# 敵対的攻撃に対する密検索モデルの脆弱性分析

薄羽　皐太† 　加藤　誠†† 　藤田　澄男†††

† 筑波大学 大学院人間総合科学学術院　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
††† ヤフー株式会社　〒 102–8282 東京都千代田区紀尾井町 1-3
E-mail: †k-ush@klis.tsukuba.ac.jp, ††mpkato@acm.org, †††sufujita@yahoo-corp.jp

**あらまし**　本論文では、大規模言語モデルをベースとした密検索モデルが、敵対的な攻撃に対してどの程度脆弱であるかを調査する。近年の密検索モデルでは、大規模言語モデルでの入力トークン数の制限から、長い文書は一定の長さで分割した上で、一部の高適合な文章の適合度を文書全体の適合度として用いる。本論文では，この密検索モデルにおける適合度計算のフレームワークを利用した敵対的な攻撃手法を提案し，攻撃の適応を通して既存の密検索モデルの脆弱性を評価する．具体的な攻撃手法として，ランキング中の上位 1 位に位置する文書を文章に分割して、高適合であった文章を推定する．そして，その文章を下位の文書に挿入することで攻撃を行う。実験では文書検索データセットを用い，攻撃手法に対する既存の密検索モデルの脆弱性をと伝統的な検索モデルとの比較を通して評価した。

**キーワード**　検索モデル, 敵対的攻撃, 深層学習,beyond accuracy

## 1　は じ め に

アドホック検索は，クエリを入力として，クエリと関連する文書を文書集合から探し出すタスクである．近年の大規模言語モデルの登場により，深層学習ベースのアドホック検索モデルも急速に発展してきている．現在では事前に文書の密ベクトル表現を計算しておくことで，推論時にはクエリのベクトル表現の計算と，ベクトルの類似度の計算だけに抑えられる密検索が注目を集めている．

こうした深層学習モデルを実際に運用していくためには，モデルが安定した動作をすることが求められるが，深層学習モデルはしばしば敵対的な攻撃に対して脆弱であることが報告されている．敵対的な攻撃とは，通常の入力に僅かな摂動を加えることで，深層学習モデルの出力を意図的に変化させることを指す．Szegedy らは画像分類タスクにおいて，人が認知できないほど僅かな摂動を入力画像に加えることで，意図的に入力を誤らせることができることを示した [1]．敵対的サンプルと呼ばれるこの悪意のある入力の研究は画像分類の分野から始まり，近年では自然言語処理分野での研究も盛んに行われている．情報検索モデルが敵対的な攻撃に対して脆弱であると，あるクエリについて恣意的に文書の順位を操作することが可能になる。これはウェブ検索のシナリオにおいて，攻撃者が自身の運営するサイトへの流入を増やすために，関連性の低いクエリについて恣意的に自身のサイトの順位を向上させるために悪用される恐れがある [2]. 近年注目されている密検索モデルは，文書の一部分のみに特定のクエリについて高適合な文章を挿入する攻撃に対して脆弱である可能性がある．密検索モデルでは，モデルの入力制限のため，長い文書は複数の文章に分割し，最も高適合であった文章の適合度を利用することが行われている [3, 4]. したがって，一部分だけ高適合になるように文章を挿入することで，文書全体の適合度も向上してしまう可能性が考えられるが，密検索モデルの敵対的攻撃に対する脆弱性の評価はまだ十分に行われていない．

本論文では，密検索モデルの脆弱性の調査を行う．密検索モデルにおける適合度の集約を利用した密検索モデルへのシンプルな攻撃手法を提案し，提案手法を密検索モデルに対して適応することで脆弱性を評価する．密検索モデルに対して，単純な攻撃手法によって文書の順位を向上させることができるのであれば，密検索モデルは脆弱であると言える．特に，短い文章の挿入によって効果的に文書の順位が向上する場合は敵対的攻撃の検知も難しくなり，攻撃に対応することがより困難になる．

密検索モデルに対して敵対的攻撃をするために，検索結果として出力されるランク付けされた文書リストを利用する．文書リスト中の高適合であった文書は，少なくともその文書中のいずれかの文章が高適合であったことを意味する．そのため，高適合であった文書中から，高適合であったと思われる箇所を推定することができれば，その箇所の文章を摂動として抽出し，任意の文書に付け加えることでその文書の順位を向上させることができるのではないかと考えた．高適合であった文書として，検索結果中の最上位に位置していた文書を利用し，摂動の抽出に利用する．高適合であった文章の推定では，ランダムに決定する手法や，TextRank を使った中心性の高い文章を利用する手法，BM25 によってクエリとの適合度を計算して高適合であった文章を利用する手法を提案する．

