



DialFill Framework: Towards Full Exploitation of Retrieved Information in Conversational AI

薛, 強

(Degree)

博士 (工学)

(Date of Degree)

2025-09-25

(Date of Publication)

2027-09-25

(Resource Type)

doctoral thesis

(Report Number)

甲第9328号

(URL)

<https://hdl.handle.net/20.500.14094/0100498368>

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。



(別紙様式 3)

論文内容の要旨

氏 名 薛強

専 攻 情報科学

論文題目 (外国語の場合は, その和訳を併記すること。)

DialFill Framework: Towards Full Exploitation of
Retrieved Information in Conversational AI

(DialFill フレームワーク : 対話 AI における検索情報の完
全活用に向けて)

指導教員 滝口哲也

(注) 2, 000 字～4, 000 字でまとめること。

Open-domain dialogue systems based on large language models have developed rapidly, showing strong abilities in fluent response generation. However, there remain three fundamental and persistent problems: the tendency to hallucinate facts, failure to effectively use new evidence retrieved at run time, and frequent repetition or overuse of generic expressions. These issues reflect a deeper gap between the static knowledge stored in model parameters and the real-world need for integrating up-to-date, diverse information from outside sources. Traditional retrieval-augmented generation (RAG) methods attempt to bridge this gap by providing external evidence as input to the model. Yet, RAG still cannot guarantee the inclusion of critical facts, nor does it offer fine control over whether evidence should be copied directly or rephrased. Current methods for enforcing evidence use, such as constrained decoding, often cause a heavy computational load or loss of response fluency, while other approaches focused on stylistic control fail to solve the knowledge-drop problem. Despite the introduction of copy mechanisms, pointer-generators, or additional classifiers, none of these techniques deliver a unified, efficient, and controllable solution that ensures faithful grounding, fluency, and diversity in real time. Therefore, it remains an open question how to design a dialogue generation framework that can (i) always include required knowledge, (ii) allow flexible phrasing, (iii) reduce repetition, and (iv) maintain computational efficiency suitable for real-world deployment.

Chapter 1 briefly reviews the evolution of open-domain dialogue systems and discusses the tension between factual grounding and natural language flexibility. It highlights the limitations of existing retrieval and generation approaches, setting the stage for the main research questions of this thesis.

Chapter 2 provides a short survey of relevant background, including retrieval-based and generation-based dialogue architectures, knowledge selection, text infilling objectives, and recent techniques for repetition control. This chapter also standardizes terminology and notation for later chapters.

Chapter 3 introduces DialFill v1, the initial implementation of the proposed Dialogue Filling framework. In this version, the system treats response generation as a span-infilling task: a dual-encoder model retrieves a candidate response from a large index, and the most important entity or phrase is anchored by masking its surrounding context. The generator, based on a fine-tuned Transformer, is required to fill the masked parts, ensuring the evidence is always present in the output. Five different masking patterns are explored to balance lexical diversity and factual precision.

Experimental results on DailyDialog and Reddit show that DialFill v1 reduces hallucination rates and increases response diversity compared to simple retrieval-concatenation methods. However, its reliance on a single retrieved sentence makes it fragile when the retriever fails to find relevant evidence, revealing the need for a more robust framework.

Chapter 4 presents DialFill v2, which extends the method to handle multiple retrieved snippets and more complex knowledge sources, such as triples or longer passages. This version uses a unified model to predict keywords, draft masked responses, and perform dialogue infilling, all supervised with a multi-task objective. An adaptive mask-search algorithm dynamically determines how much context should be masked around the keyword based on token probabilities, enabling the system to build context-sensitive templates without costly search. Experiments on Wizard-of-Wikipedia and OpenDialKG show that v2 improves knowledge grounding and maintains efficient decoding, but some repetition problems reappear when similar entities are retrieved from different sources.

Chapter 5 explores a new generative-classification mechanism that directly connects response generation and quality control. By introducing a special JudgerToken—an auxiliary discriminator token (e.g., $\forall \text{texttt}\{\text{BAD}\}$)—into the output vocabulary and training on synthetic examples that emphasize undesirable repetition, the model learns to interpret the JudgerToken’s probability as an intrinsic signal of overall response quality. This approach allows the model to make binary judgments about candidate responses—such as whether a reply is repetitive—without the need for separate classifier networks or additional computational cost. The core idea of using output logits for quality scoring, pioneered in this chapter, sets the foundation for later developments in the framework and is crucial for enabling real-time response control in the next stage.

