LLMを用いたKyutechコーパスのトピック分類

古野 雅人[1]　嶋田 和孝[2]

概要：組織内の意思決定を行う会議では，内容を記録・共有のために議事録が広く用いられている．議事録の自動作成には，発話を話題に応じてトピック単位で整理・分類することが重要である．本研究では，発話単位のトピック分類に取り組む．複数人議論コーパスである Kyutech コーパスを対象に，28 種類のトピックタグに対して，人的コストを抑えたトピック分類の実現を目的とする．LLM を用いることで低コストで柔軟な分類が可能になる一方，多数ラベルの分類を発話のみから行うことは容易ではない．そこで，議論資料を外部知識として参照する RAG と，少数のタグ付き発話例を提示する Few-shot を導入し，その有効性を比較，検証する．

キーワード：トピック分類，マルチラベル分類，大規模言語モデル，議論マイニング

LLM-based Topic Classification on the Kyutech Corpus

Abstract: Meetings are essential, and minutes are widely used to record and share content.To generate meeting minutes, utterances are organized and classified by topic.This study focuses on topic classification for individual utterances.It aims to classify 28 topic tags in a multi-party discussion corpus while minimizing human annotation costs.Although large language models enable flexible classification low-cost, it is difficult to classify many labels based only on utterances.Therefore, this study uses Retrieval-Augmented Generation , which refers to discussion materials as external knowledge.Furthermore, it adopts few-shot learning, which uses a small number of labeled examples to improve classification performance.

Keywords: Topic Classification, Multi-Label Classification, Large Language Models, Argument Mining

1. はじめに

組織内の意思決定や意見形成を行う会議では，内容を記録・共有する手段として議事録が広く用いられている．しかし，人手による議事録の作成には，発言内容の整理や要点の抽出に多大な時間と労力を要するという問題がある．このような背景から，議事録の自動作成技術が注目されている [1].

理解しやすい議論記録を実現するためには，発話を単に時系列で記録するだけでなく，話題ごとに整理することが重要である．会議では複数の参加者が異なる観点から発言するため話題が混在しやすく，議論の流れを把握することが困難になる．したがって，議論構造を明示的に整理する技術が求められる．

このような技術の一例として，西山らは議論の可視化により話題ごとの発言量や参加者の発話状況を把握可能にし，ファシリテーションを支援できることを示している [2] ．また，Chen らは議論のトピックを分析・可視化するフィードバックシステムにより，参加者の議論を支援できることを示している [3] ．このことからも，議論構造の整理は議論の理解や進行を支援する上で不可欠である．特に，発話を話題単位で分類することは，議論の流れを把握するための基盤的な処理であり，要約や可視化など実現する上で重要な要素である．そこで本研究では，議論構造の把握に寄与する基盤技術として，発話単位のトピック分類に着目する．

従来の機械学習に基づくトピック分類手法では，モデルの性能が学習データの量に強く依存するため，高い分類精

[1] 九州工業大学 大学院情報工学府
Department of Creative Informatics, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
[2] 九州工業大学 大学院情報工学研究院 知能情報工学研究系
Department of Artificial Intelligence, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN

度にするためには大量の正解ラベル付きのデータが求められる．しかし，発話単位のアノテーションには多大な人的コストを要する．さらに，会議対話は文脈への依存度が高く，単一の発話のみからトピックを正確に判断することが難しいという問題がある．

近年，大規模言語モデル（Large Language Model：LLM）が自然言語処理分野で注目を集めており，様々なタスク [4,5] に利用されている．LLM は高い文脈理解能力を有し，単一の発話だけでなく前後の発話を入力として扱うことで，文脈を踏まえた推定が可能である．また，追加学習を行うことなく分類タスクへ適用可能であることから，低コストなトピック分類を実現する手段として利用できる．

しかし，LLM を単に適用するだけでは，発話に含まれる情報のみに依存した推定となり，判断に必要な知識が不足する場合がある．特に会議では，議論資料などの事前に共有された情報が参照されることが多く，そこに含まれる知識が発言の解釈に影響を及ぼすことがある．そのため，発話のみを対象とした分類では，トピックの判別が困難になる可能性がある．

そこで本研究では，LLM の文脈理解能力を活かしつつ，発話に加えて利用可能な情報を統合的に活用することで，トピック分類の精度向上を目指す．具体的には，Retrieval-Augmented Generation（RAG）[6] と In-Context Learning（ICL）の 2 つのアプローチを用いる．RAG は外部情報を検索し，LLM に与えることで推論精度の向上を図る手法である．RAG は関連情報を動的に取得し，発話と併せて解釈することで，発話だけでは不足しがちな情報を補完できる．ICL は，LLM に少量の事例を与えることでタスクへの適応を促す手法である．本研究では，これらのアプローチを適用し，トピック分類における有効性を検証する．

