プラグマティクス拡張と検索モデル学習による対話履歴検索の精度改善

増田湧真 [1]　滝口哲也 [2]　有木康雄 [2]　岩﨑雄介 [3]　小島幹 [3]　山本昌輝 [3]
[1] 神戸大学工学部情報知能工学科
[2] 神戸大学大学院システム情報学研究科
[3] 株式会社デンソーテン
2265075t@stu.kobe-u.ac.jp
{takigu, ariki}@kobe-u.ac.jp
{yuusuke.iwasaki.j7p, motoki.kojima.j3h, masaki.yamamoto.j3h}@jpgr.denso.com

概要

Dense Retrieval はクエリ表現のみから関連文書を同定するため，対話履歴検索のようにクエリが短く含意が多い場面では「クエリ理解不足」が性能を制限する．本研究では，クエリのプラグマティクスを **INTENT / BACKGROUND / CONSTRAINTS** に分解して明示化し，それに基づいてクエリを拡張する検索手法を提案する．さらに，拡張クエリを用いて検索モデルを対照学習でファインチューニングし，拡張と検索モデル学習を一体で最適化する．提案法は単なる表層的 rewrite と異なり，プラグマティクス要素単位で拡張を設計できる点に新規性がある．小〜中規模モデルにおいて，プラグマティクス拡張により対話履歴検索の性能改善を確認した．

1　はじめに

対話システムでは，現在の発話（クエリ）に対して過去の対話履歴から関連情報を検索する処理が重要となる．本研究は，この対話履歴検索におけるクエリ理解の不足に着目し，プラグマティクス拡張と検索モデル学習を組み合わせて Dense Retrieval（DR）の精度を向上させる．

1.1　対話履歴検索の課題

対話では発話が短く，省略や文脈依存が多いため，表層的な語句一致や単純な意味類似度だけではユーザ意図や参照対象を適切に特定できない．このことが対話履歴検索の主な性能低下要因となる．

1.2　Dense Retrieval の課題

DR に用いられる埋め込みモデルは，文脈・含意・指示語といった会話特有の暗黙情報を扱うことが苦手である．短いクエリでは検索の手がかりが不足するため，表面的な言い換えでは性能改善が限定的であり，意図・前提・制約といった深い意味情報を検索表現として補う必要がある．

1.3　本研究の目的と貢献

本研究は，クエリに不足する意味情報をプラグマティクス推定により補完し，深層的な検索表現へ変換することを目的とする．さらに，拡張クエリに適応した DR モデルを弱教師ありの対照学習で再調整することで，履歴チャンクをより適切に検索できるようにする．本研究の貢献を以下に示す．

- **プラグマティクス拡張による深層的クエリ強化**: LLM により意図（INTENT），前提（BACKGROUND），制約（CONSTRAINTS）を推定し，クエリへ明示的に付与することで検索表現を強化する．
- **拡張クエリに適応する Dense Retrieval モデルの学習**: クエリ拡張により意味空間が変化するため，対照学習を用いて DR モデルを再調整し，拡張クエリから適切に文脈を検索できるよう最適化する．

提案手法は，Recall，All Hit，MRR，nDCG の各指標で一貫した性能向上を確認した．

2　関連研究

2.1　対話履歴検索

対話履歴検索（Conversational Search）は，現在のクエリに対して過去の会話から関連文脈を取得するタスクであり，近年は Dense Retrieval が主流となっている [1]．ChatGPT Retriever [2] や HAConvDR [3] など LLM を活用した手法も提案されているが，対話に内在するプラグマティクス情報（意図・前提・制約）を明示的に扱う枠組みは十分に整備されていない．

2.2　LLM を用いたクエリ拡張

LLM によるクエリ拡張は，曖昧な発話の補完や検索手がかりの付与を目的として利用される．Query2doc [4] は query-to-document 生成に基づく手法を示し，対話型検索における LLM 書き換えの分析 [5] も報告されている．しかし，意図・背景・制約といった多面的情報を体系的に拡張へ反映する研究は限定的である．

2.3　Dense Retrieval と弱教師あり学習

Dense Retrieval の性能向上には，擬似データを用いた弱教師ありの対照学習が広く用いられる．RocketQAv2 [6] や GPL [7] は，擬似クエリ生成と hard negatives を組み合わせ，ラベルなしデータから高品質な retriever を学習可能であることを示した．

