分類モデル i の推定性能に基づく重みである．この重みは各モデルが元々持つ信頼度に応じて設定されるハイパーパラメータである．

式 (1) は２つの独立した効果をもつ総和の積として設計されている．式 (1) の一つ目の総和はカテゴリ c がどれだけ強く評価されているかを示し，高スコアを与える分類モデルが多いほど大きな値を取る．つまり両手法の両方で上位の結果となったカテゴリほど高い値を取り，逆に片方のモデルのみで上位のカテゴリは低い値を取ると言える．これにより機械学習モデルだけが高いスコアを出すような，特定のキーワードに過度に引っ張られるケースを低いスコアとして見積もることができる．式 (1) の二つ目の総和はカテゴリ c を返す分類モデル群の総合的な信頼度を示す．元々の分類器の性能として，機械学習モデルは LLM より高い性能を持つことが知られているため, w_i の値について機械学習モデルを LLM より高く設定する．これにより，式(1)の一つ目の総和の特性も踏まえると, 式全体として「強く評価され，かつ信頼できるシステムに支持されているカテゴリほど上位になる」という特性を持つような分類器を構築することができる．

4 評価実験

提案手法の有効性を検証するため，発話意図推定タスクにおける評価実験を行った．本章では，実験設定，評価結果，および考察について述べる．

4.1 実験設定

データセット: ドコモのチャット応対窓口に寄せられた問い合わせログをもとに作成した 28 クラスの意図分類データセットを使用した．このデータセットを学習用，評価用に分割し，それぞれモデル構築と性能評価に利用した．

評価指標: 本実験では，各手法が予測した意図カテゴリが，正解ラベルと一致した割合を示す正解率（Accuracy）を評価指標として用いた．

比較手法: 提案手法の有効性を多角的に評価するため，以下の 4 つの手法を比較対象とした．

1. **LLM 単体**: 3.2.2 節で述べた方法で，LLM（GPT-4）のみを用いて分類を行う．
2. **SVM 単体:** 3.2.1 節で述べた方法で，SVM 分類器のみを用いて分類を行う．
3. **RRF (LLM＋SVM):**代表的なリランキング手法である RRF を用いて，各カテゴリの最終スコアを，SVM と LLM の出力ランキングを統合する形として算出する．
4. **提案手法:** 3.3 節で述べた WMNZ ベースのスコア統合手法．式（1）における w_i については SVM は 1.25，LLM は 0.75 とした．

4.2 評価結果

表 1 に，各手法における発話意図推定の正解率を示す．

表 1：各手法の正解率（Accuracy）

手法	正解率
LLM 単体	63%
SVM 単体	68%
RRF(SVM＋LLM)	72%
提案手法	76%

実験結果から，まずベースラインである単体モデルの性能を見ると，ドメイン特化の学習を行った SVM 単体が 68%の正解率を達成し，汎用的な LLM 単体の 63%を上回った．これは本タスクがドメイン固有の専門知識を必要とすることを示唆している．

次にハイブリッド手法を見ると，RRF は 72%，提案手法は 76%の正解率を達成し，いずれも単体モデルの性能を大きく上回った．これにより特性の異なる両モデルの結果を統合することの有効性が確認された．そして比較対象としたすべての手法の中で，本研究の提案手法が最も高い正解率を達成した．

4.3 WMNZ によるスコア統合

提案手法が最も高い性能を示した要因は，その精緻なスコア統合メカニズムにあると考えられる．RRF は，SVM と LLM の出力を組み合わせることで SVM 単体を上回る性能を示したが，その統合は各モデルが出力した「順位」情報のみに依存する．そのため，SVM がキーワードに過剰反応して誤ったカテゴリに高い順位を付けた場合でも，その影響を十分に抑制することが難しい．それに対し，本提案手法は，(1)モデルの基本信頼度に基づく重み付け，そして(2)両モデルの予測順位が高いほど高いスコアを取る，という２つの情報を複合的に利用する．これにより SVM が単独で高スコア・高順位を出力したとしても，LLM の評価が低ければ結果的なスコアは高くならず，誤分類が抑制される．逆に，SVM と LLM の両方が正解カテゴリを高く評価した場合には，その候補のスコアがさらに強調され，確信度の高い予測が可能となる．この動的かつ多角的なスコア統合が，RRF との差を生み，最も優れた性能達成に繋がったと考察される．

5 おわりに

本研究では，機械学習モデル（SVM）と大規模言語モデル（LLM）の相補的な特性に着目し，両者の出力を WMNZ ベースの手法で統合するハイブリッドな発話意図推定手法を提案した．実際の問い合わせデータを用いた評価実験の結果，本手法は各単体モデルや代表的なリランキング手法である RRF を上回り，最も高い分類精度を達成した．この結果は，モデル全体の信頼度に基づく静的な重み付けと，入力ごとの予測一致度を考慮した動的なスコア統合が，発話意図推定の精度向上に有効であることを実証するものである．今後の展望として，本手法の他ドメインへの適用可能性の検証や，重み・ボーナススコアといったハイパーパラメータの自動最適化手法の検討を進めていきたい．

参考文献

[1] Hongshen Chen, et al. A Survey on Dialogue Systems: Recent Advances and New Frontiers. **Acm Sigkdd Explorations Newsletter**, Vol. 19, No. 2, pp.23-35, 2017.

[2] Dilek Hakkani-Tür, et al. Multi-domain Joint Semantic Frame Parsing Using Bi-directional RNN-LSTM. **Interspeech**, pp.715-719, 2016.

[3] Timo Schick, et al. Exploiting Cloze-questions for Few-shot Text Classification and Natural Language Inference. **Proceedings of the 16th conference of the European chapter of the association for computational linguistics**, pp.255-269, 2021.

[4] Shengli Wu, et al. Data Fusion with Estimated Weights. **European Conference on Information Retrieval**, pp.275-286, 2005.

[5] Corinna Cortes, et al. Support-Vector Networks. **Machine learning**, Vol. 20, No. 3, pp.273-297, 1995.

[6] Jacob Devlin, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. **Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics**, pp.4171-4186, 2019.

[7] Suchin Gururangan, et al. Annotation Artifacts in Natural Language Inference Data. **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics**, Vol. 2, pp. 107-112, 2018.

[8] Tom Brown, et al. Language Models are Few-shot Learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877-1901, 2020.

[9] Takeshi Kojima, et al. Large Language Models are Zero-shot Reasoners. **Advances in neural information processing systems**, Vol. 35, pp. 22199-22213, 2022.

[10] Zihao Zhao, et al. Calibrate before use: Improving Few-shot Performance of Language Models.

International conference on machine learning, pp. 12697-12706, 2021.

[11] Thomas G Dietterich. Ensemble Methods in Machine Learning. **International workshop on multiple classifier systems**, pp. 1-15, 2000.

[12] Gordon V Cormack, et al. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. **Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval**, pp. 758-759, 2009.