アウトプット型学習のための主張抽出とRAGに基づく訂正・補足情報の生成

米村　琉衣†　　中井香那子††　　山本　岳洋††

† 兵庫県立大学 社会情報科学部　〒651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 兵庫県立大学 大学院情報科学研究科　〒651–2197 兵庫県神戸市西区学園西町 8–2–1
E-mail: †{fa22p101,ad24c041}@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本研究では，議論の発話内容から話者の主張の抽出と，その主張の中に誤った情報が含まれている場合は訂正文を，不足情報がある場合は補足文の生成に取り組む．学生が行うアウトプット型学習を目的としたプレゼンテーションにおける発話には，誤った理解や不足した背景情報を含む主張が見られることがあり，これらは学生の学習の妨げの要因となる．提案システムでは，発表音声から大規模言語モデル（LLM）を用いて発表者および質問者の主張を抽出する．また，その後発表の根拠となる文書を外部データとした検索拡張生成（RAG）を用いて訂正文と補足文を生成する．実験のため，大学生または大学院生 11 名を対象に実施した論文紹介のプレゼンテーションを用いてデータを収集した．収集したデータに対して音声認識モデルの違いによる主張抽出精度の差と訂正文および補足文の生成精度を，適合率，再現率，F_1 値を用いて評価した．その結果，主張抽出の精度は音声認識の精度に影響されることを確認した．また，提案システムは主張抽出を行わない場合と比較して，訂正文および補足文の生成において，適合率，再現率，F_1 値のいずれも向上することを示した．

キーワード　アウトプット型学習，RAG，主張抽出

1　はじめに

近年の中等教育および高等教育においては，学習者が主体的に学習活動に参加することで，知識の理解を深め，思考力や表現力を育成することを目的とした学習方法であるアクティブ・ラーニングが盛んに行われている [1] [2]．具体的には，授業で習った内容を生徒が各自でまとめて他の生徒に発表するプレゼンテーションなどのアウトプット型学習がある．このような学習方法は授業を聞き知識を身に着けるインプット型学習よりも学習者の知識の獲得に役立つと考えられる．

しかし，アウトプット型学習には課題も存在する．特に，指導教員などの学習を支援する立場にある者が存在しない場合，あるいは指導教員が生徒の発表分野に関する知識を十分に有していない場合，学習者が発表した知識が間違っていたり，議論の最中に突発的に発言をしてしまうことがあると考えられる．このような学習者が突発的に発言する場面では，学習者が自律的に調べた情報の正確性を即座に判断することは困難であり，誤った情報や十分ではない情報を共有してしまうリスクがある．さらに，このような誤情報は発表を聞いている他の学習者にも共有されてしまう可能性があり，学習内容全体の理解に悪影響を及ぼす恐れがある．

そこで本研究では，アウトプット型学習を支援することを目的として，発表および質疑応答の音声から外部文書を用いた検索拡張生成（RAG）に基づいて訂正文および補足文を生成するシステムを提案する．

本システムの実現にあたり，実際の発話には曖昧な表現や言い直しが多く含まれるため，発話全体を対象として正確な訂正または補足を生成することは困難であると考えられる．そこで発話の中から話者が伝えようとしている主張を抽出することにより，訂正文および補足文生成の精度向上を目指す．本システムにより，学習者が共有した情報の誤りの訂正や不足情報を補足することで，より正確で理解の深い学習環境の実現を目指す．

例えば，大学の研究室にて情報検索における nDCG という評価指標についてプレゼンテーション形式で発表を行い，学生のみで学習した情報を共有したとする．その際，図 1 のように発表者が誤った情報を発信した際に，「nDCG の算出式における対数の底を変更するとスコアが変化する.」という主張を抽出した後，誤りの認識と訂正を行い，「発表で用いている nDCG の算出式の対数の底を変化させると DCG のスコアは変わりますが nDCG のスコアは変化しません.」という訂正文や，「nDCG の定義は 2 種類ある」という補足文を画面に返すシステムを実現する．

提案システムでは，まず発表および質疑応答の音声に対して音声認識を行い，得られた文字起こしテキストから大規模言語モデル（LLM）を用いて話者の主張を抽出する．次に，抽出された主張を入力として，発表の根拠となる文書を外部データとした RAG を用い，主張に誤りが含まれる場合には訂正文を，不足情報がある場合には補足文を生成する．

本研究では，研究室内で実施された論文紹介プレゼンテーションを対象として実験を行い，発話内容からの主張抽出精度と主張抽出の有無が訂正文および補足文生成の精度に与える影響を評価した．実験の結果，主張抽出の精度は音声認識の精度

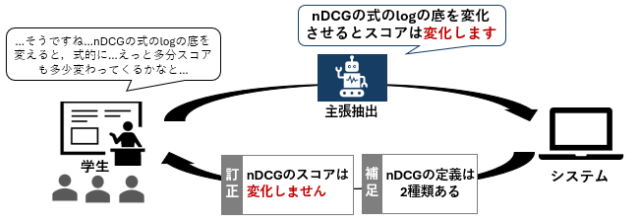


図 1　提案システムの概要.

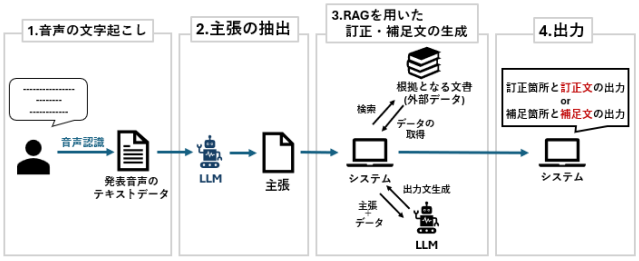


図 2　提案システムの流れ.

