

アドホック検索における モデルマージによる破滅的忘却の抑制

谷口 壮平[†] 佐々木泰河^{††} 山本 岳洋^{††} 大島 裕明^{††} 藤田 澄男^{†††}

[†] 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

^{†††} LINE ヤフー株式会社 〒 102-8282 東京都千代田区紀尾井町 1-3

E-mail: [†]{fa22v053,ad24l026}@guh.u-hyogo.ac.jp, ^{††}t.yamamoto@sis.u-hyogo.ac.jp,

^{†††}ohshima@ai.u-hyogo.ac.jp, ^{††††}sufujita@lycorp.co.jp

あらまし 本研究では、アドホック検索タスクにおけるモデルマージ用いた破滅的忘却の抑制効果を検証する。モデルマージは、複数の学習済みモデルのパラメータを統合し、追加の訓練コストを抑えながら各モデルの利点を兼ね備えた新しいモデルを構築する技術である。我々は、追加学習で生じる破滅的忘却という課題をモデルマージによって抑制することで、既存の検索能力を保持したまま追加学習が実現できると考えた。そこで本研究では、まず汎用的な埋め込みモデル（ベースモデル）と、それを代表的なアドホック検索データセットでファインチューニングしたドメイン特化モデルを構築し、両者をモデルマージしたマージモデルとの性能を比較検証した。nDCG@10 による評価の結果、ドメイン特化モデルでは多くの学習対象外ドメインで破滅的忘却が観測されたが、マージモデルではこの低下したすべてのドメインで nDCG@10 スコアが回復した。さらに、一部のドメインではマージモデルが学習対象ドメイン自体の性能も向上させ、モデルマージがアドホック検索タスクにおける破滅的忘却の抑制に有効な手段であることを実証した。

キーワード 情報検索、モデルマージ、破滅的忘却、大規模言語モデル

1 はじめに

アドホック検索タスクは、情報検索分野における基本的なタスクである。これは、ユーザから与えられた自由形式のテキストクエリに対して、事前に用意された文書集合の中からそのクエリに適合する文書を特定し、その適合度の高い順に順位を付けて提示することを目的とする。このタスクの性能評価には、ランキングの上位における関連性の順序の精度を測る nDCG [7] などの評価指標が標準的に用いられる。アドホック検索タスクは、Web 検索やレビュー検索、専門的な論文検索において中心的な役割を果たしている。さらに、大規模言語モデル (LLM) の外部知識を活用する能力を高める技術として注目を集める検索拡張生成 (RAG) [9] にも活用されており、アドホック検索の重要性は一層高まっている。

実際の検索システムの性能を維持・向上させるには、日々生まれる新たな文書や変化するユーザの情報要求に対応するため、モデルを継続的に更新・学習させることが不可欠である。しかし深層学習モデルには、継続事前学習や追加学習によって新たな知識を学習する際に、過去に獲得した知識を忘却してしまう、それに関連するタスクの性能が低下してしまうという減少が課題として存在する。この現象は**破滅的忘却** (Catastrophic Forgetting) [13] として知られており、知識の獲得と既存知識の保持を両立させるうえでの主要な障害となっている。

一方、複数の異なる学習済みモデルを組み合わせ、各モデ

ルの得意分野や獲得した知識を融合し、新たなモデルを構築するモデルマージ [21] という手法が注目されている。特に、モデルのパラメータを直接統合することで、一から LLM を構築するコストや追加学習に必要な計算資源を削減できる点に大きな利点がある。

本研究は、アドホック検索タスクにおいて、モデルマージが破滅的忘却の抑制に有効な手法として機能するかを検証することを目的としている。具体的には、一般的な**ベースモデル**と、これに対し特定ドメインのデータでファインチューニングした**ドメイン特化モデル**を構築し、この2つのモデルを統合した**マージモデル**を生成する。その際、ベースモデルとドメイン特化モデルをマージする比率をマージ係数として定義し、マージ係数を段階的に変化させることで最適な比率を探索する。これら3種類のモデルの性能を、ドメイン特化モデルの構築に使用したドメインとその他のドメインの評価セットを用いて比較分析することで、以下の2点を明らかにする。

RQ1: 破滅的忘却が発生した場合、モデルマージによるその抑制は有効か。

RQ2: モデルマージを行う際、ベースモデルとドメイン特化モデルのマージ係数の変化によって性能 (nDCG スコア) はどのように推移するか。また、最も望ましい性能を示す最適なマージ係数は存在するか。