実験では，近年の密検索モデルを対象に，提案する攻撃手法の性能の評価を行なった．データセットにはアドホック文書検索タスクである TREC Robust 2004 のデータセットを利用した．密検索モデルには ANCE を用い，ベースラインの検索モデルとして BM25 を用いた．文書中から高適合である文章を推定する手法のそれぞれを用いた場合の攻撃性能を比較した．さらに，伝統的なスパミングであるクエリ追加による攻撃を ANCE

に適応し，ANCE の脆弱性を評価した [5]．その結果，提案する攻撃手法に対して ANCE は BM25 より堅牢であることが判明した．一方で，BM25 との比較において，クエリ追加による攻撃は部分的に ANCE に対してより効果的であった．

この論文における我々の貢献を以下に示す：(1) 密検索フレームワークを利用した密検索モデルに対するシンプルな攻撃手法を提案した．(2) 実験を行い，提案した手法を用いて密検索モデルの脆弱性を評価した．(3) 伝統的なスパム手法に対する密検索モデルの脆弱性を評価した．

本論文の構成は以下の通りである．2 節では深層学習モデルに対する敵対的攻撃に関する関連研究について述べる．3 節では問題設定を説明し，密検索モデルへの攻撃手法について述べる．4 節では実験結果を示す．最後に，5 節では今後の課題と共に本論文の結論を述べる．

## 2 関 連 研 究

### 2.1 敵対的サンプル

深層学習技術の発展と共に，深層学習モデルの脆弱性の研究も進んできている．初期には Computer vision の分野で入力に微小な変化を加えることで，対象の深層学習モデルの出力を意図的に誤らせることができることが明らかになった [1, 6]．このような，出力を意図的に誤らせる入力は敵対的サンプルと呼ばれ，テキストを含むその他の分野にも応用されてきている [7–9]．敵対的攻撃の問題設定は，攻撃対象のモデルや環境の知り得ることに基づいて，ホワイトボックスとブラックボックスに分けられる．

攻撃対象のモデルに直接アクセスできる問題設定はホワイトボックスと呼ばれ，モデルのパラメータの情報を利用し，損失が大きくなるように摂動を生成する．ブラックボックスではモデルの出力は取得可能だが，モデルのパラメータなどその他の情報については知り得ない設定である．Jin らはテキスト分類タスクにおいて，BERT を対象にしたブラックボックス設定下での攻撃手法を提案し，予測精度の高い分類モデルであっても敵対的サンプルに対して脆弱であることを示した [8]．

### 2.2 アドホック検索タスクでの敵対的攻撃

アドホック検索タスクにおける敵対的攻撃手法も登場してきている．Raval と Verma はアドホック検索モデルを対象に，ブラックボックス設定下で遺伝的アルゴリズムを用いて敵対的サンプルを作成し，文書の順位を下げる攻撃を行なった [10]．Wang らは，アドホック検索モデルを対象に，文書の順位を下げる攻撃だけでなく，向上させる攻撃も行い，BERT ベースの検索モデルの脆弱性を示した [11]．敵対的サンプルの生成はホワイトボックス設定で行い，生成される敵対的サンプルの分析を通して検索モデルの振る舞いを分析した．ブラックホックス設定下で攻撃を行なった研究では，攻撃対象の検索モデルの出力を模倣するような検索モデルを構築する手法が挙げられる [12, 13]．これらの手法では，作成した模倣モデルへの攻撃を，攻撃対象のモデルに転移させることで攻撃を行い，攻撃手法の性能とモデルの脆弱性を評価した．

これらのアドホック検索タスクにおける敵対的攻撃手法は，敵対的サンプルの生成に大きな計算コストがかかる．特に，ブラックボックス設定下で敵対的サンプルを生成する手法では，最初に攻撃対象の検索モデルを模倣するような検索モデルを構築する必要がある．本論文では，クエリと文書を別々にエンコードしてそれぞれのベクトル表現の類似度から適合度を計算する密検索モデルを対象とし，密検索のフレームワークを利用したシンプルな攻撃手法を提案する．そのため，本論文で提案する手法では攻撃対象の検索モデルを模倣するような検索モデルを構築することは行わない．