Chapter 6 proposes DialFill v3, which fully integrates retrieval, evidence selection, template construction, and degeneration mitigation into a single, end-to-end logits-level workflow. In this version, all named entities and possible mask configurations from the retrieved knowledge are paired with GOOD and BAD control tokens. The system evaluates every candidate combination using a single forward pass, and selects the template and anchor that minimize the relative probability of BAD responses. This joint scoring approach allows precise control over which evidence is included and how much surrounding text is generated. DialFill v3 achieves these

advances without changing the backbone model or adding significant computational cost. The method represents a unified and practical solution for retrieval-grounded dialogue generation, supporting accurate, diverse, and efficient responses in open-domain conversational AI.

Chapter 7 summarizes the main findings and discusses future research directions.

氏名	薛 強		
論文 題目	DialFill Framework: Towards Full Exploitation of Retrieved Information in Conversational AI (DialFill フレームワーク：対話 AI における検索情報の完全活用に向けて)		
審査 委員	区 分	職 名	氏 名
	主 査	教授	滝口 哲也
	副 査	教授	大川 剛直
	副 査	教授	佐野 英樹
	副 査	教授	藤井 信忠
	副 査		
要 旨			
本研究は、近年急速に進化を遂げている大規模言語モデル (Large Language Model: LLM) を用いたオープンドメイン対話システムに対して、検索情報を統合的かつ十分に活用できる新たなフレームワーク「DialFill」を提案し、その理論的背景、段階的な拡張、応用の可能性について検討したものである。 近年、オープンドメイン対話システムは、Transformer アーキテクチャに基づく大規模言語モデル (LLM) の発展により、自然言語生成と文脈理解の両面で、人間に近い性能を見せ始めている。しかしその一方で、「モデル内部の静的知識」と「外部の動的情報」とが分離しているという構造的な限界は依然として顕著である。このため、事実に基づかない幻覚 (hallucination)、新たに得られた Evidence の不十分な活用、さらには応答の反復や陳腐化といった深刻な問題が未解決のまま残されている。従来手法では、これら本質的な課題に対して、部分的かつ限定的な対応策しか提示できていないのが現状である。 こうした状況を背景に、Retrieval-Augmented Generation (RAG) 手法など、検索と生成を統合する研究が活発化している。具体的には、ユーザ発話や対話履歴に基づいて外部知識 (例: Wikipedia や知識グラフなど) を検索し、それを生成モデルに入力することで、応答の事実性や一貫性を高めるアプローチである。しかし、Retrieved Evidence を確実に応答に反映させることや、Evidence の引用と再表現のバランス、多様性・流暢性・推論効率の確保といった観点では、いまだ多くの技術的・理論的な課題が残されている。 本研究では、こうした既存手法の課題を克服し、Retrieved Evidence の構造的な特徴と自然言語生成の柔軟性を両立させる新しいフレームワーク「DialFill」を提案している。DialFill では、単に検索知識をそのまま応答に連結・引用するのではなく、Evidence Snippets の中から重要な情報を部分的なアンカーとして活用し、その周辺文脈をマスクすることで、生成モデルに穴埋めを行わせる。これにより、知識の強制的な反映と、流暢かつ多様な応答の両立を目指している。本論文では、DialFill の理論的な基盤、段階的な拡張、応用の可能性、そして残された課題について、多角的に検証している。 第1章では、オープンドメイン対話システムの発展の経緯を詳しく述べている。初期のエンコーダ・デコーダ型モデルから、現在主流の Transformer ベース大規模モデルに至るまでの技術的進化と、それに伴って明らかになった課題を整理し、検索情報を取り込む手法と純粋な生成手法の両方に存在する限界を明示している。これにより、事実の正確さと自然な表現のあいだにあるギャップを浮き彫りにし、本論文で取り組むべき主要な研究課題を明確にしている。 第2章では、対話システムを構成する基本要素として、検索型および生成型のアーキテクチャ、知識の検索と選択、テキストの穴埋め手法、反復制御技術など、関連分野の研究を詳しく調査している。特に、知識ベースを活用したシステム、最新の制御技術や効率化手法についても触れ、本研究で使用する専門用語や記号の統一を図り、以降の章の議論をスムーズに進める準備をしている。 第3章では、DialFill v1 を提案し、対話応答生成を「文章の生成」ではなく「部分穴埋め」の課題として再定義している。DialFill v1 では、検索システムで取得した応答文から重要なエンティティやキーワードを抽出し、それらの周辺文脈を柔軟にマスクし、Transformer 系の生成モデルでマスク部分を補完する方式を採用している。この手法では、Evidence を強制的に応答に含めつつ、表現の多様性と流暢さも確保できる点が特徴である。右側・左側・中央・全体・二段階などの複数のマスキング戦略を実装し、それ			