本研究の主な実験結果は以下の通りである。

RQ1: ドメイン特化モデルでは学習対象外ドメインにおいて

nDCG スコアが低下したが、モデルマージを適用したマージモデルは、評価した全ドメインでこの低下したスコアを回復させることに成功し、破滅的忘却の抑制手法として機能することを示した。

RQ2: マージ係数の変化にともなってモデルの性能は変動し、最適なマージ係数は対象とするドメインごとに異なることが示された。また、マージモデルはファインチューニングにおける過学習を緩和する効果も持つことが明らかになった。

2 関連研究

2.1 LLM によるアドホック検索

アドホック検索タスクを解くための主要なアプローチには、単語の一致に基づく疎検索と、意味的な関連性を捉える密検索の2つがある。疎検索の代表例は TF-IDF [6] や BM25 [14] であり、密検索は LLM を用いてクエリと文書の埋め込みベクトルを生成するアプローチが一般的である。

密検索に用いられる LLM には、クエリと文書を個別にベクトル化するバイエンコーダモデルと、結合してベクトル化するクロスエンコーダモデルなどがある。バイエンコーダモデルはクロスエンコーダモデルよりも一般に精度は劣るものの、事前にインデックスを構築することで高速に検索することができる。

近年では、LLM を活用することにより密検索の性能は向上しており、特に対照学習に基づいて訓練された e5 [20], Contriver [5] といったバイエンコーダモデルが、主要なテストコレクションで優れた性能を示すことが報告されている。本研究では、密検索モデルを対象とし、その中でもバイエンコーダモデルである e5 を用いて実験を行う。

2.2 モデルマージ

モデルマージとは、複数の学習済みモデルの得意分野や獲得した知識を融合し、新たなモデルを構築する技術である。モデルマージには、モデルの重みを直接操作し、線形補間などの手法で統合する重みレベルでのマージと、モデルの内部構造（層）を組み合わせて構築する層レベルでのマージがある。

パラメータ空間でのマージは、基本的に同じアーキテクチャを持つモデルの統合に限定される。代表的な手法として、パラメータの平均化を行う Model Soup [21] のほか、SLERP [16], TIES [22], DARE [8], Task Arithmetic [3] など、多様な手法が提案されている。

本研究で対象とするアドホック検索タスクにおいても、モデルマージの有用性が示されている。我々はこれまでに、汎用的な検索モデルと検索能力を持たないドメイン特化モデルの重みを線形補間によってマージすることで、そのドメインに対して高い検索性能を持つモデルを構築できることを示してきた [25] [15]。

本研究ではこのモデルマージの有効性を、元のモデルの検索能力の維持という観点に着目して検証し、アドホック検索タスクへ適用する。

2.3 破滅的忘却

破滅的忘却 (Catastrophic Forgetting) は、ニューラルネットワークが新しいタスクを学習する際に、以前学習した知識や性能を失う現象である。この現象は、ネットワークの重みが複数のタスクで共有され、新しい学習による重み更新が以前の学習内容を上書きしてしまうことが原因であり、現代の LLM や深層学習においても継続学習・ファインチューニングにおける主要な課題となっている。

情報検索の分野においても、破滅的忘却の問題が報告されている。Lovón-Melgarejo らは、深層学習ベースのモデルが異なるドメイン間で順次学習を行う際に破滅的忘却が発生することを実験的に示し、特に意味的マッチング能力の高いモデルほど、新しいドメインに適応した際に以前のドメインの知識を失う傾向がより強く現れることを明らかにした [10]。さらに、正則化ベースのアプローチでこの現象に対処することが可能であることを示したが、このアプローチは学習プロセス自体の変更を伴うという課題が残る。

これに対し、学習プロセスそのものを変更せず、モデルマージによって破滅的忘却を抑制するアプローチも提案されている。例えば岩川らは、LLM の翻訳タスクにおいて、継続事前学習の前後のモデルを線形補間によりマージすることで、一般タスクの性能を維持しながら翻訳能力が向上したモデルを構築可能であることを報告している [24]。また、アドホック検索の領域においても、Han らは時間的な制約を含むクエリ (Temporal IR) への対応と汎用的な検索能力を両立させる手法として、モデルマージの有効性を示している [4]。しかし、この研究は「2015 年に」といった時間的な表現を含む限定的な状況に特化したものであり、検索タスク全般におけるドメイン対応と破滅的忘却の抑制については十分に検証されていない。そこで本研究では、モデルマージによる手法をより広範なアドホック検索タスクに応用し、ドメイン特化による性能向上と、学習対象外ドメインにおける破滅的忘却の抑制をより一般的な状況下で両立できるかを図る。