に影響されることを確認した．また，主張抽出を導入することで，訂正文および補足文の生成精度が向上することを確認した．

2　関 連 研 究

2.1　LLM を用いた議論支援

LLM を用いた議論支援システムは，多く提案されている．Imamura らは，ブレインストーミング中の議論の書き起こしから埋め込みベクトルを生成し，関連論文を提示する Serendipity Wall を提案している [3]．木下らは，市民参加型 Web 議論を対象に，情報提示の必要性を判定したうえで，LLM による関連情報の検索・要約とファクトチェックを行う情報推薦システムを提案している [4]．

教育分野では，Steinert らが，教育学的理論に基づくプロンプトを用いて学習者の回答に形成的フィードバックを生成する学習支援システム LEAP を提案している [5]．また，Kirstein らは，LLM と RAG を組み合わせ，議事録と補足資料を統合する要約パイプラインを提案し，要約の質向上を示している [6]．加えて，Yun らは，教育理論に基づく専門家ルールベースと LLM を統合したハイブリッド AI エージェントを開発し，教室内の複雑な対話シーケンスを自動分析する手法を提案している [7]．

さらに，創造的協調作業を支援する研究として，Shi らの IdeaWall [8]，Andolina らの InspirationWall [9]，中らの AIR-VAS [10]，Fede らの The Idea Machine [11]，および IdeaExpander [12] が提案されており，議論内容に関連する情報や刺激を提示することで参加者を支援している．

しかしながら，教育分野において，議論の発話内容に基づいて訂正情報や補足情報を提示することを目的とした議論支援システムの提案は少ない．

2.2　LLM を用いた主張抽出やファクトチェック

発話や文章から主張を抽出する研究や，ファクトチェックに関する研究が行われている．

Panchendrarajan らは，自動ファクトチェックにおける主張検出を対象に，単言語・多言語・クロスリンガル設定における既存手法を，検証可能性や優先度などの観点から体系的に整理している [13]．Ullrich らは，文脈付きテキストから事実確認対象となる主張を生成的に抽出する手法を整理・比較し，主張抽出の品質を評価する自動評価指標 Ffact を提案している [14]．また，Tran らは教室内ディスカッションを対象に，LLM を用いて主張構造のエンドツーエンド抽出を行い，少量のアノテーションでも高い性能を示している [15]．

ファクトチェック支援に関して，Gupta らは日常会話中の誤情報をリアルタイムに検知し，ウェアラブル端末による非言語的フィードバックを行うシステム Factually を提案している [16]．さらに，Venktes らはライブ音声ストリームを対象に，発言の書き起こしと話者同定を行い，主張の真偽を数秒以内に検証するシステム LIVEFC を提案している [17]．

一方で，Rashkin らが示すように，既存研究の多くは書き言葉を中心に，真偽判断に有効な言語的特徴の分析に焦点を当てている [18]．そのため，人間の発言，特に発表や議論といった口語的な発話から主張を抽出し，その内容に対して体系的にファクトチェックを行う研究は著者らの知る限り少ない．また，発話内容は冗長表現や言い直しを含むことが多く，発話全体から正確な訂正情報や補足情報を生成することは困難であるため，正確な訂正情報や補足情報を提示するには，発話から話者の主張を正しく抽出する必要がある．本研究は，この点に着目し，議論に対する訂正および補足を行う新たな支援手法を提案するものである．

3　主張抽出に基づく訂正文および補足文の生成

3.1　システムの流れ

本研究ではアウトプット型学習を支援するために，発言に対する訂正と補足を行うシステムを提案する．提案システムの流れを図 2 に示す．まず，発表や質疑の音声の文字起こしを行い，文字起こしテキストから LLM を用いて主張の抽出を行う．その後，得られた主張を入力として，RAG より発表内容における補足が必要な箇所や訂正が必要な箇所を自動的に検出し，対応する訂正文および補足文を生成して画面上に出力する．

3.2　音声認識による発話内容のテキスト化

提案システムにおける主張抽出と訂正文および補足文の生成の入力として用いるため，発話音声に対して音声認識を行い，発話内容をテキスト化する．音声は，話者が一つの発話を話し終えたタイミングで音声ファイルを区切るために，無音区間に基づいて分割した．具体的には，音量が 45dB 以下の状態が 2 秒以上継続した区間を無音区間と定義し，音声ファイルを無音区間ごとに wav ファイルとして分割した．その後，分割された各音声ファイルに対して音声認識モデルを適用し文字起こしを行う．

3.3 発話内容からの主張抽出

発表内容に含まれる誤りや不足情報に対して適切な訂正文および補足文を生成するためには，発話全体から検証対象となる情報を特定する必要があると考えられる．そのため，発話内容から話者が提示している重要な説明や断定的な内容を抽出し，訂正文および補足文生成の対象となる単位を明確にする．

無音区間ごとに分割された発表音声から得た文字起こしデータに含まれる話者の主張を自動で取り出すために，LLM を使用する．本研究における主張とは話者が発表の中で「伝えたい中心的な内容」や「説明したいポイント」をまとめた文のことであり，訂正文および補足文生成における基本的な単位となる．

例えば，nDCG には複数の定義が存在するが，そのうち一般的に用いられている Burges らによる nDCG の定義 [19] について発表が行われたとする．その際，「nDCG の算出式における対数の底を変更するとスコアが変化する」といった性質に関する主張や，「nDCG では対数の底は定義によると後ほどキャンセルされるのでなんでもよい」といった主張などを，発話内容から取り出すことを想定している．