### 2.3 ウェブスパム

ウェブ検索において，あるウェブページの順位を不当に操作する行為や操作された文書の検出や対策は長年研究されており，Gyongyi と Garcia-Molina はこのようなスパミング手法を Term spaming と Linking spamming に分類した [5]．term spamming はクエリに含まれる語を，ウェブページのボディや URL に挿入することで，あるクエリに対して順位を向上させる技術である．また他のウェブページからフレーズや文をコピーして切り貼りするスパムも発見されており，その検出手法も提案されている [5, 14, 15]．

本研究は，ランキングの順位を不当に上昇させる目的で文書を操作する手法を扱う点で，これらの研究と類似している．しかし，これらの研究では，スパミングの対象となる検索モデルとして，TFIDF や BM25 といった語彙的な類似度によって適合度を計算するモデルを前提とおり，本研究では密検索モデルを対象としている点で異なる．密検索モデルでは，クエリと文書を密ベクトルに埋め込むことで，意味的な類似度による適合度の計算が可能になったと考えられており [16, 17]，単純なテキストの追加に対する密検索モデルの脆弱性を扱った研究は見られない．本研究では密検索モデルにおいて一般的な適合度計算方法を利用した攻撃手法を提案する．

## 3 提 案 手 法

### 3.1 問 題 設 定

アドホック検索タスクでは，$D$ を文書集合とし，与えられたクエリ $q$ に基づいて，$D$ に含まれる文書 $d_i$ の適合度 $s_i = f(q, d_i)$ を検索モデル $f$ によって推定し，適合度に基づいてランク付けされた文書リスト $l = (d_1, d_2, \ldots, d_N)$ を得る．ただし，$s_1 \geqq s_2 \geqq \ldots \geqq s_N$ である．本論文では，$l$ と $q$ と順位を向上させたい文書 $d_t$ $(1 \leqq t \leqq N)$ が与えられた時，摂動を追加する関数 $m$ を用いて摂動 $p = m(l, q)$ を得る．$d_t$ と摂動 $p$ から敵対的サンプルを生成する関数 $h$ を用いて敵対的サンプル $d'_t = h(p, d_t)$ を生成し，$d'_t$ の順位を向上させる問題を解く．ここでは $h(p, d) = p \oplus d$ とする．ただし，$\oplus$ は文字列の結合を意味する．

現実のアドホック検索モデルの利用において，検索モデルのパラメータにアクセス可能であることは稀である．そのため，

敵対的サンプルの生成時に $f$ に直接アクセスすることはできず，クエリとその検索結果である文書のランキングのみ利用可能なブラックボックス設定で行う．

### 3.2 提案手法

本論文で提案する攻撃手法の概要を図 1 に示す．密検索モデルでは，文書とクエリをそれぞれエンコードし，ベクトル表現を得る．そうして得られたベクトル表現の類似度を適合度として計算し，ランキングに用いる．文書とクエリをエンコードする際に，近年では Transformer をベースとした密検索モデルが多く提案されているが，このような密検索モデルでは入力可能な文字列の長さには制限がある．そのため，長い文書をエンコードする際には，文書を複数の文章に分割した上でそれぞれの文章をエンコードする．複数のベクトル表現が得られた場合，それぞれのベクトル表現とクエリのベクトル表現の類似度を計算し，次のように類似度の最大値を文書の適合度として利用することがよく行われる [3, 4, 18, 19].

$$s_i = \max_{p \in P_i} f(q, p)$$
$$f(q, p) = \mathrm{sim}(E(q), E(p))$$
$$P_i = H(d_i, k)$$