3 実験設定

3.1 実験の概要

本研究は、汎用的検索モデルを特定ドメインでファインチューニングする際に生じる、元のモデルが獲得していた検索能力や性能の低下という課題に対処する。具体的には、元の汎用的検索モデルの検索能力や性能の維持と、特定ドメインにおける検索性能の向上を両立させる、パラメータ空間でのモデルマージによるアプローチの有効性を検証する。

検証のため、以下の3種類のモデルを用意し、その性能を評価する。

ベースモデル: 汎用的なドメインで学習された検索モデル。本研究では e5-base-v2¹を用いる。

ドメイン特化モデル: ベースモデルを特定ドメインでファイン

1: <https://huggingface.co/intfloat/e5-base-v2>

表 1 テストコレクションの統計情報.

	クエリ			コーパス
	訓練	検証	テスト	
NFCorpus	2,590	324	323	3,633
SciFact	648	161	300	5,183
FiQA	5,500	500	648	57,638
FEVER	109,810	6,666	6,666	5,416,568
HotpotQA	85,000	5,447	7,405	5,233,329

チューニングしたモデル.

マージモデル: ベースモデルとドメイン特化モデルをパラメータ空間で統合した, 本研究の提案モデル.

これらのモデルを使用して, マージによる破滅的忘却の抑制とドメイン特化性能の向上の効果を, 後述するテストコレクションを用いて定量的に評価する.

3.2 テストコレクション

本実験では, アドホック検索タスクの評価ベンチマークとして広く利用されている BEIR (Benchmarking-IR) [17] を用いた. BEIR は医療検索, 金融検索, 質問応答, 事実検証など多岐にわたるドメインとタスクをカバーしており, 異なるドメインやデータセット間での検索モデルの性能を評価することができる. この BEIR のフレームワークから, 本研究の目的である元のモデルの検索能力の維持とドメイン特化性能の向上を検証するため, 以下の 6 つのデータセットを抽出した.

特定ドメイン検索テストコレクション

- **FEVER** [18]: 自然文の主張の真偽を検証するため, Wikipedia の記事から根拠となる文書を検索する能力を評価するデータセット.
- **FiQA** [12]: 金融・経済分野の質問と回答のデータを含み, 特定分野の専門用語と知識理解の能力を評価するデータセット.
- **HotpotQA** [23]: 複数の Wikipedia 記事から情報を統合し, 多段階の推論を必要とする複雑な質問に回答する能力を評価するデータセット.
- **NFCorpus** [2]: 栄養学及び健康科学分野の専門的な文書とクエリからなり, 専門的な情報の検索能力を評価するデータセット.
- **SciFact** [19]: 科学論文の要旨に基づき, 特定の主張の根拠となる文書を検索するタスクを評価するデータセット.

モデルの学習と決定には, それぞれ各データセットの訓練および検証データを用いる. ただし, SciFact には公式で検証データが存在しないため, 訓練データをランダムに 8:2 の比率で分割し, それぞれ訓練, 検証データとして利用した.

3.3 ドメイン特化モデルの構築

3.3.1 学習対象ドメインと学習対象外ドメイン

ドメイン特化モデルは, ベースモデルを前節で示した特定ドメインデータセット (FEVER, FiQA, HotpotQA, NFCorpus, SciFact) の訓練データを用いてファインチューニングするこ

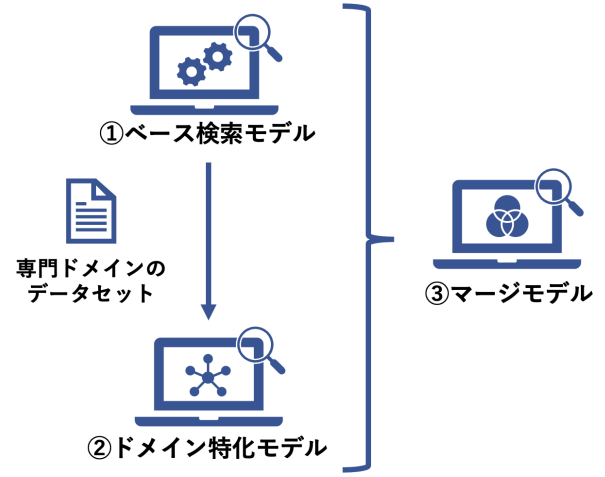


図 1 モデルマージ構築の流れ.