以下は, 実際に LLM に与えたプロンプトであり，以下のプロンプトにおいて文字起こしデータは変数 transcription に格納されている．

主張抽出のプロンプト

あなたは発表会の記録を整理する専門アシスタントです．
以下は，発表者と質問者の途中の会話を文字起こししたテキストです．
このテキストから，実際に述べた要点だけを抽出してください．
推測や補完は禁止です．話されていないことを想像して書かないでください．

【出力形式】
- 発話者が説明・主張・条件・前提・結果・評価・意見・補足として述べた内容を箇条書きにして下さい．
- 箇条書きの結果のみを出力してください．

【制約条件】
- 実際の発話に存在しない情報を補完・推測しないこと．
- 言い換えはしてもよいが，意味を変えない．
- 発表者や質問者など役割ごとにまとめることはせず一括で出力すること．
-可能な限り多く，重複しない主張を抽出すること．

=== 文字起こし ===
transcription

主張抽出プロンプトは，発話内容から実際に述べられた要点のみを安定して抽出するため以下の特徴を持つように設計した．

- **推測と補完の禁止**
  「推測や補完は禁止」と明示することによって，入力テキストに存在しない情報が LLM の憶測によって出力されることを防ぐ．
- **意味の保持を重視した言い換えの許可**
  意味を変えない範囲での言い換えを許可することによって，発話の冗長性を減らし，要点を簡潔に表現できるようにする．
- **役割ごとに分けない一括出力**
  発表者や質問者といった役割で分けずに一括で出力させることで，質問文のみを主張として抽出することを防ぎ，発話全体の文脈に基づかない不必要な補足の生成を減らす．

例えば，「提示する応答っていうのがそのまま表示するものが 1 つで事実性が低い情報にハイライトしたものが 1 つで事実性が高い情報にハイライトしたものが 1 つでその提案手法である事実性が低い情報を隠したものっていうのが 1 つで最後事実性が低い情報っていうのを曖昧な表現に変更したものっていうのを 1 つで用意しています」という発話から「提示する応答は、そのまま表示するもの、事実性が低いものまたは高いものにハイライトしたもの，隠したもの，曖昧な表現に変えたものである」という主張を抽出する．

3.4 主張に対する関連文章の取得

主張に対する関連文章を取得するために根拠となる文書を埋め込みベクトルに変換し，RAG を用いる際に検索できるようにする．RAG に用いる外部データとして今回の研究では論文の PDF ファイルを対象とした．まず，PDF を読み込みテキストチャンクに分割する．チャンク分割を行うことで，主張内容と対応する局所的な文脈を含む記述を効率的に検索でき，発表内容を根拠となる文書の記述に基づいて検証したうえで，訂正文および補足文を生成することが可能となる．次に，得られたテキストチャンクに対して，埋め込みモデルを用いて埋め込みを行う．各チャンクは独立してベクトル化され，論文内の内容を表現する検索用ベクトルとして保存される．検索時には，入力となる主張文の埋め込みベクトルとのコサイン類似度を用いて各チャンクとの適合度を算出し，適合度の高い上位 k 件のチャンクを取得する．

3.5 訂正文および補足文の生成

主張抽出によって得られた主張文に対して，発表内容の誤りや情報不足を明らかにするため RAG を用いて訂正文および補足文の生成を行う．例えば，「nDCG の算出式における対数の底を変更するとスコアが変化する.」という主張に対し，nDCG の算出式の対数の底を変化させても nDCG のスコアは変化しないという点を示した訂正文を提示する．また，nDCG の定義は 2 種類存在することなどを補足文として示すことを想定している．

本研究では，訂正と補足を同一の処理として扱わず，訂正用および補足用に別々のプロンプトを設計し，個別に処理を行う．これは，誤りの修正と情報の追加では判断基準や出力内容が異なるため，個別のプロンプトで生成した方が訂正文および補足文生成の精度が高くなると考えたからである．

以下は訂正文および補足文の生成のために与えたプロンプトである．以下のプロンプトにおいて生成された主張は変数 claim に，埋め込みベクトルに変換された外部データは変数 context に格納されている．

訂正文生成のプロンプト

対象テキスト:
claim
参照可能な外部情報:
context
このテキストは論文紹介のプレゼンテーションの内容を音声から書き起こし，主張をまとめたものです．
発表者の主張が外部データである論文の情報と明らかに異なっている場合や，一般的な知識や常識と異なる場合訂正が必要な箇所とみなし，以下を簡潔に示してください：

・誤っている箇所の抜粋
・誤りの理由
・正しい情報

【制約条件】
-文章が異なっていても意味が同じ場合は訂正箇所とはみなさないでください．
-誤字がある場合でも意味が通じる場合は訂正箇所とはみなさないでください．
-訂正が不要な場合は「(何も出力しない)」で返してください．
-可能な限り多く，重複しない訂正を抽出すること．

補足文生成のプロンプト

対象テキスト:
claim
参照可能な外部情報:
context
このテキストは論文紹介のプレゼンテーションの内容を音声から書き起こし，主張をまとめたものです．
外部データの論文を紹介するにあたって重要な前提知識・定義・根拠・背景説明が欠けている場合，以下を含む補足情報を簡潔に示してください：

・補足が必要な箇所の抜粋
・補足情報

【制約条件】
-補足を行う際，誤った情報の訂正は行わないでください．
-補足が不要な場合は「(何も出力しない)」で返してください．
-可能な限り多く，重複しない訂正を抽出すること．

