文字起こしテキストから得た質問のタグ推定

自見仁太朗[1]　大野正樹[2]　橋本泰一[2]　嶋田和孝[1]
[1]九州工業大学 大学院情報工学研究院　[2]RevComm
jimi.jintaro102@mail.kyutech.jp　{masaki.ono,hashimoto}@revcomm.co.jp
shimada@ai.kyutech.ac.jp

概要

本論文では，営業員と顧客の会話の文字起こしテキストから得られた質問のタグを推定するタスクに取り組む．質問には冗長性や言い間違えを含まれるため，その内容を理解することが難しい．私たちは大規模言語モデルによって質問のトピックを推定することで，質問を簡潔な文書に変換した．さらに，質問の内容を理解することに適した単語のみをタグとして獲得した．評価では，実際の営業員と顧客の会話から得た約 1 万件の質問とオープンなモデルである Llama2 を用いた．

1　はじめに

音声認識技術と自然言語処理技術の発展により，それらの技術を応用して電話対応をサポートすることが期待される．重要な応用の一つは，顧客と営業員のやり取りを分析し，質問における傾向を認識することである．これは営業員に既存の問題に対する知見を提供し，販売戦略の更新につながる．

本論文では，営業員と顧客の会話の文字起こしテキストから得られた質問のタグを推定するタスクに取り組む．タグとはその質問の内容を識別できる目印を指す．本論文の手法の概要を図 1 に示す．質問に含まれている「InsideSales」と「案件」は，KeyBERT[1] によってキーワードとしては認識されたが，これらは質問の内容を示すには十分ではないと本論文では考えた．そして 2 段階の方法でタグを抽出することを試みた．

本論文が対象とする，営業員と顧客の会話の文字起こしテキストには 2 つの特徴がある．1 つ目は，テキストが冗長性や言い間違えを含み，Web 記事などの人手で整理されたテキストよりもより複雑であることである．2 つ目は，多様な話題を含むことである．これは一般的に知られており，発話の意図を特定する意図検出タスクのデータセットである

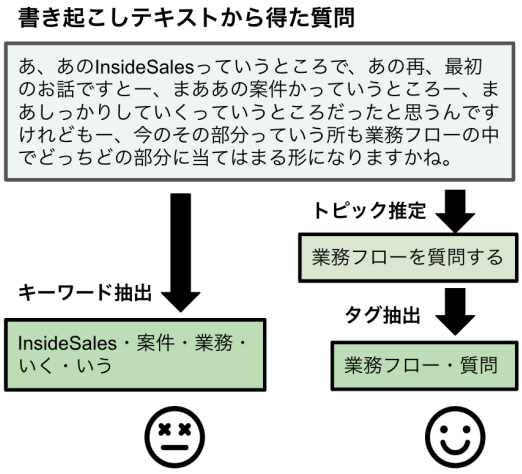

図 1　本論文の手法の概要

Banking77 [2] と Clinic150 [3] は，発話に対して 70 種類以上の分類を持つ．また，私たちが ChatGPT[1)] によって，営業員と顧客の会話から得た質問群に対してトピック推定を行ったところ，10,334 の質問から 7,235 のトピックが生成された．

1 つ目の特徴に対処するために，私たちは大規模言語モデル（LLM）によって質問のトピックを推定することで，簡潔で言い間違えがないテキストを生成する．近年の研究は，LLM が誤りを含む文書を正しい文書に変換できること [4] と，LLM によるアノテーションは人手のアノテーションと同等であることを報告している [5][6][7]．これらは，LLM が複雑なテキストを理解する能力があることを示唆する．

2 つ目の課題に対処するために，質問の傾向を分析する要求を，質問からタグを抽出するタスクとして扱う．一般的に，意図検出タスクは教師あり分類問題として定式化されるが，本論文の対象はラベルなしテキストである．LLM によってアノテーションのコストを低減し教師データを作成すること [8] が考えられるが，その場合には分類先のクラスを手動で決定する必要があるため，多数のクラスを扱う