とで構築する. 本研究では, そのドメイン特化モデルを構築するために使用した訓練データのドメインを学習対象ドメイン, それ以外のドメインを学習対象外ドメインと呼ぶ. 例えば, FEVER をファインチューニングすることで構築したドメイン特化モデルにおいては, 学習対象ドメインは FEVER であり, その他のドメインが学習対象外ドメインとなる. また, 一般ドメインとして汎用的な検索能力の維持度合いを評価するためのデータセットである MS MARCO [1] を使用し, これはすべてのドメイン特化モデルの学習対象外ドメインとした.

3.3.2 ファインチューニング

ファインチューニングには, 公式で提供されている適合文書 (正例) だけでなく, 非適合文書 (負例) も必要となる. 本研究では, BM25 を用いて検索した上位 30 件のラベリングされていない文書を, モデルにとって判別が困難な負例 (ハードネガティブ) として利用した. 学習には, 検索モデルの訓練フレームワークである Tevatron [11] を用いた. ハイパーパラメータは公式の設定を参考に調整し, エポック数は検証データにおける nDCG@10 の値を基準に決定した. 具体的には, 最大 20epoch の学習過程において, 検証データでの nDCG@10 が最大となったチェックポイントでのモデルをドメイン特化モデルとして採用した.

3.4 マージモデルの構築

本研究の提案モデルであるマージモデルは, ベースモデルとドメイン特化モデルの 2 つをパラメータ空間の線形補完で統合することによって構築する. マージモデルを構築する流れを図 1 に示す. 線形マージは, モデルを構成するすべてのパラメータに対して, 同一の比率で線形和を求める. モデルの任意のパラメータ i におけるマージ後の重み $\theta_i^{(\text{merged})}$ は, ベースモデルの対応する重み $\theta_i^{(\text{base})}$ と, ドメイン特化モデルの対応する重み $\theta_i^{(\text{domain})}$ を用いて, 以下の式で表現される.

$$\theta_i^{(\text{merged})} = (1 - \alpha)\theta_i^{(\text{base})} + \alpha\theta_i^{(\text{domain})}$$

ここで, α ($0 \leq \alpha \leq 1$) は全てのパラメータに共通して適用されるハイパーパラメータであり, 本研究ではマージ係数と呼

ぶ。実験においては、RQ1の検証時には α を0.5に固定し、両モデルの重みの平均を取ることでモデルを構築し、RQ2では最適な α の値を探索的に検証する。

マージモデルの構築により、ベースモデルの検索能力を維持しつつ、ドメイン特化モデルが獲得した特定ドメインでの検索能力を統合し、破滅的忘却を抑制した高性能な検索モデルの実現を目指す。

4 実験結果

本節では、1章で述べた2つのリサーチクエスチョンに対する実験結果を述べる。評価指標にはnDCG@10を用いた。また、2つのモデル間の性能の比較には有意水準を5%とした対応のある t 検定を用いた。

4.1 RQ1: モデルマージによりアドホック検索における破滅的忘却の抑制は有効か？

まず、ドメイン特化モデルを構築する過程において破滅的忘却が発生するかを確認した。具体的には、ベースモデルと各ドメイン特化モデルの性能を、学習対象ドメインおよび学習対象外ドメインのそれぞれで評価・比較した。その結果を表2に示す。表の最左列はドメイン特化モデルの構築（ファインチューニング）に用いたデータセットを示し、各列は評価対象のデータセットを示している。それぞれの数値について、ベースモデルのスコアに対して有意に上回っているものは太字、有意に下回っているものは下線で示している。

FEVERのドメイン特化モデルに着目すると、学習対象外ドメインにおけるnDCG@10のスコアはすべてベースモデルに対して有意に低下しており、破滅的忘却が発生しているといえる。その他のドメイン特化モデルに関しても、学習対象外ドメインのほとんどで性能が低下しており、アドホック検索タスクにおいても破滅的忘却が発生することが確認された。

次に、3.4節で構築したマージモデルの性能を評価し、破滅的忘却の抑制効果を検証した。表2のFEVERにおけるマージモデルの結果を見ると、ドメイン特化モデルで得られた学習対象ドメインでの性能を維持しつつ（ドメイン特化モデル：0.887、マージモデル：0.889）、学習対象外ドメインでの性能をベースモデルと同等まで回復させることに成功している。たとえば、MS MARCOではベースモデルの0.703に対し、マージモデルは0.702を記録している。他のすべてのドメインにおいても同様の傾向が見られることから、アドホック検索タスクにおいて、モデルマージは破滅的忘却を抑制するための有効な手法であることが示唆される。