ただし，$E$ はエンコーダであり，$H$ は文書を $k$ の長さの文章の集合 $P_i$ に分割する関数である．密検索において各文章の適合度計算は独立しているため，文書を複数の文章に分割した際に，どれか 1 つの文章の適合度が高くなれば，文書の適合度も高くなる．本論文ではこのような文書の一部を使った適合性度計算を利用し，検索結果中の高適合であった文書 $d^*$ から高適合であった文章 $p^*$ を推定し，$p^*$ を $d_t$ に加えることで，$d_t$ の順位を向上させる．高適合であった文章を推定する関数 $g$ を用いて，$d'_t$ を次のように生成する．

$$p^* = g(d^*)$$
$$d^* = d_1$$

このようにして得られた $d'_t$ は $d_t$ と置き換えることで，敵対的攻撃を行う．最上位の文書を利用するため，検索モデルと同一のトークナイザを用いて文書の分割を行い，最大値の適合度を得た文章を摂動として用いた場合には，$d'_t$ は 1 位に順位付けられる．ただし，トークナイザが異なる場合や分割する長さが異なる場合,$p^*$ が実際に高適合を得た文章と異なっていた場合にはその限りではない．

#### 3.2.1 高適合文章推定

検索結果から各文書の高適合であった文章の情報が得られることは一般的ではない．そのため，検索結果で最上位であった文書のうちどの部分が高適合であったのかを推定する必要がある．本論文では，$g$ として以下の 3 つの手法を用いる．

Random　検索結果の最上位の文書を文章に分割し，一様分布にランダムに 1 つの文章を摂動とする手法.

表 1　データセットの概要.

| | 文書数 | クエリ数 | 平均文書長 | 平均クエリ長 |
|---|---|---|---|---|
| TREC Robust 2004 | 528,155 | 250 | 2120.255 | 3.616 |

TextRank　文書の中で中心性の高い文章はその他の文章とよく類似しているため，文書中の中で平均的に高いスコアを得やすいという仮定の元で，次のように TextRank を用いて最もスコアが高かった文章を摂動とする手法.

$$p^* = \arg\max_{p \in P} \mathrm{TextRank}(p, P_i)$$

ただし，TextRank は $P_i$ における $p$ の TextRank のスコアを出力する関数である．

BM25　検索結果の最上位の文書を文章に分割し，次のように BM25 を用いてクエリとのスコアが最も高かった文章を摂動とする手法.

$$p^* = \arg\max_{p \in P} \mathrm{BM25}(q, p)$$

ただし，BM25 は $p$ の $q$ に対する BM25 のスコアを出力する関数である．

## 4 実　　験

### 4.1 データセット

提案手法の評価にはアドホック文書検索データセットである TREC Robust 2004 を用いた．このデータセットの統計情報を表 1 に示す.

### 4.2 実験設定

攻撃対象のモデルとして，アドホック文書検索タスクである MS MARCO （a large scale MAchine Reading COmprehension dataset） Document Ranking [20] のデータセットを用いて学習された ANCE (Approximatenearest neighbor Negative Contrastive Learnin) [3] と BM25 を用いて比較する．長い文書の適合度計算は文書を 512 トークン単位で文章に分割する．各文書ごとにクエリとの適合度計算を行い，最も高い適合度を文書の類似度とする．また，それぞれの攻撃対象モデルで上位 100 件の文書を取得し，$t = 100$ として 100 位に位置していた文書の摂動追加後順位の変動を測る．

高適合であったパッセージを推定するために，文書の分割単位を決定する必要がある．ただし，密検索モデルの情報を得られないブラックボックス設定であるため，いくつかの分割単位を設定し，それぞれでの分割単位での攻撃の効果を比較する．実験では，文書の分割単位長として 512, 256, 128, 64, 32 の 5 つを設定し，それぞれの分割単位長で文書の分割を行う．文書の分割に利用するトークナイザとして，Hugging Face[1]が公開している BERT-base-uncased のトークナイザを用いた.