3　提案手法

本研究は，(1) クエリのプラグマティクス（INTENT / BACKGROUND / CONSTRAINTS）を明示化して拡張し，(2) 拡張クエリを用いた弱教師あり学習により検索モデルを最適化することで，対話履歴検索を高精度化する．

3.1　問題設定

タスクを「クエリに最も関連する過去チャンクを検索する問題」と定義する．対話履歴をチャンク集合 $C = \{c_i\}$ とし，検索スコア $s(q, c)$ に基づき

$$\hat{c} = \arg\max_{c \in C} s(q, c) \tag{1}$$

を返す．Dense Retrieval では q と c をエンコードし，内積等で $s(\cdot)$ を計算する．

3.2　クエリのプラグマティクス拡張

LLM によりクエリ q から次の 3 要素を推定する：

- **INTENT**: 達成したい目的
- **BACKGROUND**: 想定される状況・参照対象
- **CONSTRAINTS**: 時間・距離・費用等の条件

省略や暗黙知を補い，検索の手がかりとなる意味情報を付与する．

推定した 3 要素を用い，拡張クエリ $\tilde{q}$ を生成する．本研究では以下の 2 形式を用いる：

- **prefix 型**：
  `INTENT: ... / BACKGROUND: ... / CONSTRAINTS: ... / QUERY: q`
- **rewrite 型**：意図・背景・制約を踏まえて単一文へ再構成する．

これにより，表層に現れない検索意図を補完し，検索エンコーダが利用可能な意味情報を増強することで semantic gap を縮小し，チャンクとの整合的な意味空間を形成する．

3.3　検索モデルの弱教師あり学習

拡張クエリに適応したエンコーダを学習するため，pseudo-query を用いた弱教師ありデータで対照学習を行う．図 1 に学習フローを示す．

3.3.1　擬似クエリと正例の生成

各チャンク c_i に対し LLM に以下を行わせる：

- そのチャンクが検索における正解となる抽象的クエリ q_i を生成（A.2 に具体例記載）
- q_i をプラグマティクス拡張し $\tilde{q}_i$ を得る
- $(\tilde{q}_i, c_i)$ を正例ペアとして用いる

3.3.2　二段階の対照学習

InfoNCE 損失を用いて，バッチ内で正例を近づけ，その他を負例として遠ざける：

$$L = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s(\tilde{q}_i, c_i)/\tau)}{\sum_{j=1}^{B} \exp(s(\tilde{q}_i, c_j)/\tau)} \tag{2}$$

ここで，B はバッチサイズであり，τ は温度パラメータである．これにより拡張クエリを基準とした意味空間を構築する．次に一段階モデルで検索し，類似性は高いが正例ではないチャンクを hard negative として追加学習する．これを二段階目モデルと呼ぶことにする．対話データでは同一文書内

チャンクが上位に入りやすいため，偽負例を避ける目的で同一文書の候補をマスクする．この工夫により負例の質が向上し，学習が安定する．

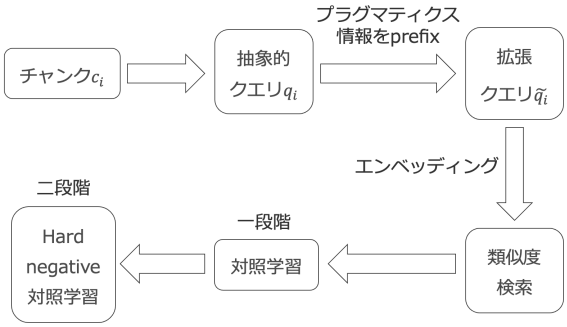

図 1　提案手法の学習フロー

4　実験設定

4.1　データセットと評価

本研究では LiHua-World [8] の 1 年分の対話データを用いる．評価には付属の既存 QAE に加え，全ての検索チャンクを元に LLM により作成した QAE（これを抽象的 QAE と呼ぶ． A.3 参照）も使用し，Recall@10，All Hit@10，MRR@10，nDCG@10 を主指標とする．抽象的 QAE については @40 も参考として報告する．