4.2 RQ2: モデルのマージ割合と破滅的忘却の関係

マージモデル構築時の α を0.0から1.0まで0.1刻みで変化させてその性能の推移を比較することで、それぞれのドメイン特化モデルについて最適なマージ割合を分析した。図2は、ベースモデルにおけるnDCG@10を基準とした相対スコアの推移を示している。ここで、相対スコアとは、ベースモデル（ $\alpha = 0.0$ ）でのnDCG@10を S_{base} 、あるモデルのnDCG@10

を S_{model} としたとき、以下の式で求められる値である。

$$REL_{model} = \frac{S_{model} - S_{base}}{S_{base}}$$

FEVERを学習対象としたグラフに着目すると、 $\alpha = 0.2$ 付近までは、学習対象ドメインだけでなく学習対象外ドメインでも、ベースモデルを上回る性能向上が確認できる。しかし、 $\alpha = 0.3$ を超えてドメイン特化モデルの影響が大きくなるにつれて、学習対象ドメインのスコアは上昇し続ける一方で、学習対象外のドメインはすべて低下している。このことから、特化モデルの重みが大きくなるにつれて破滅的忘却が顕著になることが読み取れる。

他のドメインにおいても、 α がある値より大きくなるにつれて学習対象ドメインのスコアは上昇し、学習対象外のドメインでのスコアが低下した。 α を中程度の値（0.4～0.6）に設定することで、単一ドメインへの過学習を避けながら、モデル全体の検索能力を底上げできることが示唆される。

ここで、複数のドメインにおける性能を1つの指標で総合的に比較分析するため、調和平均に基づいた新たな手法を導入した。調和平均には算術平均とは異なり、極端に低い値に強く引きずられる性質がある。そのため、特定ドメインで突出した高スコアを記録しても、別のドメインで破滅的忘却による性能低下が生じていれば、総合スコアは低く算出される。この性質は、本研究の目的である破滅的忘却の抑制を評価するのに適している。評価対象とするドメイン群 D の一般的な調和平均 H_D は次式で定義される。

$$H_D = \frac{|D|}{\sum_{d \in D} \frac{1}{S_d}}$$

$|D|$ は評価対象とするドメインの総数を、 S_d はドメイン d における評価スコアを表す。

しかし、通常の調和平均ではドメインの数が増えるほど各ドメインの影響力が相対的に低下するため、ファインチューニングの目的である学習対象ドメインにおける検索能力の向上がスコアに反映されにくくなる課題がある。そこで本研究では、学習対象ドメインの影響力を高めた新たな指標 H'_D を導入する。まず、学習対象外ドメイン群 $D \setminus d_{tgt}$ における調和平均 $H_{D \setminus d_{tgt}}$ を算出し、それと学習対象ドメインのスコア $S_{d_{tgt}}$ をさらに調和平均することで、総合スコアとする。

$$H'_D = \frac{2}{\frac{1}{H_{D \setminus d_{tgt}}} + \frac{1}{S_{d_{tgt}}}}$$

これにより、ドメインの数にかかわらず、学習対象ドメインの影響力を学習対象外ドメイン群全体と同等まで高めており、学習対象ドメインでの検索能力の向上を重視した評価が可能となる。学習対象ドメインでの検索能力を獲得し、かつ学習対象外ドメインでの性能を維持できているかを判定する評価基準となっている。

この指標 H'_D を用いて性能推移を評価した結果を図3に示す。どのドメインにおいても、 α が0.0（ベースモデル）や1.0

表 2 ベースモデル, ドメイン特化モデルおよびマージモデルの比較. ベースモデルのスコアに対して有意 ($p < 0.05$) に上回っているものは太字, 下回っているものは下線で示している.

学習データセット	モデル	評価データセット					
		MS MARCO	FEVER	FiQA	HotpotQA	NFCorpus	SciFact
ベース		0.703	0.850	0.399	0.692	0.354	0.720
FEVER	ドメイン特化	<u>0.652</u>	0.887	<u>0.359</u>	<u>0.622</u>	<u>0.328</u>	<u>0.656</u>
	マージ	0.702	0.889	0.403	<u>0.683</u>	0.349	0.719
FiQA	ドメイン特化	0.668	<u>0.798</u>	0.414	<u>0.679</u>	0.364	0.704
	マージ	0.705	0.851	0.427	<u>0.690</u>	0.359	0.719
HotpotQA	ドメイン特化	<u>0.613</u>	<u>0.755</u>	<u>0.320</u>	0.692	<u>0.320</u>	<u>0.651</u>
	マージ	0.682	<u>0.843</u>	<u>0.380</u>	0.729	0.354	0.719
NFCorpus	ドメイン特化	<u>0.612</u>	<u>0.582</u>	<u>0.271</u>	<u>0.606</u>	0.385	<u>0.667</u>
	マージ	0.694	<u>0.812</u>	<u>0.385</u>	<u>0.677</u>	0.386	0.735
SciFact	ドメイン特化	0.686	<u>0.804</u>	0.393	<u>0.683</u>	0.360	0.788
	マージ	0.705	<u>0.839</u>	0.403	0.691	0.364	0.773