訂正文の生成では，抽出された主張が外部データである論文の内容，あるいは一般的な知識や常識と明らかに異なる場合に，訂正が必要な箇所として検出する．プロンプトでは，誤っている箇所の抜粋，誤りの理由，および正しい情報を簡潔に出力するよう指示する．一方で表現が異なっていても意味が同じ場合は訂正対象としない．もし訂正が不要な場合には何も出力しないように指示している．

補足文の生成では，主張自体に誤りはないものの論文を理解する上で重要な前提知識，定義，根拠，背景説明が不足している場合に補足文を生成する．プロンプトでは，補足が必要な箇所の抜粋と対応する補足文を簡潔に示すよう指示する．また補足生成では誤った情報の訂正は行わず，訂正文の生成の際と同様に補足が不要な場合には何も出力しないものとする．

4 実　　験

本研究では，提案手法の有用性を検証するため，主張抽出の精度評価と，主張抽出を行うことによって訂正文および補足文の生成精度が向上するかどうかを評価した．

4.1 実験に用いたデータ

研究室内で行われた発表を対象にデータを収集し，実験に用いた．具体的には，著者らが所属する兵庫県立大学山本研究室に所属している大学生または大学院生 11 名を対象とし，各自が選んだ論文に関する発表を実施した．実験の際，すべての参加者から研究目的およびデータの利用に関する説明を事前に行い，同意書への署名を得た．データの収集は，2025 年 11 月 7 日に実施した．

発表者には，質疑応答を含めて約 10 分間で発表を行うよう求め，その際に作成したスライドを基に内容を説明してもらった．この発表において，発表および質疑応答における音声データ，スライド映像，ならびに発表資料（スライド）と各自が選んだ対象論文を取得した．実験では，収集した発表および質疑応答の音声データから音声を文字起こししたテキストを入力として，提案手法による訂正文および補足文の生成を行った．

4.2 実 験 設 定

本実験では，3 節で述べた提案システムを用いて，主張抽出と訂正文および補足文生成の性能を評価した．

音声認識モデル: 音声認識モデルには，OpenAI が提供する Whisper（API）[1]およびローカル環境で動作する Whisper（Local）[2]（Multilingual model，medium）の 2 種類を使用した．Whisper（API）は，著者らが試した限り高い文字起こし精度を示した一方で，API を利用するためコストが発生する．一方，Whisper（Local）はローカル環境で動作するため利用コストはかからないが，文字起こし精度は Whisper（API）と比較して低い傾向がある．このような精度とコストのトレードオフを考慮し，本研究では両者を

1：https://openai.com/ja-JP/index/introducing-chatgpt-and-whisper-apis/
2：https://github.com/openai/whisper

比較対象として採用した．

主張抽出のモデル: 主張抽出には，Google が開発したローカルで動作する LLM である gemma3:12b [20] を使用した．

埋め込みモデル: RAG における外部データの埋め込みの際，PDF からのテキスト抽出には Python ライブラリ fitz を用いた．抽出したテキストは，文脈のまとまりを考慮しつつ 200 字ごとに分割し，各テキストを 1 チャンクとして扱った．各チャンクは Google による埋め込みモデルであり多言語に特化している embeddinggemma:300m[3]を用いて埋め込みベクトルに変換し，検索用データベースとして保存した．検索時には，入力となる主張文を同様に埋め込みベクトルに変換し，各チャンクとのコサイン類似度に基づいて適合度を算出し，上位 5 件のチャンクを取得した．

訂正文および補足文生成のモデル: 訂正文および補足文の生成には，gemma3:12b と OpenAI の GPT-4o-2024-11-20（GPT-4o）[4]の 2 つのモデルを使用した．GPT-4o は，文脈を安定して理解でき，訂正文および補足文生成においてプロンプトにより忠実に出力を生成できると考えたため，本研究では GPT-4o を採用した．一方で，gemma3:12b はローカル環境で動作可能なモデルであり，外部 API を利用せずに運用できるという利点を有する．本研究では，提案手法が特定のモデルに依存せず有効に機能するかを検証するため，性能特性の異なるこれら 2 種類のモデルを用いて比較を行った．なお，生成結果の再現性を確保するため，すべての生成において temperature は 0 に統一した．

4.3 評価指標

4.3.1 主張抽出の評価指標

主張抽出の性能を評価するため，人手により作成した理想的な主張の集合と，システムによって抽出された主張の集合を用いて評価を行った．まず，発話内容から抽出されるべき m 個の理想的な主張の集合を $A = \{a_1, a_2, \ldots, a_m\}$ と定義する．ここで，各要素 a_i（$1 \leqq i \leqq m$）は，人手により作成された主張の 1 単位を表す．なお，理想的な主張の集合は，著者が作成したものである．次に，実際に抽出された n 個の主張の集合を $B = \{b_1, b_2, \ldots, b_n\}$ と定義する．ここで，各要素 b_j（$1 \leqq j \leqq n$）は，システムが出力した主張の 1 単位を表す．集合 A と B の要素間で，意味的に同一であると人手で判断された対応関係の集合を $A \cap B$ とする．なお，文面が完全に一致しない場合であっても，意味が同一であると判断できる場合は一致とみなす．この一致判定は，理想的な主張の集合とシステム出力の集合が対応しているかどうかを著者が人手で判断することにより行った．

このとき，主張抽出における適合率（Precision），再現率（Recall），および F_1 値をそれぞれ以下の式で求める．

$$\text{Precision} = \frac{|A \cap B|}{|B|}$$

$$\text{Recall} = \frac{|A \cap B|}{|A|}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.3.2 訂正および補足文生成の評価指標