1）　https://chat.openai.com/

表 1 質問とトピックとタグの例：(解析-分析) は「解析」と「分析」の 2 つのタグがマージされていることを示す

質問文	トピック	タグ
ちょっとすみません，わたくしの方で**人名**のえーとスケジュールが把握できませんでして，えーとー確認次第折り返しさせていただいてもよろしいでしょうか	スケジュール確認の依頼	スケジュール確認，依頼
この MiiTel とこに活用していきたい分析の部分であったりだとか，そういった機能の部分で何かございますか.	MiiTel の分析機能についての質問	質問,MiiTel,(解析-分析) 機能
ごめんなさい，御社内でのご検討状況進捗あればと思いまして，お電話差し上げたんですけれども，何かございますでしょうか.	御社内の検討状況を質問する	質問, 検討, 御社, 社内

場合に実装が困難である．

本論文は 2 段階の手法を質問に適用した．1 つ目は，トピック推定であり，LLM を使用して与えられた質問のトピックを推定する．ここでは，LLM の出力を制限するために In-Context Learning（ICL）[9] をする．さらに精度を向上させるために，推論を複数回行い [10]，尤度に着目して誤って推定されたトピックを削除する（Self Consistency）．2 つ目は，タグの抽出であり，与えられたトピックからタグとして考えられる単語を導出する．トピック推定の出力はある程度統一されたフォーマットを持つため，ここではルールベースの手法を適用する．入力と各段階における出力の例を表 1 に示す．

実際の営業員と顧客の会話から得た約 1 万件の質問を対象にして評価を行った．Llama2 [2] を使用して，冗長性や言い間違えを含む日本語のテキストから，80%の正解率でトピックを抽出することができた．ICL が LLM の出力を制限することに効果があった．

2 提案手法

2.1 トピック推定

本節では，質問からそのトピックを抽出する手法について説明する．この手法は，冗長性や言い間違えを含んだテキストを短く綺麗なテキストに変換し，タグ抽出の精度を上げる役割がある．質問応答の分野では，要約手法によって不要な情報を取り除き，精度が向上したことが報告されている [11]．

本論文の目的は質問の傾向を分析することであり，質問から得られたトピックの集約を容易にするためにトピックの表層系が似通っている必要がある．そこで，LLM の出力を制御するために ICL を適用した．ICL は入力と出力の例をプロンプトに含ませることで，LLM に出力の形式や内容を伝える手法である．先行研究で ICL が様々なタスクにおける LLM の性能を向上させたことが報告されている [12]．ICL の副作用としてプロンプトのトークン数が増え，計算量を増加することが挙げられる．そこで，計算の効率を上げるために，本論文では 1 つのプロンプトに複数の入力を含ませた．

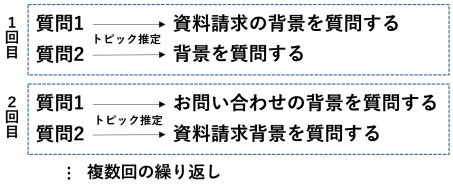

図 2 トピック推定を複数回行った場合の出力

トピック推定の精度を上げるために 2 つの手法を使用した．1 つ目は，与えられた質問に対してトピックが導出される尤度を計算し，その値が低いトピックを扱わないことである．2 つ目は，複数回の推論を行うことであり，その例を図 2 に示す．プロンプトの詳細については付録 A に記述する．

2.2 タグ抽出

本節では，トピック群からタグ候補を抽出し，質問にタグを付与する手法について説明する．トピック抽出における出力が一定のフォーマットを持つために，本論文では，タグ候補の抽出のためにルールベースの手法を使用した．

タグは，一目で質問の内容を識別できる目印であり，トピックに存在するキーワードと等価ではないと本論文では考えた．例えば，トピックである「ツール導入の状況について質問する」から，「ツール導入」と「質問」をタグとして抽出したい．「状況」はそれだけでは情報が乏しく，タグとしては不適切である．本論文ではストップワードを使用することで，対象外のタグ候補を特定する．

2） https://ai.meta.com/llama/

タグ抽出は，4 つのステップで構成される．はじめに，係り受け解析によってチャンク単位での分割を行い，ルールを適用してタグ候補を得る．ルールは，1) 名詞・動詞・助動詞以外の語の削除，2) 非自立語のみで構成されるチャンクの削除，3) 文末のサ変動詞の削除，4) 数列・人名の正規化，5) 時制の統一，の 5 つである．