4.2　学習データの作成

3.3.1 で述べた方法により，各チャンク c_i から LLM により pseudo-query q_i を生成し，これをプラグマティクス拡張して得た $\tilde{q}_i$ と組み合わせて $(\tilde{q}_i, c_i)$ を正例として用いる．

4.3　学習手法

本研究では all-MiniLM-L6-v2, e5-base-v2, bge-base-en-v1.5 を使用し，in-batch 対照学習の後に hard negative（10 件）による追加学習を行う．これらを選定した理由は次の通りである．(i) MiniLM 系は軽量で一般領域の DR ベースラインとして広く用いられており，プラグマティクス拡張の効果を最も純粋に観察できる．(ii) e5-base-v2 は “text embedding の標準的フォーマット（query/document の prefix 仕様）” を持ち，情報検索に特化した事前学習が施されているため，拡張クエリとの整合性の違いを評価するのに適している．(iii) bge-base-en-v1.5 は近年高性能な英語向け汎用埋め込みモデルとして評価されており，MiniLM・e5 と異なる構造のモデルに対しても提案手法が有効に働くかを検証する目的で採用した．

4.4　比較手法

以下の 4 条件を比較する：

- **ベースライン**: 拡張を行っていない抽象的クエリを既存 Dense Retriever に入力．
- **拡張のみ**: 抽象的クエリをプラグマティクス拡張するが，モデルは学習しない．
- **学習のみ**: 弱教師ありデータでモデルの二段階学習を実施．推論は抽象的クエリ．
- **拡張 + 学習（提案法）**：拡張クエリを使用し，弱教師あり二段階学習したモデルを用いる．

5　実験結果と考察

5.1　既存 QAE に対する検索性能

まず，LiHua-World 付属の既存 QAE で検索性能を確認した（表 1）．この結果は，今回作成した抽象的 QAE による検索性能と比較することにより，抽象的クエリによる検索難易度が大幅に上昇していることを示す意味で重要である．

5.2　抽象的 QAE に対する検索性能

次に，抽象的な評価データセットに対する検索性能を評価した．抽象的 QAE は，抽象的で多義的な検索タスクとなっているため，Dense Retrieval の限界を評価するのに適している．

まず@10 での抽象的 QAE に対する結果を表 2 に示す．抽象的なクエリでは表層手がかりが少なく，@10 の網羅性が低下しやすい．@10 のみではモデルの検索能力を十分に評価できない可能性がある．@40 は @10 より網羅性が高く，関連文脈をより多く含められる（表 2）．

5.3　プラグマティクス拡張と学習の効果

5.3.1　拡張形式の比較

rewrite による検索結果は付録（表 5）に示す．本論文では，学習・推論の整合が取りやすい prefix を主に用いる．

表 1　既存 QAE に対する検索性能

モデル	Recall@10	All Hit@10	MRR@10	nDCG@10
all-MiniLM	0.679	0.656	0.486	0.523
e5-base	0.789	0.769	0.608	0.643
bge-base	0.781	0.768	0.605	0.641

表 2　抽象的 QAE に対する検索性能

@10				
モデル	Recall@10	All Hit@10	MRR@10	nDCG@10
all-MiniLM	0.298	0.123	0.256	0.216
e5-base	0.299	0.119	0.287	0.235
bge-base	0.311	0.097	0.293	0.239
@40				
モデル	Recall@40	All Hit@40	MRR@40	nDCG@40
all-MiniLM	0.464	0.229	0.298	0.285
e5-base	0.515	0.282	0.304	0.299
bge-base	0.447	0.212	0.267	0.261

5.3.2　prefix 形式での性能

抽象的 QAE に対して，プラグマティクス拡張と検索モデル学習を行った結果を表 3 に示す．学習・推論時ともに拡張形式は prefix 型を用いた．提案手法により，全モデルで Recall@10，All Hit@10，MRR@10，nDCG@10 の各指標が，表 2 に比べ一貫して改善されているのが分かる．e5-base-v2 では，アブレーションと二段階学習の効果を表 4 に示す．拡張のみ，学習のみのいずれも改善が見られるが，拡張と学習を組み合わせた場合に最も高い性能が得られた．