訂正文および補足文生成の性能を評価するため，主張抽出と同様に，人手で作成した理想的な出力例と，システムによる生成結果を比較する．訂正文と補足文は異なるタスクであるが，同一の評価方法を用いた．まず，理想的な o 個の訂正文または補足文の集合を $C = \{c_1, c_2, \ldots, c_o\}$ と定義する．ここで，各要素 c_k（$1 \leqq k \leqq o$）は，人手により作成された訂正文または補足文の 1 単位を表す．なお，理想的な訂正文および補足文の集合は，著者が作成したものである．次に，提案手法または比較手法によって生成された p 個の訂正文または補足文の集合を $D = \{d_1, d_2, \ldots, d_p\}$ と定義する．ここで，各要素 d_l（$1 \leqq l \leqq p$）は，システムが生成した訂正文または補足文の 1 単位を表す．集合 C と D の要素間で，意味的に同一であると人手で判断された対応関係の集合を $C \cap D$ とする．文面が完全に一致しない場合であっても，訂正内容や補足内容の意味が同一であると判断できる場合は一致とみなす．この一致判定は，理想的な訂正文および補足文の集合とシステム出力の集合が対応しているかどうかを著者と指導教員が人手で判断することにより行った．このとき，訂正文および補足文生成における適合率，再現率，および F_1 値を 4.3.1 節と同様に求めた．

4.4 比較手法

実験では，提案手法の有用性を検証するため，主張抽出を行うことで訂正文および補足文生成の精度が向上するかという点と，使用するモデルの違いが生成精度に与える影響という 2 点に着目し，主張抽出と訂正文および補足文生成の各段階において条件を整理した比較を行った．

4.4.1 主張抽出の比較手法

主張抽出の比較手法として，音声認識手法の違いによる影響を検証するため，Whisper（API）と Whisper（Local）の 2 種類の音声認識結果を用いて主張抽出を行い，それぞれの抽出精度を比較した．この際，主張抽出に用いる LLM やプロンプトは同一の条件とした．

4.4.2 訂正および補足文生成の比較手法

訂正文および補足文生成の比較では，以下の手法を対象とした．

はじめに，主張抽出の有無による比較を行った．一つ目は，Whisper（API）および Whisper（Local）によって文字起こしされたテキストから主張抽出を行い，得られた主張を入力として訂正文および補足文を生成する手法である．二つ目は，主張抽出を行わず，音声を文字起こししたテキストを直接入力として訂正文および補足文を生成する手法である．主張抽出を行わない手法は，発話内容全体をそのまま入力として生成を行う一般的な方法に相当するため，提案手法の有効性を検証する比較対象として適切であると考えられる．この比較により，発話内容から主張を抽出することが，訂正文および補足文生成の精度向上に寄与するかを明らかにする．

3：https://huggingface.co/google/embeddinggemma-300m

4：https://platform.openai.com/docs/models/gpt-4o?snapshot=gpt-4o-2024-11-20

表 1　主張抽出の評価結果.

音声認識のモデル	適合率	再現率	F_1 値
Whisper（Local）	0.55	0.60	0.56
Whisper（API）	0.65	0.71	0.67

次に，使用するモデルの違いが生成精度に与える影響を検証する．訂正文および補足文生成に用いる LLM として GPT-4o および gemma3:12b を使用し，同一条件下で比較を行った．これにより，提案手法の性能が特定のモデルに依存するか，あるいは複数のモデルにおいて一貫した傾向が得られるかを分析する．

また，これらの比較においては，使用する外部データや RAG の設定など，主張抽出およびモデル以外の条件を可能な限り統一することで，各要因が生成精度に与える影響を個別に評価できるようにした．

4.5 実験結果

提案手法による実験結果について述べる．主張抽出の精度評価には，研究室内で実施した論文紹介プレゼンテーションに参加した 11 名分のデータを用いた．一方で，訂正文および補足文の生成精度の評価については，評価データの作成に詳細な評価が必要なことから，2 名分のデータを対象として評価を行った．

4.5.1　主張抽出の精度の結果

表 1 に，主張抽出の評価結果を示す．表中の値は 11 名分データそれぞれについて主張抽出の評価を行いマクロ平均をとったものである．Whisper（API）を用いた場合，適合率 0.65，再現率 0.71，F_1 値 0.67 となり，Whisper（Local）を用いた場合の適合率 0.55，再現率 0.60，F_1 値 0.56 と比較して，すべての指標において高い値を示した．

4.5.2　主張抽出の有無による訂正および補足生成の結果

表 2 および表 3 に，訂正文および補足文生成の評価結果を示す．表中の値は 2 名分データそれぞれについて訂正文および補足文生成の評価を行いマクロ平均をとったものである．

まず訂正文生成について見ると，主張抽出を行わずに文字起こしテキストを直接入力した場合はモデルを問わず適合率，再現率，F_1 値はいずれも 0.00 となり，有効な訂正文を生成できなかった．一方で，主張抽出を行った場合，GPT-4o を用いた条件において，適合率 0.17，再現率 0.25，F_1 値 0.20 となり，訂正文生成が可能であることが確認された．

次に補足文生成については，表 3 に示すように，主張抽出を行った条件で高い精度が得られた．特に，GPT-4o を用いた場合には，適合率 0.65，再現率 0.58，F_1 値 0.59 と最も高い値を示した．これに対し，主張抽出を行わない場合や，補足文生成に gemma3:12b を用いた場合には，F_1 値が低下する傾向が見られた．

表 2　訂正文生成の評価結果.

主張抽出	音声認識モデル	訂正文生成のモデル	適合率	再現率	F_1 値
なし	Whisper（Local）	gemma3:12b	0.00	0.00	0.00
あり	Whisper（Local）	gemma3:12b	0.00	0.00	0.00
なし	Whisper（API）	GPT-4o	0.00	0.00	0.00
あり	Whisper（API）	GPT-4o	0.17	0.25	0.20

表 3　補足文生成の評価結果.