次に，前ステップで作成したタグ候補から不適切なものを取り除くために，名詞を対象にしてストップワードを生成する．本論文では，「複合名詞や名詞句の後方に頻繁に現れる語は，他の名詞に情報を補強されることが多いため，単体では情報が乏しい」と仮定した．そして，スコアリング関数を定義し，その値が閾値を超えたものをストップワードとし，タグ候補から取り除いた．係り受けに着目して語の曖昧性を定義した先行研究として SPIQA[13] があり，ここでは質問応答システムへの入力において修飾語のない語は情報が乏しいと判断した．

スコアリング関数の定義を示す．ある名詞 x がトピック群に出現する回数を n，複合名詞の先頭以外に出現した回数を n_{h1}，語 x が名詞句の先頭以外に出現した回数を n_{h2} とする．また，名詞 x が抽出されたタグ候補の先頭以外に出現する回数を m_h とし，タグ候補中の先頭以外に出現する名詞全ての出現回数を m_{all} とする．

$$score_x(n, n_{h1}, n_{h2}, m_x, m_{all}) = \frac{n_{h1}}{n} * \frac{n_{h2}}{n} * \frac{m_h}{m_{all}} * 100 \quad (1)$$

3 つ目のステップで，1 語から構成されるタグを対象に，似た語義を持つタグをマージする．word2vec[14] によってタグをベクトルに変換し，cos 類似度が閾値以上であるタグのペアをマージする．

最後に，タグ候補の要素を質問のトピックと文字列照合し，質問にタグを付与する．このとき，重複を避けるため最大長のタグを付与する．

3 評価

この節では営業員と顧客の会話の文字起こしテキストから得られた質問に対して，トピック推定とタグ抽出を適用した結果を報告する．質問の件数は 10,334 件であり，これは 2017 年から 2023 年に行われた会話から得られた．言語モデルとして Llama 2 70B[3)]（Llama2）を選び，AWQ[15] によって量子化し，vLLM[16] によって実行した．タグ抽出における係り受け解析器として CaboCha[4)] を使用した．また，ストップワードスコアの閾値を 1.0，マージの際の cos 類似度の閾値を 0.75 に設定した．

3.1 トピック推定

精度 LLM によって推定されたトピックの正確性を評価するために，100 件のトピックと質問のペアをランダムに取得し，人手でその紐付けが正しいか判断した．その結果，正解率 80%であった．ここでは 10,334 件の質問を対象にして推論を 4 度行い，推論結果を全て使用した．1 つの質問に対して異なる複数のトピックが紐づいても正解と判断した．例えば，ある質問に対して，「業務内容を質問する」と「お客様の業務を質問する」のトピックが導出された場合，それらは どちらも正しいトピックと判断する．本来であれば，適合率と再現率の指標を用いて出力の質と量のトレードオフを考察すべきであるが，ある質問から生成され得るトピックを網羅することができず，再現率を計算することが難しい．

私たちは尤度に着目して，質問に対する誤ったトピックを特定しようと試みた．この試みの評価を行うために，10,334 件の質問とトピックのペアの尤度を測り，尤度が上位 25%のペアと下位 25%のペアからそれぞれランダムに 50 件を獲得した．そして，人手でその紐付けが正しいか判断したところ，上位 25%のペアの正解率が 84%であり，下位 25%のペアの正解率が 56%であった．これは尤度と質問とトピックの紐付けの正確さが関連することを示す．

ICL における例の数 ICL における例の数を増やすことで性能が上がることが一般に知られており，本論文では，ICL における例の数と出力されるトピック数の関係を評価した．評価結果を表 2 に示す．出力数とは回答が得られた質問であり，欠損率とは回答が得られなかった質問の割合である．ここでは 10,334 件の質問を対象にして推論を 1 度行い，1 つのプロンプトに 8 つの質問を含ませた．また，トピック数とは出力から重複を取り除いた，ユニークなトピックの数である．

表 2 を見ると，例の数が増えるたびに欠損率とトピック数が減っている．これは，例が増えることで私たちの意図を LLM が理解し，出力されるトピックの表層形が集約されていることを示す．例を 2 つ与えた場合の欠損率は 0.005 であり，8 個の入力にほぼ全て回答している．

計算時間 本論文では，計算の効率を上げるために，1 つのプロンプトに複数の入力を含ませた．そ

3） https://huggingface.co/meta-llama/Llama-2-70b
4） https://taku910.github.io/cabocha/