2. 関連研究

2.1 機械学習モデルによるトピック分類

トピック分類タスクは，自然言語処理分野における重要なタスクの一つである [7]． 従来の研究では，GRU [8] , BERT [9] など，多様な機械学習モデルを用いたトピック分類手法が提案されている．近年、会話データの特性を考慮したトピック分類手法が提案されている [10] ．川嵜ら [11] は，Kyutech コーパス [12] を対象として複数の機械学習モデルによるトピック分類を行い，ナイーブベイズ分類器が高い分類性能を示すことを報告している．

しかし，これらの従来手法の多くは，発話単体または限定的な文脈情報に基づいて分類を行うため，複数発話にまたがる依存関係や話題の遷移といった会話特有の構造を捉えることが難しい．これらの課題を踏まえ，本研究では長い文脈の理解と柔軟な推論が可能な LLM を用いたトピック分類に取り組む．

2.2 RAG

RAG は，データベースなどの外部の知識ベース（外部情報）から関連情報を検索し，その検索結果を LLM に与えることで，生成や推論の精度を向上させる手法である．RAG では，あらかじめ外部情報を検索し，その検索結果を LLM に与えた上で生成を行うため，LLM の事前学習に含まれない情報であっても活用できるという特徴を持つ．この特性により，RAG は特に与えられた質問に対して、適切な回答を生成する QA タスク（Question Answering Task）において有効であり，外部知識を参照することで回答精度が向上することが報告されている [6, 13] ．さらに近年では，対話タスクを対象として，会話の流れや過去の発話内容を考慮した検索により精度を高めている [14] ．このように RAG は多くのタスクに利用されている [15, 16] ．本研究においても，議論資料を RAG で提供することで，LLM が発話の背景に関する知識を適切に考慮し，より文脈に即したトピック分類や生成を行えるようになると期待される．

2.3 In-Context Learning

In-Context Learning（ICL）は，大規模言語モデル（LLM）に対して少数の入力例とその正解ラベル（Few-shot）をプロンプト内に提示することで，追加学習を行うことなくタスクへの適応を促す学習手法である．LLM は提示された例を文脈として参照し，入力と出力の対応関係や判断基準を推定しながら推論を行うため，柔軟な分類が可能となる．Brown ら [17] は，GPT-3 において少数の例を与えるだけで多様な自然言語処理タスクに対応できることを示し，ICL の有効性を報告している．

一方で，ICL の性能は選択方法に大きく依存するという問題がある．ランダムに選択した例よりも，入力文と意味的に近い例を提示した方が性能が向上することが報告されており，Few-shot 選択の重要性が指摘されている [18] ．

また，ICL の性能は提示する例の質に大きく影響されることが報告されている [19] ．しかし，高品質な例を準備するためには人的コストが伴うという問題がある．そのため，本研究では実データに加えて生成データを Few-shot として活用し，Few-shot 選択手法の違いが本タスクのトピック分類性能に与える影響を検証する．

3. データセット

本節では，本研究で使用したデータセットについて説明する．3.1 節では Kyutech コーパスの概要について説明する．3.2 節では Kyutech コーパスに対して施した前処理について説明する．

3.1 Kyutech コーパス

本研究では，複数人議論コーパスとして Yamamura ら [12] によって作成された Kyutech コーパスを使用す

表 1: Kyutech コーパスの対話の例

参加者	発話	必須	サブ 1	サブ 2
⋮	⋮	⋮	⋮	⋮
D	あんまり今のある店ってちょっと高齢者向け高齢者向けってのがどういうのかがちょっと良く分からんけど	People	Exists	–
D	カフェとかちょっと濃い目のがっつりしたラーメンとかまあ余り高齢者向けではないのかなっていう	People	Exists	–
C	結局高齢者向けだったらもうパスタになりません	People	CandZ	–
⋮	⋮	⋮	⋮	⋮