1：https://huggingface.co

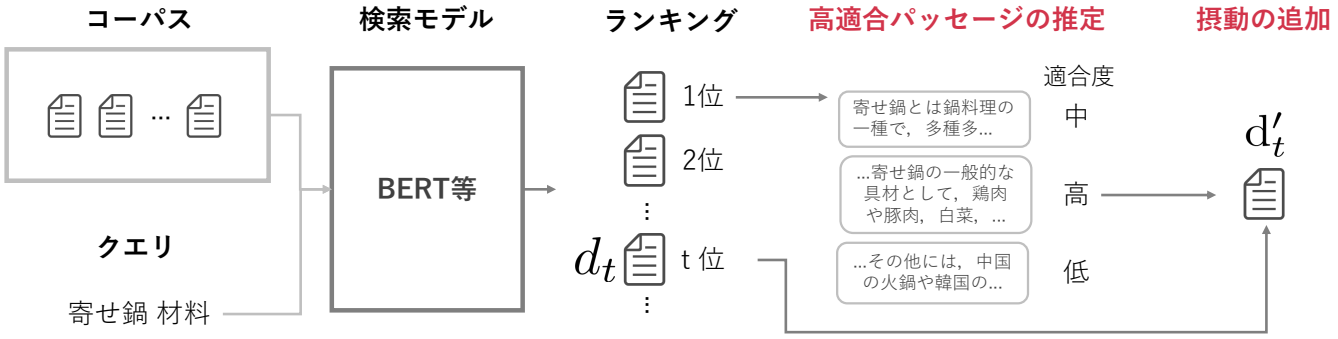


図 1　提案する攻撃手法の概要図

表 2　各分割長・検索モデル・文章推定手法における MPR と APR.

| | | MPR | | APR | |
|---|---|---|---|---|---|
| 分割長 | 文章推定手法 | ANCE | BM25 | ANCE | BM25 |
| 32 | Random | 100 | 87 | 79.360 | 70.828 |
| | TextRank | 98 | 75 | 74.852 | 65.104 |
| | BM25 | 58 | 15 | 58.764 | 24.668 |
| 64 | Random | 91 | 59 | 69.244 | 56.832 |
| | TextRank | 70 | 44 | 63.180 | 48.512 |
| | BM25 | 29 | 10 | 43.212 | 17.708 |
| 128 | Random | 42 | 21 | 48.736 | 38.084 |
| | TextRank | 20 | 14 | 41.524 | 30.424 |
| | BM25 | 10 | 7 | 27.112 | 12.824 |
| 256 | Random | 24 | 10 | 42.336 | 27.224 |
| | TextRank | 8 | 7 | 27.456 | 19.720 |
| | BM25 | 5 | 4 | 21.468 | 8.676 |
| 512 | Random | 9 | 5 | 31.220 | 15.768 |
| | TextRank | 7 | 4 | 29.500 | 9.524 |
| | BM25 | 5 | 3 | 23.040 | 5.864 |

#### 4.2.1　評価指標

最初に攻撃対象の検索モデルを用いて各クエリについて検索を行い，検索結果を得る．その後検索結果を用いて敵対的サンプルを作成し，$d_t$ を敵対的サンプルである $d'_t$ と置き換える．再度同じクエリで検索し，順位の変化を測る．具体的には，以下の指標を評価指標として用いる．

APR (average promoted rank)　全てのクエリについて，再検索を行なった後の $d'_t$ の順位の平均．

MPR (median promoted rank　全てのクエリについて，再検索を行なった後の $d'_t$ の順位の中央値．

$d'_t$ の順位を向上させることが目的であり，これらの評価指標の値が小さいほど効果的な攻撃であったことがわかる．

### 4.3　実験結果

#### 4.3.1　検索モデルへの攻撃の効果

実験結果を表 2 に示す．ANCE を対象とした結果では，分割長が 256 で文章推定手法に BM25 を用いた場合に MPR・APR ともに最小となっている．分割長が 32 の場合は，ANCE に対して文章推定手法が BM25 の時に APR で約 58.8, MPR で 58 となっており，平均して約 59 位の順位の上昇が得られた．分割長が 512 の場合では，ANCE に対して文章推定手法が BM25 の時では，APR で約 23.0，MPR で 5 となっており，平均して 23 位の順位の上昇が得られた．ANCE に対する攻撃のうち，分割長が 512 で文章推定手法が BM25 と TextRank は，分割長が 256 であった場合の結果を下回っているが，それ以外の全ての分割長・検索モデル・文章推定手法において，分割長が大きくなるほど MPR・APR ともに小さくなっている．一方で，分割単位長が大きくなることは摂動が大きくなり元々の文書との差分が大きくなることを意味しており，攻撃に気づかれ易くなる．このため，攻撃のもっともらしさを考慮すると，追加する摂動が元の文書と比較してどの程度の大きさであったかを調査する必要がある．図 3 に文書長の分布を示す．この図から，約 57%の文書が 512 トークン以下の文書長となっている．平均文書長では表 1 より約 2120 であり，512 トークンはこのうちの約 24%であるが，データセット中の半数以上の文書において，512 トークンの追加は元々の文書長以上のテキストの追加となっていることがわかる．