表 2 In-Context Learning の効果

例の数	出力数	欠損率	トピック数
1	8304	0.196	4437
2	10286	0.005	5785
4	10311	0.002	4339
6	10325	0.001	3506

表 3 トピック推定の計算時間の変化

例数	質問数	入力 トークン数	出力 トークン数	実行 時間
1	1	252.81	30	1.58
6	1	729.81	30	2.39
1	8	797.36	240	10.03
6	8	1274.36	240	11.54

の試みを評価するために，1 つのプロンプトが含む ICL の例と入力の数を変えた場合の実行時間を表 3 に示す．ランダムに選んだ質問によってプロンプトを作成することを 20 回行い，その平均によって入力トークン数を計算した．1 つのトピックのトークン数を 30 として出力トークン数を計算した．

表 3 を見ると，1 つのプロンプトに 6 つの例と 1 つの質問を含む場合の実行時間は 2.39 秒である．また，6 つの例と 8 つの質問の場合に実行時間は 11.54 秒である．この場合に 1 質問あたりの実行時間は 1.44 秒であり，これは 1 つの質問を含む場合の実行時間の 60%である．よって，1 つのプロンプトで複数の入力を使用することで実行時間が短縮した．ミニバッチによって推論を行うことでより実行時間が短くなるが，本論文では報告しない．

3.2 タグ抽出

この節ではタグ抽出の評価を報告する．10,334 件の質問を対象に，Llama 2 によって推論を 4 度行い，その全てを結果の対象にしてタグ抽出を行った．

トピック抽出の効果 営業員と顧客の会話の文字起こしテキストは冗長性や言い間違えを含むため，本論文ではトピック推定を行い，質問を簡素で言い間違えが少ないテキストに変換した．その試みを評価するために，質問とトピックのそれぞれからタグ候補を抽出して比較した．出現頻度が高いタグ候補を表 4 に示す．質問から得られたタグ候補は「いう」「あの」などであり，これらは情報が乏しい．

ストップワード 本論文では，タグは一目で質問の内容を識別できる目印であると考えた．そして，

表 4 質問とトピックから得られたタグ候補上位 5 件

抽出対象	タグ候補
質問	背景，ー，いう，あの．ござる
トピック	質問，確認，お客様，内容，契約

複合名詞や名詞句における出現位置に着目してタグとして的確ではない語を特定し，ストップワードとしてまとめた．ここでは，ストップワードの例を示し，その内容を論じる．

出現頻度が高いストップワードは，「状況」「内容」「方法」である．本論文では単語をタグの候補としており，これらの単語は質問の内容を示していないと考える．そのため，出現頻度が高いストップワードは，本論文の考えに沿っている．

タグのマージ 本論文では，word2vec を使用して似た語義を持つタグをマージした．この試みの効果を論じるために，マージされたタグと同じタグが得られたトピック群の例を 2 つ示す．

1. 【タグ】（スマートフォン-モバイル）（利用-使用）【トピック】「スマートフォンでの利用について質問する」，「スマートフォンの利用状況を質問する」，「モバイルの使用状況を質問する」
2. 【タグ】（利点-メリット-デメリット）【トピック】「利点を質問する」，「デメリットを質問する」，「メリットを質問する」

1 つ目の例では，「スマートフォン」と「モバイル」，「利用」と「使用」がマージされており，3 つのトピック群から同様のタグを抽出できた．2 つ目の例では，「利点」「メリット」「デメリット」がマージされている．「メリット」と「デメリット」が出現する文脈は似ているが，これらは正反対の語義を持つためマージしない方が良いだろう．word2vec によるタグのマージは，対義語を扱えないことがある．

4 まとめ

本論文では，営業員と顧客の会話の文字起こしテキストから得られた質問のタグを推定するタスクに取り組んだ．LLM によって冗長性や言い間違えを含むテキストを簡潔で言い間違えがないテキストに変換した．タグは一目で質問の内容を識別できる目印であると考え，複合名詞や名詞句における出現位置に着目してタグ として的確ではない語を特定した．評価では，実際の営業員と顧客の会話から得た約 1 万件の質問とオープンなモデルであ る Llama2 を用いた．

参考文献

[1] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020.