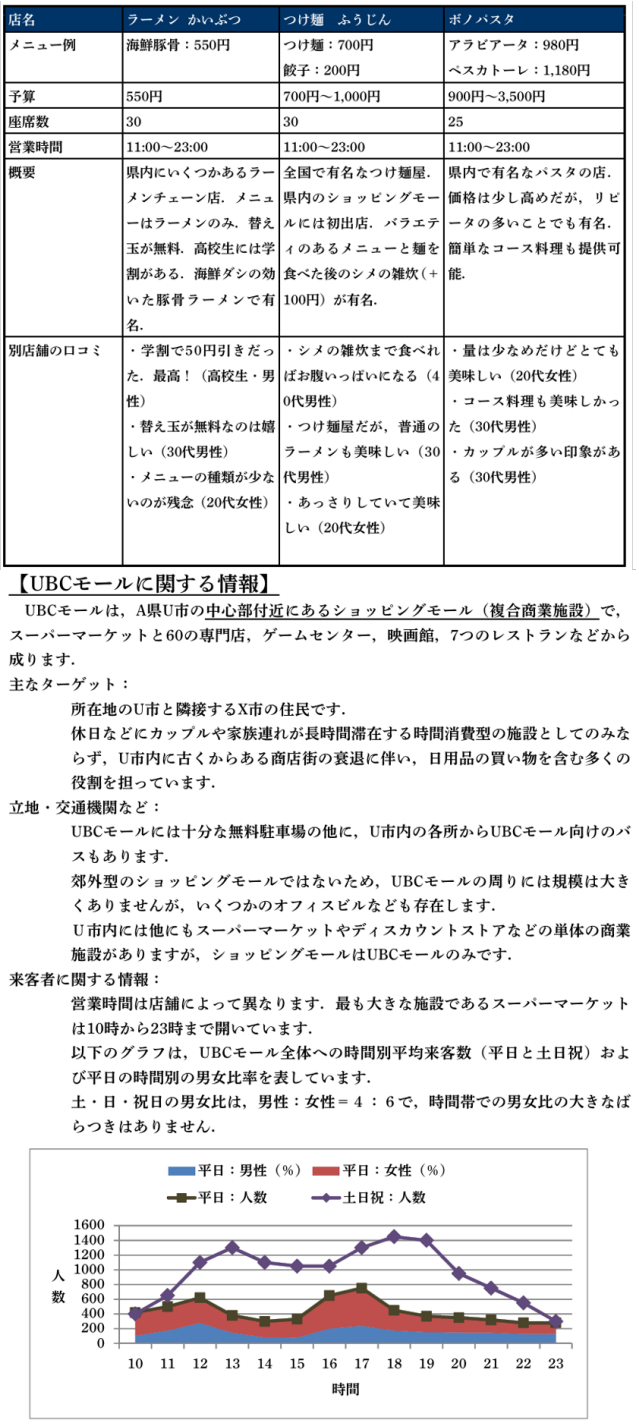
る．Kyutech コーパスは，4 名の参加者（A，B，C，D）による意思決定タスク対話である．参加者は架空のショッピングモールの経営者という役割で，ショッピングモールのレストラン街に新規出店するレストランを 3 件の候補の中から 1 つ選択するというものである．参加者は対話の前にショッピングモールの情報，ショッピングモールの位置する市の人口などの統計情報，隣接する町や市の統計情報，候補店の情報，既存店の情報などが書かれた 10 ページほどの資料が渡される．議論に用いた資料の一部を図 1 に示す．参加者はこの資料を 10 分間黙読した後，20 分の対話を行う．Kyutech コーパスは，4 つの異なる設定（設定 1〜4）に基づいて収集された対話を収録したデータセットである．設定 1 には 3 対話が含まれ，設定 2〜4 にはそれぞれ 2 対話が含まれており，全体で 9 対話から構成される．

Kyutech コーパスの対話の例を表 1 に示す．各発話には人手によるアノテーションで，参加者の話者 ID，発話の開始時間，発話の終了時間，最大 3 個のトピックタグ（必須タグ 1 つ，サブタグ最大 2 つ）が付与されている．必須タグは発話の主要な話題を表すラベルとして必ず付与される一方，サブタグは発話に含まれる副次的な話題や補足的な内容を表す場合に任意で付与される．本研究において，必須タグとサブタグは区別しない．なお，表 1 では発話の開始時間，発話の終了時間の情報は省略している．Kyutech コーパスにおけるトピックタグ一覧を表 2 に示す．Kyutech コーパスで利用される店舗名は，すべて架空の名称である．候補店，既存店，閉店する店舗は設定 1〜4 ごとに異なる．

3.2 Kyutech コーパスの発話文の整形

本研究では川嵜らの手法にならい，Kyutech コーパスに対してノイズとなる発話文を取り除く処理を行ったデータセットを用いる．本節では行われた処理について説明する．

Kyutech コーパスの発話の中には聞き取ることが不可能であった発話を表す「(?)」など，文としての情報を持たない文が存在する．これらを取り除いた結果，計 172 発話のデータが削除され，最終的なデータ数は 3,120 発話となった．Kyutech コーパスのトピックタグごとのデータ数を表 3 に示す．



店名	ラーメン かいぶつ	つけ麺 ふうじん	ボノパスタ
メニュー例	海鮮豚骨：550円	つけ麺：700円 餃子：200円	アラビアータ：980円 ペスカトーレ：1,180円
予算	550円	700円〜1,000円	900円〜3,500円
座席数	30	30	25
営業時間	11:00〜23:00	11:00〜23:00	11:00〜23:00
概要	県内にいくつかあるラーメンチェーン店．メニューはラーメンのみ．替え玉が無料．高校生には学割がある．海鮮ダシの効いた豚骨ラーメンで有名．	全国で有名なつけ麺屋．県内のショッピングモールには初出店．バラエティのあるメニューと麺を食べた後のシメの雑炊（+100円）が有名．	県内で有名なパスタの店．価格は少し高めだが，リピータの多いことでも有名．簡単なコース料理も提供可能．
別店舗の口コミ	・学割で50円引きだった．最高！（高校生・男性） ・替え玉が無料なのは嬉しい（30代男性） ・メニューの種類が少ないのが残念（20代女性）	・シメの雑炊まで食べればお腹いっぱいになる（40代男性） ・つけ麺屋だが，普通のラーメンも美味しい（30代男性） ・あっさりしていて美味しい（20代女性）	・量は少なめだけどとても美味しい（20代女性） ・コース料理も美味しかった（30代男性） ・カップルが多い印象がある（30代男性）