氏名	薛 強
ぞれが Evidence の反映性、多様性、流暢性にどのように貢献するかを実験で検証している。DailyDialog や Reddit といった実データを用いた結果、多様性指標や BLEU スコアで顕著な向上が見られ、Evidence の反映と自然な生成が両立できる新しい枠組みの有効性を示している。ただし、単一の Retrieved Evidence に依存しているため、検索システムの失敗時には応答全体の品質が大きく低下するという課題も明らかにしている。 第4章では、DialFill v2 を提案している。これは、DialFill v1 の課題を踏まえて、複数の Evidence Snippets やより複雑な知識構造（トリプルや長文パッセージ）に対応するための拡張を行ったものである。Retrieved Evidence の多元化と統合、そしてマルチタスク学習を通じて、キーワード予測・マスク応答ドラフトの作成・対話穴埋めを単一モデル内で統合的に各 Evidence Snippet から重要語句を抽出し、適応的マスク探索アルゴリズムによりマスク処理している。Wizard-of-Wikipedia や OpenDialKG といった大規模知識対話データセットを用いた評価では、応答の正解性、知識の正確性、検索知識の活用率などの指標で、従来手法を上回る成果を確認している。特に、複数の Evidence 統合による冗長性の自動削減や、文脈に応じたテンプレート構築による柔軟な生成制御が可能となった点が新しい成果である。一方で、類似したエンティティが複数取得された場合の応答の反復、文脈に合わない Evidence の混入といった新たな課題も明らかにしている。 第5章では、DialFill v3 の実現に向けた基礎的な検討として、対話応答の生成とその品質判定を統合するための、新しい生成・分類融合メカニズムの設計と検証を目的としている。これまでの DialFill v2 では、Retrieved Evidence の選択やマスク範囲の決定は、主にランダムまたは単純なヒューリスティック手法に頼っており、応答の品質を細かく制御することが難しいという課題があった。一方で、これらのプロセスに外部の分類器や追加のネットワークを導入すれば、応答の品質や知識の統合性を高める効果が期待できるが、その反面、計算コストの増大や推論効率の低下といった実用上の問題も新たに発生してくる。そこで本章では、モデルの語彙空間に「<GOOD>」「<BAD>」といった判別用のトークンを追加し、モデル自身が出力トークンのロジット値を使って、応答の品質を直接判定できる「単一トークン分類機構（JudgerToken）」を提案している。この手法では、外部の分類器を用いずに、生成モデルの内部だけで、応答候補に対する品質スコアの付与、反復の抑制などを同時に実現できる点が特徴である。これにより、複数の応答候補を軽量の計算コストで比較・制御できるだけでなく、単一パスでのリアルタイムな応答生成も可能になる。実験では、代表的な生成・分類モデルと比較し、応答の品質評価、負例の生成と判定戦略の有効性、応答の多様性や反復抑制性能などについて詳細に分析している。その結果、提案手法は従来手法と比べて同等以上の応答品質と反復抑制効果を示し、DialFill v3 の単一エンドツーエンド実装に向けた有力な基盤技術となることを示している。 第6章では、提案する DialFill v3 の全体的な設計と各手法の詳細について述べている。DialFill v3 は、Evidence の選択、エンティティ抽出、テンプレート構築、マスク範囲の最適化、分類メカニズム構築、そして応答生成までの一連の処理を、単一の Transformer モデル上でエンドツーエンドに統合している。これにより、応答生成全体の効率化と知識統合の一貫性向上を実現している。従来はランダムや単純なヒューリスティックな手法で処理されていた Evidence 選択やマスクの最適化を、モデル内部で一括してスコアリングすることで効率化し、知識の反映性と生成効率の両立を実現している。また、すべての処理を単一の Transformer 構造で完結させることで、外部の分類器や追加ネットワークを使う必要がなくなり、パラメータの増加や推論の遅延を防ぐことができ、実際の運用にも適した軽量の構成となっている。実験では、複数の知識対話データセットを用いて、知識の活用度、多様性、流暢さ、生成効率といった自動評価指標において、DialFill v3 が従来手法に比べて一貫した性能向上を示すことを確認している。一方で、Evidence Snippets 間の信頼性の評価や、知識間の矛盾、ユーザ要求への動的な適応、マルチモーダル対応、説明性や責任ある AI 設計といった課題はまだ残されており、今後の発展的な研究テーマとして検討が必要であることを述べている。 第7章では、これまでの研究成果をまとめ、将来に向けた研究の可能性や展望について議論している。 以上のように、本研究は、大規模言語モデルを用いたオープンドメイン対話システムに関して、検索情報を統合的かつ十分に活用できる新たなフレームワーク「DialFill」について研究したものであり、対話 AI における検索情報の完全活用のための深層学習アプローチについて重要な知見を得たものとして価値ある集積である。提出された論文はシステム情報学研究科学位論文評価基準を満たしており、学位申請者の薛強は、博士（工学）の学位を得る資格があると認める。	