BM25 との比較において，全ての分割長・文章推定手法で BM25 の方が MPR・APR ともに小さくなる結果となった．特に文章推定手法の Random と TextRank において，BM25 の方が MPR・APR ともに小さくなったことから，提案する攻撃は BM25 に対してより効果的であることが示された．

以上の結果から，高適合であったパッセージを推定して追加

表 3　文書を 512 トークンで文章に分割した時に生成される文章の個数と，その文章数に分割された文書数の割合．8 個以上に分割された文書は合算して表示している．

| 文章数 | 文書の割合 |
|---|---|
| 1 | 57.052% |
| 2 | 24.986% |
| 3 | 11.670% |
| 4 | 3.404% |
| 5 | 1.285% |
| 6 | 0.555% |
| 7 | 0.302% |
| ≧ 8 | 0.746% |

表 4　ANCE における，分割長 512 の時の各文章推定手法の正解率．ここで正解率とは，文章推定手法によって推定された文章が，実際に適合度が最大値となった文章であった割合．

| 文章推定手法 | 正解率 |
|---|---|
| BM25 | 0.844 |
| TextRank | 0.740 |
| Random | 0.648 |

する攻撃手法に対して，BM25 との比較において ANCE はより堅牢であることが判明した．

**4.3.2**　文章推定手法の性能の影響

表 4 に，ANCE を攻撃対象とした時の高適合であった文章の推定手法の性能を示す．Random の正解率が 0.5 以上になっているが，これは図 3 より，文章数が 1 になっている文書の割合が高いことに起因していると思われる．この表から，BM25 の正解率が最も高いことがわかる．表 2 より，全ての攻撃手法と検索モデルにおいて，Random で APR と MPR ともに最大になっており，BM25 で最小となった．これは，高適合であった文章の推定手法の性能が，攻撃の性能に影響していることを示している．

節 4.3.1 より，ANCE を対象とした攻撃において，文章推定手法の BM25・TextRank は，分割長が 512 のときの攻撃性能が分割長が 256 の時の結果を下回っていることがわかった．一方で，文章推定手法に Random を用いた場合では，分割長が 512 の時に APR・MPR ともに最小となっている．また，BM25 を対象とした攻撃では，分割長の増加に伴い MPR・APR は一貫して減少している．このことから，分割長が大きい場合，ANCE ではこれ以上の文章推定手法の性能の向上が，攻撃の性能へ与える影響は小さいことが考えられる．

分割長が 32 の場合は，ANCE に対して文章推定手法 Random の攻撃を適応したときに APR で約 80, MPR は 100 となっており，摂動の追加が順位の向上が低い．一方で，文章推定手法が BM25 の時は APR で約 58,MPR は 58 となっており，文章推定手法が Random の時と比較して攻撃の性能が大きく向上している．同様の傾向が分割長が 64 の時にも確認できる．分割長が 64 の時では，文章推定手法として BM25 が最も高く，Random が最も低くなっており，その MPR および APR の差の絶対値は他の分割長のものより大きくなっている．そのため，追加する摂動の長さが短い場合には，高適合であった文章を推定する手法の性能が攻撃性能へ与える影響が相対的に大きいと考えれる．

**4.3.3**　クエリ追加攻撃

節 4.3.1 の実験により，ANCE が単純な文章の追加に対して BM25 より頑健であることが判明した．一方で，BM25 による高適合文章の推定が，ANCE に対する攻撃の性能を向上させることも確認できた．BM25 による適合度計算では，クエリ語を多く含むと高適合となりやすいため，ANCE に対してクエリ語を含む文章の追加は有効であることが考えられる．そこで，クエリ語の追加がどの程度 ANCE に影響を与えるのかを調査するため，さらに単純な攻撃として，クエリを直接的に文書に追加する攻撃手法を適応する．この攻撃手法はウェブスパムとして古くから用いられている手法である [5]．BM25 に対しても同様の攻撃を適応して ANCE と BM25 に対する攻撃結果を比較する．BM25 は適合度の計算方法上，クエリを文書に追加によってその文書のスコアは上昇する．しかし，密検索モデルでは文書とクエリの密ベクトルから適合度を計算しており，意味的な類似度を捉えていると考えられているため [16, 17]，クエリを単純に文書に挿入する攻撃が ANCE の適合度計算に与える影響は自明ではない．