【UBCモールに関する情報】

UBCモールは，A県U市の中心部付近にあるショッピングモール（複合商業施設）で，スーパーマーケットと60の専門店，ゲームセンター，映画館，7つのレストランなどから成ります．

主なターゲット：

所在地のU市と隣接するX市の住民です．
休日などにカップルや家族連れが長時間滞在する時間消費型の施設としてのみならず，U市内に古くからある商店街の衰退に伴い，日用品の買い物を含む多くの役割を担っています．

立地・交通機関など：

UBCモールには十分な無料駐車場の他に，U市内の各所からUBCモール向けのバスもあります．
郊外型のショッピングモールではないため，UBCモールの周りには規模は大きくありませんが，いくつかのオフィスビルなども存在します．
U市内には他にもスーパーマーケットやディスカウントストアなどの単体の商業施設がありますが，ショッピングモールはUBCモールのみです．

来客者に関する情報：

営業時間は店舗によって異なります．最も大きな施設であるスーパーマーケットは10時から23時まで開いています．
以下のグラフは，UBCモール全体への時間別平均来客数（平日と土日祝）および平日の時間別の男女比率を表しています．
土・日・祝日の男女比は，男性：女性＝4：6で，時間帯での男女比の大きなばらつきはありません．

図 1: 議論に用いた資料からの抜粋（設定 1）

本研究では，架空の店名から得られる意味的情報がトピック推定に影響を与えないようにするため，各店舗名（例：かいぶつ）を記号的なラベル（例：CandX）に置換することで一般化を行う．

4. トピック分類手法

本章では，本研究で用いるトピック分類手法について述

表 2: トピックタグの一覧

トピックタグ	タグの説明
CandX	候補店 X についてのトピック
CandY	候補店 Y についてのトピック
CandZ	候補店 Z についてのトピック
CandS	複数の候補店に関連するトピック
Exist1	既存店 1 についてのトピック
Exist2	既存店 2 についてのトピック
Exist3	既存店 3 についてのトピック
Exist4	既存店 4 についてのトピック
Exist5	既存店 5 についてのトピック
Exist6	既存店 6 についてのトピック
Exists	複数の既存店に関連するトピック
Closed	閉店したレストランについてのトピック
ClEx	既存店及び閉店したレストランの両方に関連したトピック
Mall	舞台となるショッピングモール全体についてのトピック
OtherMall	他のショッピングモールについてのトピック
Area	地域や都市についてのトピック
Access	ショッピングモールへのアクセスについてのトピック
Price	価格に対するトピック
Menu	メニューに関するトピック
Atomos	雰囲気に関するトピック
Time	営業時間などに関するトピック
Seat	座席数や回転率などに関するトピック
Sell	売上に関するトピック
Location	お店の立地に関するトピック
People	ターゲットにする顧客についてのトピック
Meeting	話を進めるための議事提案や最終決定部に関するトピック
Chat	雑談 (議題には直接関係のない発話)
Vague	前後の文脈から見ても何を言っているか判別不可能な発話

表 3: トピックタグごとのデータ数

トピックタグ	データ数	トピックタグ	データ数
CandX	325	OtherMall	13
CandY	364	Area	46
CandZ	393	Access	21
CandS	107	Price	82
Exist1	8	Menu	24
Exist2	14	Atomos	0
Exist3	28	Time	8
Exist4	81	Seat	49
Exist5	0	Sell	29
Exist6	45	Location	15
Exists	22	People	787
Closed	98	Meeting	428
ClEx	8	Chat	97
Mall	20	Vague	8

べる．まず，4.1 節で示すプロンプト設計に基づき，LLM 単体によるトピック分類をベースラインとして設定する．続いて，分類精度の向上を目的として，4.2 節では議論資料を検索・参照することで背景知識を補完する RAG を導入する．さらに，4.3 節では，少数のタグ付き発話例を提