主張抽出	音声認識モデル	補足文生成のモデル	適合率	再現率	F_1 値
なし	Whisper（Local）	gemma3:12b	0.07	0.06	0.07
あり	Whisper（Local）	gemma3:12b	0.33	0.29	0.28
なし	Whisper（API）	GPT-4o	0.34	0.29	0.32
あり	Whisper（API）	GPT-4o	0.65	0.58	0.59

5 議　　論

5.1 主張抽出の議論

表 1 の結果より，主張抽出の精度は音声認識結果の品質に大きく影響されることが分かる．Whisper（API）を用いた場合は Whisper（Local）と比較して，適合率，再現率，F_1 値のすべてが高く，特に再現率が高い傾向が見られた．これは，音声認識の誤りが少ないほど，発話中の重要な説明や断定的な内容を LLM が正確に把握しやすくなるためであると考えられる．一方で，Whisper（API），Whisper（Local）のいずれにおいても，発話内容に論文由来の専門用語や略語が多く含まれる場合には，単語の認識誤りが増加する傾向が確認された．このような音声認識の誤りは，主張抽出において主張の欠落や意味の誤解釈を引き起こす要因となっていた．例えば，発表において「低リソース言語に対応した指示追従モデルを構築したい」という内容の発話が行われた場合，音声認識モデルによる文字起こしでは「この論文何をしているかと言いますとペリソース言語に対応した市立移住モデルを構築したいという論文になります」といった誤認識が生じることがあった．このような誤った文字起こし結果に基づいて主張抽出を行うと，本来の意味とは異なる，「ペリソース言語に対応した市立移住モデルを構築したい」といった意味の通らない主張が抽出されてしまった．

また，発表者が事前に作成した資料を基に，論文の背景，提案手法，実験設定，結論などを説明している発話においては，主張抽出が質疑応答の発話よりも高い精度で行われていることが確認された．これらの発話は，スライド構成に沿って論理的に説明されることが多く，定義や結果に関する断定的な表現が明確である一方で，言い直しや曖昧な言い回しが少ないという特徴を持つ．そのため，音声認識結果が一定の品質を満たしている場合，LLM が発話の要点を把握しやすく，主張として抽出すべき内容を安定して取り出すことが可能であったと考えられる．

一方で，Whisper（API）を用いた場合でも F_1 値は 0.67 にとどまっており，すべての主張を完全に抽出できているわけではない．特に，質疑応答における発話では，主張が正しく抽出されないケースが確認された．例えば，「論文の実験は倫理審査を通していたか」という質問に対し，発表者は複数の発話を経た後に「分からない」と回答していた．しかし，発話の途中に含まれる言い直しや曖昧な表現に加え，音声認識の誤りの影響により，会話の文脈が十分に反映されず，抽出された主張で

は Whisper（API）および Whisper（Local）のいずれにおいても「この論文の実験は倫理審査に通っている」という誤った内容が出力された．このように，質疑応答のような対話的で断片的な発話に対しては，音声認識の誤りと文脈理解の困難さが重なり，誤抽出が生じる可能性があることが示唆される．事前に内容を整理した発話と突発的な発話とでは，主張抽出の難易度に差があると考えられる．

5.2 主張抽出の有無による訂正文および補足文生成の議論

表 2 および表 3 の結果より，主張抽出を行うことで，訂正文および補足文の生成精度が向上することが確認された．特に補足文生成においては，主張抽出を行った条件で適合率，再現率，F_1 値が大きく改善しており，発話内容から主張を抽出することが，補足対象の明確化に有効であることが示唆される．これは，発話内容全体をそのまま入力とする場合と比較して，重要な説明や断定的な内容のみが入力として与えられることで，RAG による情報検索および生成が適切に機能したためであると考えられる．

訂正文生成においても，全体として精度は高くないものの，主張抽出を行うことで有効な訂正文が生成された事例が確認された．具体的には，発表中に「評価実験 1 と評価実験 2 の違い」について説明する場面において，発表者が不明確な情報を発言していた事例が挙げられる．この場面での文字起こしされた具体的な発話は「1 と 2 差がトピックと人数以外にあるのかどうか論文を見た感じでは，えーっと人数とテーマと，あと，参加している人が，評価実験 1 は結構大学とかその 身内で，2 はそれ以外の一般の人に参加を募ってやっているっていう差があって，他は多分ほとんど同じ内容で実験していると思います.」となっている．この発話において発話中に言い直しや補足的な説明が含まれていたため，主張抽出を行わずに文字起こしテキストを直接入力した場合には，訂正すべき主張が明確にならず，訂正文は出力されなかった．一方で，主張抽出を行った場合には，「評価実験 1 と 2 は，人数とテーマ以外に、評価実験 1 は身内，2 は一般の人に参加を募っている点が異なる.」という主張が抽出されており，この主張を入力として RAG による照合を行うことで，「評価実験 1 の参加者はファシリテーション協会の会員とその知り合いであり，評価実験 2 の参加者は外部委託業者によって募集された一般の人々である.」という適切な訂正文が生成された．この結果は，主張抽出が訂正文生成においても重要な前処理として機能することを示唆している．一方で，訂正文生成の全体的な評価値が低くなった要因として，訂正文生成のプロンプトにおいて「誤字の訂正は行わない」と指示していたにもかかわらず，音声認識の過程で生じた誤字を訂正対象として検出してしまうケースが多く見られた点が挙げられる．例えば，「正例」という語が音声認識により「精霊」と誤認識されていた場合，この誤字が訂正対象として出力される事例が確認された．このような出力は人手評価において不正解と判定されるため，適合率および再現率の低下につながったと考えられる．