[2] Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. Efficient intent detection with dual sentence encoders. In **Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020**, mar 2020. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

[3] Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, 2019.

[4] Qi Cao, Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. Unnatural error correction: GPT-4 can almost perfectly handle unnatural scrambled text. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 8898–8913, Singapore, December 2023. Association for Computational Linguistics.

[5] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? GPT-3 can help. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 4195–4205, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] Xingwei He, Zhenghao Lin, Yeyun Gong, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. Annollm: Making large language models to be better crowdsourced annotators. **arXiv preprint arXiv:2303.16854**, 2023.

[7] Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In **International Conference on Machine Learning**, pp. 26837–26867. PMLR, 2023.

[8] Zihan Wang, Tianle Wang, Dheeraj Mekala, and Jingbo Shang. A benchmark on extremely weakly supervised text classification: Reconcile seed matching and prompting approaches. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 3944–3962, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[10] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. **arXiv preprint arXiv:2203.11171**, 2022.

[11] Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. **arXiv preprint arXiv:2310.04408**, 2023.

[12] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. **arXiv preprint arXiv:2301.00234**, 2022.

[13] Chiori Hori, Takaaki Hori, Hideki Isozaki, Eisaku Maeda, Shigeru Katagiri, and Sadaoki Furui. Deriving disambiguous queries in a spoken interactive odqa system. In **2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).**, Vol. 1, pp. I–I. IEEE, 2003.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. **Advances in neural information processing systems**, Vol. 26, , 2013.

[15] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. **arXiv preprint arXiv:2306.00978**, 2023.

[16] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In **Proceedings of the 29th Symposium on Operating Systems Principles**, pp. 611–626, 2023.

A　本研究で用いたプロンプト

本論文でトピック推定を行う際に使用したプロンプトについて解説する．プロンプトは，命令と例と入力の 3 つのパートで構成される．表 5 に各パートの例とそれらを Llama2 に与えた際の出力例を示す．

命令のパートでは，LLM にタスクを説明をする．ここでは，より詳細にタスクを書くことで，出力の精度が向上することが知られている．本論文では，入力が営業に関連することを伝えた．

例のパートは，ICL における例であり，出力の形式や内容を伝えるために入力と出力のペアを示す部分である．例の数を増やすことで性能が上がることが一般に知られており，本論文では例の数と出力されるトピック数の関係を評価した．表 5 には 2 つのペアを含む例を載せた．

入力のパートでは，プロンプトで処理したいテキストを示す．本論文では営業員と顧客の会話の文字起こしテキストから得られた質問が入力に該当する．1 つのプロンプトには，複数の入力を含ませることができ，表 5 には入力される質問が 3 つである場合の入力パートを書いている．

表 5 の出力とは，Llama2 の出力例を示している．適切なプロンプトを書くことで，ユーザの指定したフォーマットに従った出力を得ることができ，この出力例では例のパートと同様のフォーマットに沿った出力が得られた．この出力例は複数の質問のトピックを含んでいるため， 各トピックを取得するためには出力をパースする必要がある．もしも，出力における区切り文字が変更された場合にはパースが失敗する．また，入力が複数ある場合には，全ての入力に対応する出力が得られないことがある．ユーザの意図に沿った出力を得るための方法として，ICL における例の数を増やすことや推論を複数回行うことが挙げられる．

表 5　プロンプトの各パートと出力の例

パート	内容
命令	あなたは営業です。下記の各発言のトピックを推定してください。
例	（質問） 1 部: はい、それではあの早速でございますが、その資料請求いただいた背景を教えていただいてもよろしいでしょうか。 2 部: あっありがとうございます。はーい、その後御社の中での MiiTel のご検討状況はいかがでしょうか （トピック） 1 部: 資料請求の背景を質問する 2 部: 検討状況を質問する
入力	（質問） 1 部: Facebook インスタグラムよりー、あの資料請求いただいた背景と教えていただいてもよろしいでしょうか。 2 部: MiiTel だったり、ほかの IP 電話という所のサービスを活用されていないっていう所で今認識は合ってるんですか。 3 部: 各いただければと思いますはいはい利用料金の部分になりますが、何かご不明な点とかは、ございますか。 （トピック）
出力	1 部: お問い合わせ背景を質問する 2 部: お客様の既存サービスを質問する 3 部: 料金に関する不明点を認知する