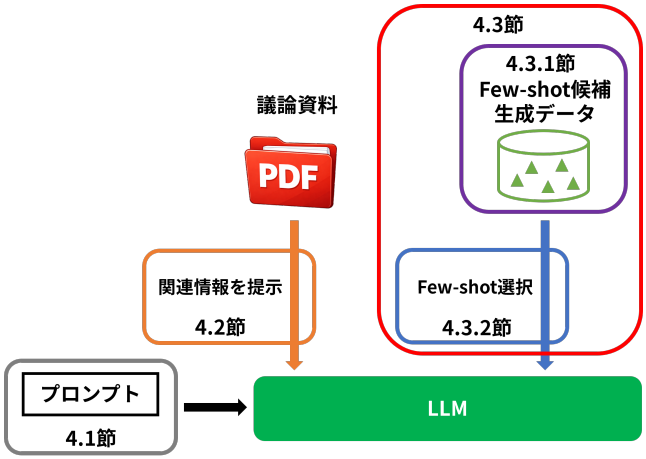


図 2: トピック分類手法の全体構成

示することで判断基準を具体化する Few-shot を適用する．Few-shot 手法では．4.3.1 節で述べる，LLM により生成した対話データを対象とする．また，4.3.2 節で述べる，ランダム選択と意味的類似度に基づく選択の 2 手法を比較する．図 2 に各手法の位置付けを示す．

4.1 プロンプト

本節では，トピック分類を行う際に使用したプロンプトについて説明する．今回 LLM に用いたプロンプトを図 3 に示す．図 3 中の<<utterance>> は推定対象の発話と，その直前の文脈として最大 9 発話を含む発話列を表している．3.1 節で示した通り，店舗を含む 10 個のトピックタグは，設定 1〜4 ごとに対応する店舗名が異なる．そのため，設定ごとのトピックタグの定義（例：CandX）と対応する店舗名（例：ラーメンかいぶつ）を LLM に明示する．そして，入力発話に対して該当すると考えられるトピックタグを 1〜3 個まで出力させる．このプロンプトを用いたトピック分類をベースラインとする．

4.2 RAG を用いたトピック分類

RAG とは，データベースなどの外部の知識ベース（外部情報）から関連情報を検索し，その検索結果を LLM に与えることで，生成や推論の精度を向上させる枠組みである．RAG を用いたトピック分類の概要を図 4 に示す．本研究では，Kyutech コーパスにおける議論資料を外部知識として用い，発話単位のトピック分類に RAG を適用する．外部情報は図 1 に示した Kyutech コーパスの議論の際に利用した資料を用いる．本研究では，OpenAI が提供する RAG の仕組みを用いてトピック分類を行う．

具体的には，各議論設定に対応する資料群を検索可能な形で整理し，分類対象となる発話と資料の類似度に基づいて関連箇所を検索する．得られた検索結果を文脈情報として LLM に付与した上で，当該発話に対するトピックラベルを推定させる．これにより，発話中に店舗名や属性が現

28 種類のトピックラベルから、会話ブロックの最後の発話に当てはまるものを最低 1 つ、最大 3 つ選んでください。
- CandX：候補店 X についてのトピック
- ・・・
- Vague：前後の文脈から見ても何を言っているか判別不可能な発話

以下の対応関係に注意してトピック分類を行ってください。
- 候補店 X：ラーメンかいぶつ
- ・・・
- 閉店したレストラン：定食和屋

以下に、時系列順に並んだ会話の一部を示します。
各行は「発話番号. 話者, 発話開始時間, 発話終了時間, 発話」の形式です。
最後の行に書かれている発話だけを対象としてトピックラベルを決めてください。
先行する発話は文脈としてのみ利用し、ラベルを出力するのは必ず最後の発話だけにしてください。
出力形式は「発話番号: ラベル」の 1 行のみとし、複数のラベルが当てはまる場合は ”/” で区切ってください。

以下が会話ブロックです。
1. A, 00:09.69, 00:18.15, <<utterance>>

図 3: トピック分類に使用したプロンプト

図 4: RAG を用いたトピック分類の概要

れる場合でも，資料に含まれる背景知識を踏まえた判断が可能となり，分類の安定化・精度向上が期待される．

こうした効果を最大化するため，本研究では検索処理に以下の制約を設けた．

- **検索実行の条件**：対象発話の文字数が 5 文字未満の場合は検索を行わない．（これは，極端に短い発話は検索クエリとしての情報量が不足し，無関係な結果を取得しやすくノイズとなる可能性があるためである．）
- **検索結果の付与数**：検索結果は発話との関連度が高い順に最大 3 件まで付与する．（参照情報を過度に増やすとノイズの混入や文脈の希薄化を招く可能性があるため，情報量と精度のバランスを考慮して上限を設定する．）
- **検索結果の利用方法**：検索結果はプロンプト本文へ直接挿入せず，同一リクエスト内で検索と推論を実行する方式を採用する．