また，訂正文生成の評価値が低くなったもう一つの要因として，一回の発表において訂正が必要な理想的な訂正文の数自体が少ない点が挙げられる．訂正すべき誤りがほとんど含まれない発表に対しても，システムが訂正文を生成しようとすることで，相対的に誤検出の影響が大きくなり，評価指標が低下した可能性がある．このことから，訂正文生成においては，訂正が本当に必要かどうかを事前に判定する仕組みの導入も重要であると考えられる．

モデル間の比較では，gemma3:12b を用いた場合に，GPT-4o と比較して訂正文および補足文生成の精度が低くなる傾向が見られた．この原因として，gemma3:12b は GPT-4o と比べてモデルサイズが小さく，複雑なプロンプトの意図を十分に理解できていなかった可能性が考えられる．特に，訂正と補足を厳密に区別する必要がある本タスクでは，プロンプト理解能力の差が出力精度に影響したと考えられる．

さらに，主張抽出を行わなかった場合には，訂正や補足を行うべき範囲が不明確となり，不適切な訂正文や補足文が生成されるケースが確認された．例えば，発表中で「例えば『1 日に何 ml の水を飲むべきですか』というクエリに対して，大人は 3000ml の水を飲むべきといったコーパスを生成したい」という説明がなされていた時に，主張抽出を行わない条件では，この例示部分に対して「男性は 3000ml，女性は 2100ml の水を飲むべきです」といった訂正文が生成された．これは，例として提示された仮定の内容が主張として誤って扱われたことによるものである．一方で，主張抽出を行った場合には，このような例示は主張として抽出されないため，不適切な訂正文および補足文の生成を防ぐことができた．この結果は，主張抽出が訂正文および補足文生成の前処理として重要な役割を果たしていることを示していると考えられる．

一方で，本研究では RAG における検索性能やエンベディング手法そのものの評価は行えておらず，訂正文および補足文の生成精度が，どの程度検索結果の品質に依存しているかについては明らかにできていない．また，訂正文および補足文生成に用いたプロンプト設計や，外部データの検索方法，エンベディングの粒度や単位についても，さらなる検討の余地がある．これらの要素を体系的に評価することが，訂正文および補足文生成の精度向上につながると考えられる．

5.3 学習支援への適用に関する課題

本研究では，発話内容から主張を抽出し，訂正文および補足文を生成する手法を提案し，主張抽出と訂正文および補足文生成の精度に関する基礎的な有用性を示した．しかしながら，本手法を実際の学習支援に適用するためには，生成された情報をどのような形で学習者に提示するかという点について，さらなる検討が必要である．

フィードバックのタイミングは学習成果形成に決定的な役割を果たす [21]．このことから訂正文や補足文を提示するタイミングも学習効果に大きな影響を与えると考えられる．例えば，発表中にリアルタイムで提示する場合には，発表者や聴講者の注意を分散させ，内容理解を妨げる可能性がある一方で，誤りや不足に即座に気づかせるという利点もある．一方で，発表後

にまとめて提示する場合には，誤った理解がその場で修正されないまま進行してしまう可能性がある．このように，提示タイミングの設計は，学習者の理解度や発表の目的に応じて慎重に検討する必要がある．

また，提示する情報量や表現方法も重要な課題である．訂正や補足が過剰に提示されると，学習者にとっては情報量が多くなりすぎ，どの点が重要なのか分かりにくくなる恐れがある．特にアウトプット型学習の場面では，学習者自身が考えながら説明を行うことが重要であるため，訂正や補足が一方的に与えられすぎると，主体的な学びを阻害する可能性も考えられる．そのため，訂正と補足の重要度に応じて提示量を調整したり，簡潔な要約として提示したりするなど，情報の取捨選択が求められる．

さらに，本研究では生成された訂正文および補足文の正確性を評価しているが，それらが実際に学習者の理解や学習成果にどのような影響を与えるかについては検証できていない．例えば，訂正文を提示された学習者が，自身の誤りをどの程度理解し，その後の学習に活かすことができているか，あるいは補足文が理解の深化につながっているかといった教育的効果については，別途評価が必要である．

このように，本手法を学習支援システムとして実運用するためには，主張抽出や訂正文および補足文生成の精度向上に加えて，提示タイミング，提示量および学習者への影響といった観点から総合的な検討が必要である．今後は，実際の教育現場での利用を想定したインタフェース設計や運用方法の検討を行い，学習者の理解促進や主体的な学びにどのように寄与するかを実証的に評価していくことが課題である．

6　まとめと今後の課題

本研究では，アウトプット型学習における発表内容の理解支援を目的として，発話音声から話者の主張を抽出し，RAG に基づいて訂正文および補足文を生成するシステムを提案した．提案手法では，音声認識によって得られた文字起こしデータから LLM を用いて主張を抽出し，抽出された主張を単位として根拠となる文書との照合を行うことで，発表における発話に含まれる訂正文および補足文を生成する．

研究室内で行われた論文紹介のプレゼンテーションを対象とした実験の結果，主張抽出を導入することで，訂正文および補足文の生成精度が向上することを確認した．特に，主張抽出を行わずに文字起こしテキストを直接入力した場合と比較して，主張を抽出した場合には，訂正および補足の対象が明確化され，RAG による情報検索および生成が効果的に機能することが示された．また，発表資料に基づいて論理的に説明される発話においては，主張抽出が質疑応答の際の発話よりも高い精度で行われることも確認された．