マージモデルの相対NDCG@10スコア比較 (学習データセット別)

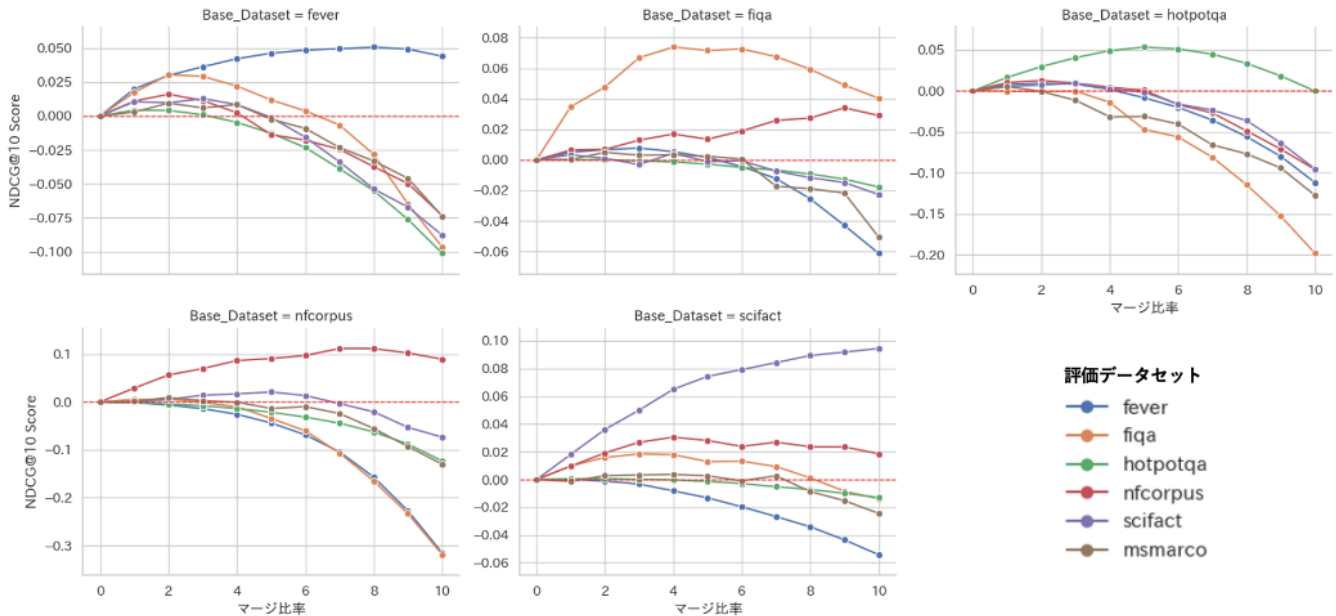


図 2 マージ係数とそのモデルの相対 nDCG@10 スコアの推移.

(ドメイン特化モデル) の設定の場合と比較して, 0.4 ~ 0.6 の範囲でマージを行ったモデルが優れた総合スコアを記録した. また, あるドメインでの最適な α はすべて 0.3 ~ 0.7 の範囲に収まっていた. 学習対象ドメインでの検索能力の獲得と破滅的忘却の抑制を両立させるという点で, モデルマージは有効な手段であるといえる. 特に, その最適なマージ係数がドメインによって異なるということは, 学習対象となるドメインとベースモデルがもともと持っている知識との親和性や, データセットの特性に依存するということを示唆している.

図 2 において, 一部のドメインではマージモデルが学習対象ドメインにおいてもドメイン特化モデル ($\alpha = 1.0$) を上回る相対スコアを示している. これは, ドメイン特化モデルがファイ

ンチューニングの過程で訓練データに過学習している可能性を示唆している.