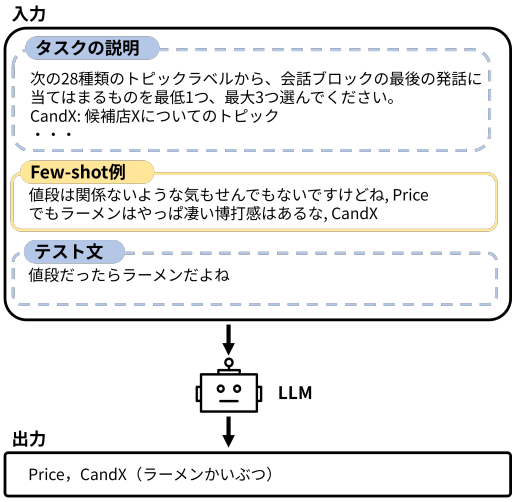


図 5: Few-shot の例

4.3 Few-shot を用いたトピック分類

Few-shot とは，LLM に対して少数の入力例とその正解をあらかじめ提示することで，タスクの形式や判断基準を示す手法である．図 5 は，本研究で用いた Few-shot プロンプトの構成例を示している．まずタスクの説明としてトピック分類の指示を与え，続いてトピックタグが付与された Few-shot を提示する．最後に分類対象となる発話を入力し，LLM は提示された Few-shot を参照しながら該当するトピックタグを出力する．なお，本研究ではプロンプトに挿入する Few-shot の数を 2 件とする．

Few-shot には主に 2 つの問題が存在する．第一に，高品質な例の準備に人的コストを要する点である．実データを用いる場合，人手アノテーションへの依存やラベル分布の偏りが生じる可能性がある．特に発話数が少ないラベルでは多様な例を確保することが難しく，Few-shot の多様性不足が分類性能の低下につながる恐れがある．そこで本研究では，この問題を軽減するため，LLM によって生成した対話を Few-shot 候補データとして活用する．具体的には，各ラベルに対応する対話例を生成し，Few-shot の候補集合を構築する．

第二に，Few-shot の選択方法によって，LLM の分類性能が大きく変化する点である．適切な事例を提示するためには，どの例を選択するかが重要となる．入力文と意味的に近い例を提示することで，より適切な判断基準を与えられることが報告されている [18] ．そこで本研究では，ランダム選択に基づく手法と意味的類似度に基づく選択手法の二つを比較する．これにより，Few-shot の選択戦略がトピック分類性能に与える影響を検証する．

まず，4.3.1 節で，対話データの生成方法について説明する．4.3.2 節では，Few-shot の選択方法について説明する．

4.3.1 対話データの生成

生成時には，4.2 節で用いた RAG と同様に，議論資料

発話番号, 話者, トピックラベル, 発話
1,A,CandX, 今回の候補店の中に『ラーメン CandX』がありますね。県内にいくつかあるラーメンチェーン店だそうです。
2,B,CandX, そうですね。メニューはラーメンのみで、替え玉が無料というのが特徴のようです。高校生には学割もあると聞きました。
3,C,CandX, 価格帯もリーズナブルで、海鮮豚骨ラーメンが550 円というのは魅力的ですね。
4,D,CandX, 営業時間は 11 時から 23 時までで、座席数は 30 席とちょうど良さそうです。
5,A,CandX, 口コミも見てみると、替え玉無料は嬉しいという意見が多いですね。高校生の男の子からも好評のようです。

図 6: 生成した発話の一部

を検索して得られた情報を根拠として LLM に与える枠組みを用いる．ただし，目的が分類ではなく対話生成であるため，検索処理の制約は生成タスクに適した形に調整している．

生成における検索処理の制約を以下に示す．

- **検索結果の付与数**：検索結果は対話生成したいラベルとの関連度が高い順に最大 3 件まで参照する．
- **検索結果の利用方法**：検索結果はプロンプト本文へ直接挿入せず，同一リクエスト内で検索と推論を実行する方式を採用する．（この設定は 4.2 節と同一である．）

議論資料を参照して対話を生成する流れは図 4 と同様である．生成では，各ラベルに対して「対象ラベル」と「ラベル定義」を与え，その範囲内の内容のみからなる対話を生成する．推測に基づく生成を抑制するため，価格・営業時間・席数・地理などの事実情報は検索結果に基づくよう指示を与える．さらに，対象ラベル以外の話題に広げないことを明示し，ラベル境界が崩れにくい対話の生成を目指す．

対話は話者集合 $\{A, B, C, D\}$ を用い，発話数を 20 発話に固定して生成する．生成した発話も 3.2 節と同様に各店舗名を記号的なラベルに置換し，一般化を行う．

実際に CandX（ラーメンかいぶつ）について生成した発話の一部を図 6 に示す．生成した発話は条件に通り出力されており，Few-shot として利用可能な形式で生成されていることを確認した．

4.3.2 Few-shot 選択手法

Few-shot の効果は提示する例の内容だけでなく，どの例を選択するかにも強く依存する．対象発話と関連性の高い例を提示することで適切な推論を促すことが期待される一方，不適切な例を提示した場合には誤った判断基準を形成する可能性がある．