表 5　クエリをそれぞれ 1 から 3 回追加する攻撃を行なった際の MPR と APR.

| | MPR | | APR | |
|---|---|---|---|---|
| 追加回数 | ANCE | BM25 | ANCE | BM25 |
| 1 | 19 | 42 | 32.800 | 45.740 |
| 2 | 10 | 4 | 22.556 | 12.032 |
| 3 | 6 | 4 | 18.468 | 10.668 |

表 5 にクエリを追加する攻撃を行なった際の MPR・APR を示す．この表から，クエリを追加する回数を増やすごとに，ANCE と BM25 ともに APR・MPR の両方が小さくなっていることがわかる．クエリの追加回数が 3 回の場合，BM25 の方が効果的に順位が上昇しているが，ANCE でも順位が大きく向上している．表 1 より平均クエリ長は約 3.6 であることから，3 回クエリを追加した場合には約 10.8 トークンの追加となる．約 10.8 トークンの追加にも関わらず, 表 2 における分割長が 512 で BM25 による高適合文章推定を行なった場合と近い攻撃性能となった．また，クエリを一回追加する場合では，BM25 と比較して ANCE の方が効果的に順位が上昇している．以上のことから，少ない回数であっても，クエリを文書に追加することで効果的に文書の順位を向上することがわかる．

## 5　ま　と　め

本論文では，大規模言語モデルをベースとした密検索モデルに対するシンプルな攻撃手法を提案し，脆弱性の評価に取り組んだ．この攻撃手法では，あるクエリについて上位に位置していた文書から高適合な文章を推定し，下位に位置していた文書に摂動として加えることで，同じクエリで改変を加えた文書を上位に位置させる手法を提案した．実験では攻撃対象の検索モデルとして ANCE と BM25 を，データセットに Trec Robust 2004 を用いて攻撃手法に対する脆弱性の比較を行なった．その結果，BM25 と比較して ANCE は提案する攻撃手法に対して相対的に堅牢であることが判明した．一方で，クエリ語を追加するような攻撃に対しては，ANCE に対しても効果的に文書の順位向上に貢献することが判明した．今後の課題として，クエリ追加に対して堅牢な密検索手法の提案，攻撃時の密検索モデルの振る舞いについての更なる分析などが挙げられる．

**謝辞**　本研究は JSPS 科研費 22H03905 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] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. International Conference on Learning Representations (ICLR), 2014.

[2] Long Lu, Roberto Perdisci, and Wenke Lee. Surf: Detecting and measuring search poisoning. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS '11, page 467–476, New York, NY, USA, 2011. Association for Computing Machinery.

[3] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. arXiv, 2020.

[4] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345, 2021.

[5] Zoltan Gyongyi and Hector Garcia-Molina. Web spam taxonomy. Technical report, 2004.

[6] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[7] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

[8] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. arXiv, 2019.

[9] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. 2017.

[10] Nisarg Raval and Manisha Verma. One word at a time: adversarial attacks on retrieval models. arXiv, 2020.

[11] Yumeng Wang, Lijun Lyu, and Avishek Anand. Bert rankers are brittle: a study using adversarial document perturbations. arXiv, 2022.

[12] Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Prada: Practical black-box adversarial attacks against neural ranking models. arXiv, 2022.

[13] Jiawei Liu, Yangyang Kang, Di Tang, Kaisong Song, Changlong Sun, Xiaofeng Wang, Wei Lu, and Xiaozhong Liu. Order-disorder: Imitation adversarial attacks for black-box neural ranking models. arXiv, 2022.

[14] Dennis Fetterly, Mark Manasse, and Marc Najork. Detecting phrase-level duplication on the world wide web. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 170–177, New York, NY, USA, 2005. Association for Computing Machinery.

[15] Carlos Castillo, Brian D Davison, et al. Adversarial web search. *Foundations and trends® in information retrieval*, 4(5):377–486, 2011.

[16] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey, 2022.

[17] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. Pretrained transformers for text ranking: Bert and beyond, 2020.

[18] Xinyu Zhang, Andrew Yates, and Jimmy Lin. Comparing score aggregation approaches for document retrieval with pretrained transformers. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II*, page 150–163, Berlin, Heidelberg, 2021. Springer-Verlag.

[19] Zhuyun Dai and Jamie Callan. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, jul 2019.

[20] 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. 2018.