これは，拡張により意味手がかりが増え，弱教師あり学習でその表現に適応した検索空間が形成されるためである．

5.4　Hard Negative Fine-tuning の効果

同一ドキュメント由来の候補を除外しない場合，偽負例が混入して学習が不安定になりやすい．本研究ではマスクを適用した設定の結果を報告する（表 4）．マスク有りだと性能低下は見られなかったが，大きな改善は見られなかった．これは生活ログでは異なるドキュメント間にも類似チャンクが存在し，マスクしても偽負例が残ることが原因と考えられる．

5.5　全体考察

実験結果から，DR の性能はクエリの抽象度に強く依存し，表層手がかりの乏しい状況では semantic gap が顕在化することが分かった．プラグマティクス拡張はこの不足を補い，検索空間を整合化する上で有効である．

表 3　提案手法の性能

モデル	Recall@10	All Hit@10	MRR@10	nDCG@10
all-MiniLM	0.332	0.123	0.316	0.251
e5-base	0.370	0.132	0.420	0.319
bge-base	0.368	0.128	0.408	0.314

表 4　アブレーションと二段階学習の効果（e5-base-v2, @10）

アブレーション				
方法	Recall@10	All Hit@10	MRR@10	nDCG@10
拡張のみ	0.299	0.119	0.287	0.235
学習のみ	0.364	0.128	0.384	0.297
拡張+学習	0.370	0.132	0.420	0.319
二段階学習				
方法	Recall@10	All Hit@10	MRR@10	nDCG@10
一段階目	0.370	0.132	0.420	0.319
二段階目	0.379	0.145	0.437	0.331

プラグマティクス拡張はこの gap を埋める役割を果たす．INTENT/BACKGROUND/CONSTRAINTS を付与することで，クエリが履歴チャンクと比較しやすい形に正規化され，抽象的クエリでも検索手がかりが大幅に増える．特に prefix 型は情報構造を保持した拡張であり，rewrite 型に比べ意味情報の損失が少ない点が有効に働いたと考えられる．

弱教師あり学習は，拡張クエリの分布に検索モデルを適応させる工程として機能する．拡張のみでは検索器が新しい表現形式に十分追従できないため改善量は限定的であるが，学習を併用するとクエリ空間とチャンク空間が整合的に再構築され，性能が大きく向上した．すなわち，本研究の枠組みは「拡張（query-level）」と「適応（model-level）」が二段で作用する点に意義がある．

また，@10 と @40 の性能差から，単一段階の Dense Retrieval では抽象的クエリに対する網羅性が不足する可能性があり，再ランキング等との組み合わせが有効であることが示唆される．

総じて，対話履歴検索の性能向上には，クエリの深層意味を補う構造化拡張と，その表現に適応する学習を併用することが重要であり，本研究はその有効性を示した．

6　おわりに

本研究では，クエリのプラグマティクス拡張と対照学習を組み合わせた対話履歴検索手法を提案し，小〜中規模モデルで一貫した性能向上を確認した．今後はチャンク側へのプラグマティクス付与や Reranking 手法との併用を検討し，応答生成への影響も分析する予定である．

参考文献

[1] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pages 6769–6781. Association for Computational Linguistics, 2020. https://aclanthology.org/2020.emnlp-main.550/.

[2] Kelong Mao, Chenlong Deng, Haonan Chen, Fengran Mo, Zheng Liu, Tetsuya Sakai, and Zhicheng Dou. Chatretriever: Adapting large language models for generalized and robust conversational dense retrieval. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pages 1227–1240, Miami, Florida, USA, 2024. Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.71/.

[3] Fengran Mo, Chen Qu, Kelong Mao, Tianyu Zhu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. History-aware conversational dense retrieval. In **Findings of the Association for Computational Linguistics: ACL 2024**, pages 13366–13378, Bangkok, Thailand, 2024. Association for Computational Linguistics. https://aclanthology.org/2024.findings-acl.792/.

[4] Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pages 9414–9423. Association for Computational Linguistics, 2023. https://aclanthology.org/2023.emnlp-main.585/.

[5] Ryo Abe, Takuma Takeoka, and Makoto Oyamada. Query rewrite for conversational search with large language models. In **Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing**, pages P8–4, 2024. In Japanese. https://www.anlp.jp/proceedings/annual_meeting/2024/pdf_dir/P8-4.pdf.