この仮説を検証するため, ファインチューニングの各エポックにおけるモデルの性能推移を分析した. 図 4 に, 各ドメインでのファインチューニングにおけるエポックごとの総合スコア H'_D の推移を示す. 緑の線は, 検証データでの nDCG@10 が最大となったエポックとして選択されたドメイン特化モデルを示している. 多くのドメインにおいて, 選択されたエポック以降も総合スコアが向上する場合もあるが, FEVER や HotpotQA, NFCorpus では, 選択されたエポックよりも前のエポックの方が総合スコアが高い. これは, 検証データでの単一ドメイン性能を最大化する基準が, 必ずしも複数ドメインにわたる汎化性

マージモデルのNDCG@10スコア比較 (学習データセット別)

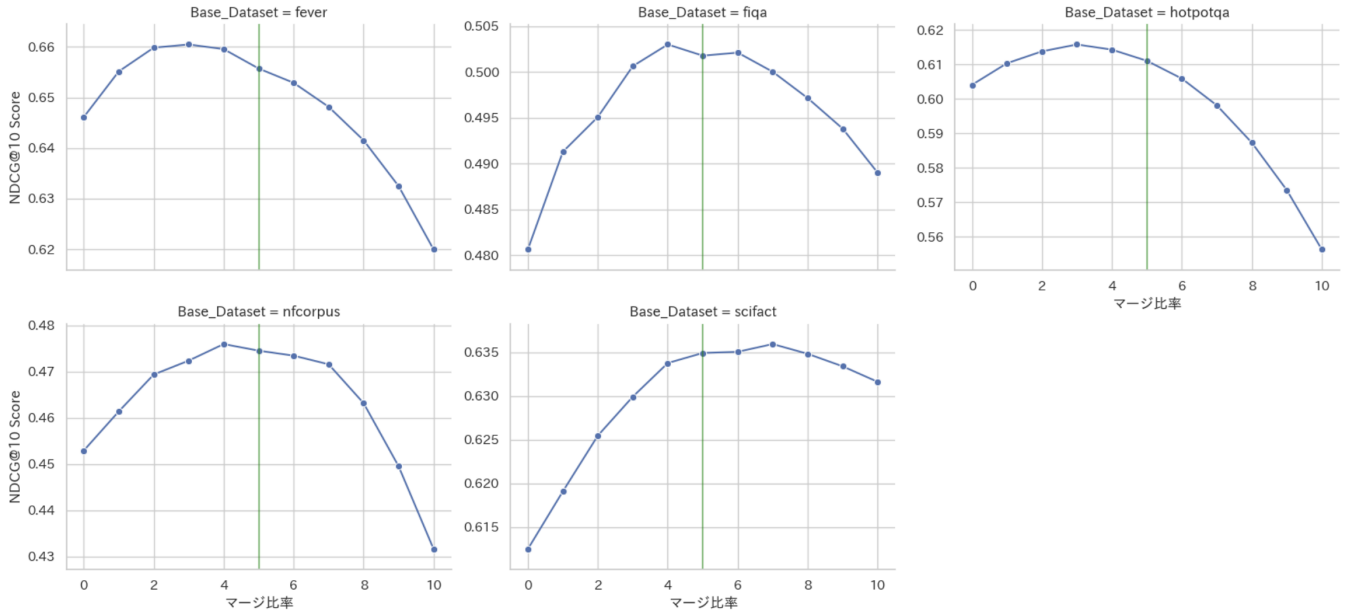


図 3 マージ係数とそのモデルの総合スコアの推移。緑の線は $\alpha = 0.5$ のモデルでの評価。

特化モデル構築時のNDCG@10調和平均スコア (学習データセット別)



図 4 ドメイン特化モデルのファインチューニングにおける学習エポックとそのモデルの総合スコアの推移。緑の線はドメイン特化モデルとして選択したモデルでの評価。

能の最大化とは一致しないことを示している。

この結果から、検証データで最良の性能を示したエポックを選択したドメイン特化モデルは、学習対象ドメインのテストデータにおいて必ずしも最適ではなく、訓練データへの過適合が生じていることが示された。一方、マージモデルはベースモデルの汎化能力を取り込むことで、この過学習を緩和し、結果として学習対象ドメインにおいても性能向上につながったと解釈できる。

したがって、モデルマージは破滅的忘却の抑制だけでなく、

ファインチューニングにおける過学習の緩和という副次的な効果も持つことが示唆される。これは、実用的なアドホック検索システムの構築において、モデルマージが単なる知識統合の手段を超えて、モデルの汎化性能を向上させる正則化的な役割も果たし得ることを意味している。

したがって、実用的なアドホック検索システムを構築する際には、単に特定のドメインデータでフルファインチューニングを行うのではなく、本研究で導入した H_D のような指標を用いて、目的のドメイン性能と汎用性のトレードオフを精緻に調整