そこで本研究では，Few-shot の選択戦略が分類性能に与える影響を検証するため，ランダム選択と意味的類似性に基づく選択の 2 手法を比較する．

- **ランダム選択**：タグ付き Few-shot 候補データから，各トピックタグごとに発話例を無作為に抽出し，Few-shot としてプロンプトに付与する．なお，全てのテストデータに対して同一の Few-shot を適用する．ランダム選択は特定の基準に依存しないため，選択バイアスを抑えつつ，比較の基準となる選択手法として位置付ける．
- **意味的類似性に基づく選択**：テスト発話と意味的に類似した発話例を Few-shot として選択する手法を用いる．まず，OpenAI の埋め込みモデルを用いて，Few-shot 候補データおよびテストデータをベクトル表現に変換する．次に，テスト発話と候補データの埋め込みベクトル間のコサイン類似度を算出し，類似度が高い上位 2 件を Few-shot として選択する．これにより，LLM は対象発話と意味的に近い Few-shot を参照しながら分類を行うことが可能となる．

5. 分類実験

本章では，4 章で述べた各手法のトピック分類性能を定量的に評価し，結果に基づいて手法間の比較と考察を行う．

5.1 実験設定

本研究では，LLM として GPT-4.1-mini[*1]を用い，生成時の温度パラメータを 0 に固定する．データセットには Kyutech コーパスを用いる．評価は 9 つの議論データを対象に行い，各議論をそれぞれテストデータとして用いる．9 回の分類実験の結果を平均し，評価する．評価指標にはマクロ平均の Precision（Pre），Recall（Rec），F-1 を用いる．Accuracy については，複数ラベルのうち一つでも一致した場合を部分一致（$\mathrm{Acc}_{\mathrm{Partial}}$）として算出する．また，正解ラベルと推定ラベルが完全に一致した場合を完全一致（$\mathrm{Acc}_{\mathrm{Exact}}$）として算出する．

比較手法として，以下の 5 種類を設定する．

- 川嵜らの研究で最高性能を示したナイーブベイズ分類器（NB）を適用した手法
- LLM 単体によるベースライン（BASE）
- 外部資料を検索・参照して背景知識を補完する RAG 手法（RAG）
- 生成データからランダムによって選択した Few-shot を用いる手法（$\mathrm{GEN}_{\mathrm{RND}}$）
- 生成データから意味的類似度によって選択した Few-shot を用いる手法（$\mathrm{GEN}_{\mathrm{SIM}}$）

5.2 実験

各手法の実験結果を表 4 に示す．表 4 より，NB と比較して LLM を用いたすべての手法は大幅に高い分類性能を示した．この結果は，LLM の導入がトピック分類における性能向上に有効であることを示唆している．

[*1] https://platform.openai.com/docs/models/gpt-4.1-mini

表 4: 実験結果

分類手法	Pre	Recall	F-1	$\text{Acc}_{\text{Partial}}$	$\text{Acc}_{\text{Exact}}$
NB	0.177	0.181	0.158	0.561	0.092
BASE	0.544	0.446	**0.442**	0.761	**0.230**
RAG	0.490	**0.477**	0.439	**0.774**	0.223
GEN_{RND}	0.552	0.364	0.400	0.692	0.198
GEN_{SIM}	**0.573**	0.376	0.410	0.676	0.191

表 5: 実データを用いた実験結果

分類手法	Pre	Recall	F-1	$\text{Acc}_{\text{Partial}}$	$\text{Acc}_{\text{Exact}}$
BASE	0.544	0.446	0.442	0.761	0.230
GEN_{SIM}	0.573	0.376	0.410	0.676	0.191
REAL_{SIM}	**0.584**	**0.456**	**0.473**	**0.775**	**0.245**

RAG は $\text{Acc}_{\text{Partial}}$，Recall で最も高い値を示した．これは，議論資料を参照することで，発話のみでは不足しやすい背景知識が補完され，適切なラベルを予測しやすくなったためと考えられる．一方で，RAG は BASE と比べて Precision が低下しており，外部資料に含まれる周辺情報まで参照されることで，不要なラベルまで付与してしまう傾向が生じたためと考えられる．この結果は，検索精度の改善や，提示する文脈量・提示方法の制御が重要であることを示唆する．

生成データを用いた Few-shot（GEN_{RND}，GEN_{SIM}）は，いずれも F-1 および Accuracy で BASE を下回った．また，GEN_{SIM} は Precision が高いものの Recall が伸びておらず，分類に必要な判断材料を十分に与えられていない可能性がある．生成対話は一定の多様性を持つものの，実際の議論に見られる言い回しや話題遷移を十分に再現できていない場合，Few-shot が有効な判断基準として機能しにくいと考えられる．

5.3 考察

生成データを用いた Few-shot は，人的コストを抑えられるという利点がある．しかし，その分類性能がどの程度まで到達可能であるかは明らかではない．そこで本節では，人手による正解データが利用できる場合，Few-shot 手法はどこまで精度向上するか（Few-shot 手法の上限精度）を確認する．具体的には，実データを用いて意味的類似度に基づき例を選択した Few-shot 手法（REAL_{SIM}）の精度を測る．なお，実データの Few-shot 候補が 2 件未満のラベルについては，取得可能な例のみを使用する．