[6] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pages 2825–2835. Association for Computational Linguistics, 2021. https://aclanthology.org/2021.emnlp-main.224/.

[7] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pages 2345–2360. Association for Computational Linguistics, 2022. https://aclanthology.org/2022.naacl-main.168/.

[8] Tianyu Fan, Jingyuan Wang, Xubin Ren, and Chao Huang. Minirag: Towards extremely simple retrieval-augmented generation, 2025. arXiv:2501.06713. https://arxiv.org/abs/2501.06713.

A 付録

A.1 プラグマティクス推定プロンプト

以下は，ユーザ発話（クエリ）から INTENT / BACKGROUND / CONSTRAINTS を推定するためのプロンプト例である．

```
You are a pragmatics-aware analyzer for the
LiHua-World QA dataset.
Goal:
Given a user query, infer what the user REALLY
wants:
- what kind of answer they are looking for,
- what background situation is plausible,
- what constraints or conditions they care
about.
IMPORTANT DOMAIN CONSTRAINTS:
- Stay within everyday-life topics (e.g.
hobby, study, work, family, outings).
- Do NOT jump to irrelevant domains.
- If the query is abstract(e.g. "any good drive
destinations?"), you are allowed to infer a
plausible everyday-life background.
Output the following JSON fields:
1. "intent":
- Short natural language description of what
type of information the user wants.
- Describe it in 1 short sentence, not a label.
- Example: "User wants suggestions for easy
day-trip drive destinations."
2. "background":
- Short natural language description of the
plausible situation or context.
- You MAY infer unstated but reasonable
everyday context,
but stay inside LiHua-World style daily life.
- Example: "They are thinking about a weekend
plan and considering a relaxed drive."
3. "constraints":
- Natural language description of explicit or
implicit constraints:
time limits, energy, money, location, mood,
skill level, etc.
- If nothing is obvious, use an empty string
"".
Output JSON ONLY (no explanation) with keys:
intent / background / constraints
— Query —
{query}
```

A.2 学習データのチャンクから抽象クエリ作成の具体例

Chunk:
"その興奮を音楽に注ぎ込んで，最高の時間を過ごしましょう！日曜日は皆さんが最高の雰囲気を持ってきてくれることを願っています！伝説的な日になるでしょう！"

Pseudo-Query:
"バンドメンバーみんなのテンションを日曜日のリハやライブ本番に向けて高めるメッセージの例が知りたい"

A.3 抽象的 QAE（Query–Answer–Evidence）の具体例

Query:
"今夜は軽めにしたい気分．どうするのがちょうどいい？"
Answer:
"ダウンタウンのカフェでスープかサンドを選ぶか，配達のサワードウを軽く焼いてオリーブオイルで食べると，外食と宅食の両方で手軽に済ませられるよ．"
Evidence (dialogue chunk):
"20260108_1100_chunk01<and>
20260317_0800_chunk01<and>
20260314_1700_chunk01"

Query:
"フェスに行くときって，持ち物はどれくらいがちょうどいい？"
Answer:
"小さめの現金，カメラ，モバイルバッテリー，薄手の上着．"
Evidence (dialogue chunk):
"20260106_2000_chunk01<and>
20260110_1400_chunk01"

A.4 推定したプラグマティクスの具体例

Query:
"メインストリート周辺を 2 時間しか滞在できない場合，最も効率的な順序は？"
Intent:
"ユーザーは，2 時間の枠内でメインストリート周辺の複数のスポットを訪問する最適な順序についてアドバイスを求めている．"
Background:
"ユーザーはメインストリート付近で短い外出を計画しており，おそらくいくつかの場所（店舗，カフェ，観光スポットなど）を想定しています．"
Constrainsts:
"総時間は約 2 時間に制限されています．場所はメインストリート周辺です．最も簡単なまたは効率的な訪問順序を好みます"

A.5 rewrite 形式での検索結果

表 5 rewrite 形式での検索結果（モデル：e5-base-v2）

方法	Recall@10	All Hit@10	MRR@10	nDCG@10
未拡張・未学習	0.299	0.119	0.287	0.235
拡張のみ（rewrite）	0.263	0.088	0.251	0.201
学習のみ（弱教師あり）	0.342	0.115	0.394	0.294
拡張+学習（最終提案）	0.372	0.132	0.416	0.318