することが不可欠であると考えられる。本実験の結果は、モデルマージがその調整を柔軟に行うための強力なフレームワークになり得ることを示している。

5 まとめと今後の課題

本研究では、アドホック検索タスクにおいて、モデルマージが破滅的忘却の抑制に有効であるかを検証した。具体的には、汎用的な検索モデル（ベースモデル）と特定ドメインでファインチューニングしたドメイン特化モデルを線形補間により統合したマージモデルを構築し、その性能を BEIR ベンチマークの 6 つのドメインで評価した。

実験の結果、ドメイン特化モデルでは学習対象外ドメインの多くにおいて nDCG@10 スコアが有意に低下し、アドホック検索タスクにおいても破滅的忘却が発生することを確認した。一方、マージモデルは評価したすべてのドメインにおいて、この低下したスコアをベースモデルと同等まで回復させることに成功し、モデルマージが破滅的忘却を抑制する有効な手法であることが示唆された。また、マージ係数を変化させた分析により、0.4 ~ 0.6 の範囲でマージを行うことで、特定ドメインでの検索性能向上と汎用性の維持を両立できることを明らかにした。これらの結果は、アドホック検索システムの構築において、フルファインチューニングの代替または補完手段としてモデルマージが有効であることを示している。最適なマージ係数はドメインによって異なり、学習対象ドメインとベースモデルの知識との親和性に依存することが示唆された。

今後の課題として、本研究では線形補間によるマージ手法を採用したが、TIES や DARE といった他のマージ手法の適用可能性を検証する必要がある。

文 献

- [1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *arXiv preprint arXiv:1611.09268*, 2016.
- [2] Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. A Full-Text Learning to Rank Dataset for Medical Information Retrieval. In *Proceedings of the 38th European Conference on IR Research*, pp. 716–722, 2016.
- [3] Marco Braga, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. Investigating Task Arithmetic for Zero-Shot Information Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2738–2743, 2025.
- [4] SeungYoon Han, Taeho Hwang, Sukmin Cho, Soyeong Jeong, Hoyun Song, Huije Lee, and Jong C. Park. Temporal Information Retrieval via Time-Specifier Model Merging. pp. 1–13, 2025.
- [5] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv preprint arXiv:2112.09118*, 2021.
- [6] Thorsten Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of the 14th International Conference on Machine*

- Learning*, Vol. 97, pp. 143–151, 1997.
- [7] Kalervo Järvelin and Jaana Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, Vol. 20, pp. 422–446, 2002.
- [8] Le Yu and Bowen Yu and Haiyang Yu and Fei Huang and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In *41st International Conference on Machine Learning*, pp. 5775–5775, 2024.
- [9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9459–9474, 2020.
- [10] Jesús Lovón-Melgarejo, Laure Soulier, Karen Pinel-Sauvagnat, and Lynda Tamine. Studying Catastrophic Forgetting in Neural Ranking Models. In *Advances in Information Retrieval*, Vol. 12656, pp. 375–390, 2021.
- [11] Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samanthan Zhan, Jamie Callan, and Jimmy Lin. Tevatron 2.0: Unified Document Retrieval Toolkit across Scale, Language, and Modality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4061–4065, 2025.
- [12] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. WWW’18 Open Challenge: Financial Opinion Mining and Question Answering. In *Companion Proceedings of the The Web Conference 2018*, pp. 1941–1942, 2018.
- [13] Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation*, Vol. 24, pp. 109–165, 1989.
- [14] Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 4, pp. 333–389, 2009.
- [15] Taiga Sasaki, Takehiro Yamamoto, Hiroaki Ohshima, and Sumio Fujita. Effect of Model Merging in Domain-Specific Ad-hoc Retrieval. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5208–5212, 2025.
- [16] Ken Shoemake. Animating Rotation with Quaternion Curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 245–254, 1985.
- [17] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Vol. 1, 2021.
- [18] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.
- [19] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020.
- [20] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei.

Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

- [21] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 23965–23998, 2022.
- [22] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. TIES-MERGING: resolving interference when merging models. In *Advances in Neural Information Processing Systems*, Vol. 36, pp. 7093–7115, 2023.
- [23] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.
- [24] 岩川光一, Haocheng Zhu, 鈴木潤, 永田昌明. モデルマージを用いた LLM 翻訳における破滅的忘却の抑制. 言語処理学会 第 31 回年次大会 発表論文集, pp. 3378–3383, 2025.
- [25] 佐々木泰河, 山本岳洋, 大島裕明, 藤田澄男. モデルマージによる複数ドメインに対応した単一密検索モデルの構築. 情報処理学会 第 160 回情報基礎とアクセス技術研究会 (SIG-IFAT) 研究報告, 2025.