実データを用いた結果を表 5 に示す．表 5 より，REAL_{SIM} はすべての評価指標で最も高い性能を示した．この結果は，Few-shot の品質が分類性能を左右する重要な要因であることを示している．したがって，生成データが実データに近い内容を生成することができれば，人的コストを抑えつつ，より高精度な分類ができる可能性がある．

6. おわりに

本研究では，Kyutech コーパスを対象として，人的コストを抑えたトピック分類の実現を目的とし，LLM を用いた分類手法を検討した．外部資料を参照する RAG と，判断基準を具体化する Few-shot を導入し，それぞれの有効性を比較評価した．

実験の結果，LLM を用いた手法はいずれもナイーブベイズ分類器より高い分類性能を示した．このことから，LLM の適用がトピック分類に有効である可能性が示唆された．RAG は人手によるアノテーションを必要とせずに比較的高い性能を示しており，議論資料を参照することで発話のみでは不足しがちな背景知識を補完できる可能性が示された．生成データを用いた Few-shot はベースラインを上回る結果には至らず，生成対話が実際の議論に見られる表現や話題遷移を十分に反映できていない可能性が考えられる．

また，実データを用いて意味的類似度に基づき例を選択した Few-shot は最も高い性能を示した．この結果は，提示する事例の質が分類性能に強く影響することを示している．したがって，実データに近い内容を生成ができれば，人的コストを抑えつつ，より高精度な分類ができる可能性がある．

今後は，検索精度の向上や提示コンテキストの最適化による RAG の改善に加え，実データの特徴をより反映した Few-shot 生成手法の検討を進めることで，低コストかつ高精度なトピック分類の実現を目指す．

謝辞

本研究は科研費 23K11368 の一部です．

参考文献

[1] Jia Jin Koay, Alexander Roustai, Xiaojin Dai, and Fei Liu. A sliding-window approach to automatic creation of meeting minutes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 68–75, 2021.

[2] 西山空良, 嶋田和孝. 議論の分析とファシリテーションのための可視化ツールの構築. HCG シンポジウム 2021, pp. I–2–2. 電子情報通信学会, 2021.

[3] Chih-Ming Chen, Ming-Chaun Li, Wen-Chien Chang, and Xian-Xu Chen. Developing a topic analysis instant feedback system to facilitate asynchronous online discussion effectiveness. *Computers & Education*, Vol. 163, p. 104095, 2021.

[4] Koki Imazato and Kazutaka Shimada. Automatic few-shot selection on in-context learning for aspect term extraction. *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 15–20, 2024.

[5] Keisuke Iwamoto and Kazutaka Shimada. Exaggeration scoring of news summaries through llm-based relative judgments. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, 2025.

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, Vol. 33, pp. 9459–9474, 2020.

[7] Yujia Wu and Jun Wan. A survey of text classification based on pre-trained language model. *Neurocomputing*, Vol. 616, p. 128921, 2025.

[8] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, 2016.

[9] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on eu legislation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6314–6322, 2019.

[10] Zhengyuan Liu, Siti Umairah Md Salleh, Hong Choon Oh, Pavitra Krishnaswamy, and Nancy Chen. Joint dialogue topic segmentation and categorization: A case study on clinical spoken conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 185–193, 2023.

[11] 川嵜慎乃介, 嶋田和孝. 機械学習モデルを用いた kyutech コーパスのトピック分類. 電子情報通信学会, 信学技報, Vol. 122, No. 99, pp. 13–18, 2022. NCL2022-3.

[12] Takashi Yamamura, Kazutaka Shimada, and Shintaro Kawahara. The Kyutech corpus and topic segmentation using a combined method. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 95–104, 2016.

[13] Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. REAR: A relevance-aware retrieval-augmented framework for open-domain question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5613–5626, Miami, Florida, USA, 2024.

[14] Zhiyu Chen, Biancen Xie, Sidarth Srinivasan, Manikandarajan Ramanathan, Rajashekar Maragoud, and Qun Liu. LLM-based dialogue labeling for multiturn adaptive RAG. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1044–1056, 2025.

[15] Yusei Fukata and Kazutaka Shimada. Augmented data generation and selection for aspect estimation. In *Proceedings of the 20th International Conference on E-Service and Knowledge Management, 2025.*, 2025.

[16] Joshua J. Woo, Andrew J. Yang, Reena J. Olsen, Sayyida S. Hasan, Danyal H. Nawabi, Benedict U. Nwachukwu, Riley J. Williams, and Prem N. Ramkumar. Custom large language models improve accuracy: Comparing retrieval augmented generation and artificial intelligence agents to noncustom models for evidence-based medicine. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, Vol. 41, pp. 565–573, 2025.

[17] 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.

[18] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, 2022.

[19] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pp. 39818–39833, 2023.