一方で，本研究にはいくつかの課題が残されている．主張抽出と訂正文および補足文生成の精度は，音声認識結果の品質に大きく依存しており，特に専門用語や略語を含む発話や，質疑応答における断片的で言い直しの多い発話に対しては，誤った抽出が生じる場合があった．また，訂正文生成においては，音声認識の誤りによる誤字が訂正対象として検出されてしまうなど，訂正が本当に必要な主張を見極めることの難しさも明らかになった．

さらに，本研究では生成された訂正文および補足文の正確性を中心に評価を行ったが，それらの情報が学習者の理解や学習成果にどのような影響を与えるかについては検証できていない．訂正や補足を提示するタイミングや提示量，提示方法によっては，学習者の注意を分散させたり，主体的な学びを阻害したりする可能性も考えられる．

今後の課題としては，質疑応答のような対話的な発話に対して複数の発話を統合して主張を抽出する手法の検討や，訂正が本当に必要かどうかを事前に判定する仕組みの導入が挙げられる．また，実際の教育現場での利用を想定し，訂正文および補足文の提示タイミングやインタフェース設計を含めた運用方法についても検討する必要がある．これらの課題に取り組むことで，アウトプット型学習をより効果的に支援する実用的な学習支援システムの構築を行っていきたい．

謝　　辞

本研究は，JSPS 科研費 JP24K03228 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] 文部科学省高等教育局. 令和 3 年度の大学における教育内容等の改革状況について（概要）. https://www.mext.go.jp/content/20230908-mxt_daigakuc01-000031526_1.pdf. 2025 年 1 月 15 日閲覧.

[2] 中央教育審議会. 新たな未来を築くための大学教育の質的転換に向けて～生涯学び続け、主体的に考える力を育成する大学へ～(答申). https://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2012/10/04/1325048_1.pdf. 2025 年 1 月 15 日閲覧.

[3] Shota Imamura, Hirotaka Hiraki, and Jun Rekimoto. Serendipity Wall: A discussion support system using real–time speech recognition and large language model. In *Proceedings of the Augmented Humans International Conference 2024*, pp. 237–247, 2024.

[4] 木下良輔, 櫻井崇貴, 白松俊. LLM を用いたファクトチェック機能の試作と Web 議論における関連情報推薦システムへの応用. 第 12 回市民共創知研究会 2023 (CCI-012), pp. 41–44, 3 2024.

[5] Steffen Steinert, Karina E. Avila, Stefan Ruzika, Jochen Kuhn, and Stefan Küchemann. Harnessing large language models to develop research–based learning assistants for formative feedback. *Smart Learning Environments*, Vol. 11, No. 62, 2024.

[6] Frederic Kirstein, Terry Ruas, Robert Kratel, and Bela Gipp. Tell me what I need to know: Exploring LLM–based (personalized) abstractive multi–source meeting summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 920–939, 2024.

[7] Yun Long and Yu Zhang. Enhanced classroom dialogue sequences analysis with a hybrid AI agent: Merging expert rule-base with large language models. *arXiv preprint arXiv:2411.08418*, 2024.

[8] Yang Shi, Yang Wang, Ye Qi, John Chen, Xiaoyao Xu, and Kwan-Liu Ma. IdeaWall: Improving creative collaboration through combinatorial visual stimuli. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative*

Work and Social Computing, pp. 594–603, 2017.

[9] Salvatore Andolina, Khalil Klouche, Diogo Cabral, Tuukka Ruotsalo, and Giulio Jacucci. InspirationWall: Supporting idea generation through automatic information exploration. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pp. 103–106, 2015.

[10] 中明理沙, 吉添衛, 服部宏充. 議論支援システム AIR-VAS への LLM に基づく議論エージェントの導入とその効果. 人工知能学会全国大会論文集, 2F6GS505, 2021.

[11] Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. The Idea Machine: LLM–based expansion, rewriting, combination, and suggestion of ideas. In *Proceedings of the 14th Conference on Creativity and Cognition*, pp. 623–627, 2022.

[12] Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. Idea Expander: Supporting group brainstorming with conversationally triggered visual thinking stimuli. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pp. 103–106, 2010.

[13] Rrubaa Panchendrarajan and Arkaitz Zubiaga. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *arXiv preprint arXiv:2401.11969*, 2024.

[14] Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. Claim extraction for fact–checking: Data, models, and automated metrics. *arXiv preprint arXiv:2502.04955*, 2025.

[15] Nhat Tran, Diane Litman, and Amanda Godley. Using large language models to analyze students' collaborative argumentation in classroom discussions. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference*, pp. 111–125, Pittsburgh, PA, USA, 2025.

[16] Chitralekha Gupta, Hanjun Wu, Praveen Sasikumar, Shreyas Sridhar, Priambudi Bagaskara, and Suranga Nanayakkara. Factually: Exploring wearable fact–checking for augmented truth discernment. In *Proceedings of the 2025 ACM Workshop on Human–AI Interaction for Augmented Reasoning*, 2025.

[17] Venktesh V and Vinay Setty. LiveFC: A system for live fact–checking of audio streams. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 1060–1063, 2024.

[18] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of Varying Shades: Analyzing language in fake news and political fact–checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2931–2937, 2017.

[19] Burges Chris, Shaked Tal, Renshaw Erin, Lazier Ari, Deeds Matt, Hamilton Nicole, and Hullender Greg. Learning to rank using gradient descent. In *ACM ICML 2005*, pp. 89–96, 2005.

[20] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[21] Hongyu Mai. The comparative effect of immediate and delayed feedback on EFL learners' engagement and willingness to collaborate. *PsyCh Journal*, pp. 1008–1017, 2025.