Ada or Bert：検索における文埋め込み計算手法の比較研究

馬　春鵬　　松田　寛
株式会社リクルート Megagon Labs, Tokyo, Japan
{ma.chunpeng, hiroshi_matsuda}@megagon.ai

概要

自然言語処理技術が使われている様々なサービスに、検索が中心的役割を果たしているものが多いが、近年の大規模言語モデルは検索に与えた影響が不明だ。本調査報告書は宿検索に対する文埋め込み計算手法を比較した。文の埋め込みの計算において、従来の BERT モデルと OpenAI 社の新しいモデルのそれぞれの強みと弱みを分析した。

1　はじめに

自然言語処理技術の進歩によって、実世界の様々なサービスが自然言語処理を導入しているが、その中でも文埋め込みを用いた意味的な類似性に基づいた検索は重要な役割を担うようになった。また、近年では大規模言語モデルの進歩が注目されている。特に、2022 年 11 月、ChatGPT[1)] という大規模言語モデルに基づいたチャットボットが公開され、生成 AI のブームが到来した。検索において、この生成 AI ブームの影響はまだ様々な疑問が残されている。例えば、ChatGPT に基づくモデルは検索結果を改善できるか？従来の手法は今の時代にまだ必要なのか？実際の応用に焦点を絞ると何が言えるか？等がある。

本報告書は検索における文埋め込み計算手法を調査する。具体的に、宿検索を題材として、従来手法（J-BERT と S-BERT、2.1 節を参照）と OpenAI 社の新たな手法（ada-002、2.1 節を参照）を比較し、それぞれの強みと弱みについて分析する。我々の観察から、以下のことがわかった。

- ada-002 は視点の広さと知識量の多さにおいて劣っている。
- ada-002 は逸脱表現に対する受容性において優れている。

これらの観察は宿検索のような特定なタスクにおいて OpenAI 社の新しいモデルだけでは不十分で、従来の手法はまだ必要であることを示唆している。

2　文埋め込みと検索

2.1　文埋め込みの計算手法

本研究は宿検索を題材として、以下の三つの計算手法を対象とする。

2.1.1　旅行領域に特化した BERT

BERT [1] は近年の人工知能分野に大きい影響力を持っている基盤モデルの一つである。文をモデルに入力すると、モデルの各々の隠れ層はベクトルを出力する。これらのベクトルは文の埋め込みとし、検索を含む様々なタスクに役に立つ。

我々は旅行領域に特化した BERT を作った。具体的には、4.1 節に記した大量の宿の口コミデータを用いて SentencePiece の vocab 構築と事前学習を行った。さらに、宿の評価文の含意関係に関するタスクでモデルの微調整を行った。このように構築したモデルは宿検索に役に立つ知識を持つと想定する。

以下、このモデルを「J-BERT」と呼ぶ。

2.1.2　旅行領域に特化した SentenceBERT

SentenceBERT [2] は文の埋め込み計算のために BERT を改良したものである。簡単に言うと、BERT で二つの文の埋め込みを別々に計算した後、二つの埋め込みを用いて、BERT のパラメーターを微調整する。こうやって学習されたモデルは文の埋め込みの計算に特化され、より良質な埋め込みを計算することが期待できる。前と同様に、我々もこのモデルを旅行領域に特化した。

以下、このモデルを「S-BERT」と呼ぶ。

1）https://chat.openai.com/

2.1.3 text-embedding-ada-002

OpenAI 社が提供する文埋め込み計算モデル[2)]。技術詳細は不明だが、検索を含む様々なタスクにおいて良い性能を示している。ただし、このモデルは（本稿執筆時点では）微調整を行なうことはできないため、汎用的検索タスクではないドメイン特化検索タスクでどの程度の性能を示すかは評価する必要がある。

以下、このモデルを「ada-002」と呼ぶ。

2.2 文埋め込みに基づいた検索

検索は文埋め込みに基づいて行う。検索は以下のように定義する。

定義 1 (検索)**.** 入力 Q に対して、文書集合 $\mathfrak{D} = \{d_1, d_2, \cdots, d_m\}$ から Q に類似度が高い部分集合 $\mathfrak{D}'(Q) \subseteq \mathfrak{D}$ を作る過程は検索と呼ばれる。ただし、$d_i (1 \leq i \leq m)$ と Q は自然言語の文であり、$|\mathfrak{D}'(Q)| << m$ である。

本研究にコサイン類似度を採用した。文書 d とクエリ Q のそれぞれの埋め込みを $\mathrm{emb}(d)$ と $\mathrm{emb}(Q)$ とし、コサイン類似度は以下の式で計算する。

$$\mathrm{sim}(d, Q) = \frac{\mathrm{emb}(d) \cdot \mathrm{emb}(Q)}{\|\mathrm{emb}(d)\| \|\mathrm{emb}(Q)\|} \tag{1}$$

3 クエリの分類体系

検索結果の質は顧客の入力（即ち、クエリ）に大いに依存しているので、検索におけてモデルの性能を比較するために、様々な尺度によってモデルを評価する必要がある。本節はクエリの分類体系を作り、実際に使われたクエリの一覧は表 2 を参照。

3.1 尺度 1：視点の広さ

一番重要な尺度は「視点の広さ」である。直感的に、もしモデルがクエリを全て読み込んで、クエリ全体の意味を解釈した上で検索すると、良い検索結果が期待できる。逆に、もしモデルがクエリの一部しか解釈しないと、検索結果の質はあまり期待できない。このように、クエリから読み取るべき顧客の意図を検索結果にどのくらい反映できているかを表すのは「視点の広さ」という概念である。

定義 2 (視点)**.** クエリに対するモデルの視点とは、モデルが検索を行う際、実際に使った単語の数である。クエリ $Q = (w_1, w_2, \cdots, w_n)$ に対して、モデル M が $W \subseteq Q$ を使って検索を行った場合、クエリ Q に対するモデル M の視点は $|W|$ である。

モデルの視点の広さを評価するために、長いクエリが理想的である。なぜなら、クエリが長くなると、より多くのトピックを適切に捉えること、トピックの組み合わせから適切なランキングを行うことが求められるからである。視点の狭いモデルは、ごく少数の単語の表層的な類似性を重視する傾向があり、クエリの全体的な意味が無視されて、検索結果の適合性が低下する。長いクエリの作成は次のような観点で行った。

- 長い修飾をつけること
- 理由をつけること
- 明示的否定を加えること
- 婉曲な打ち消し表現を加えること
- 無意味な単語を入れること

3.2 尺度 2：知識量の多さ

J-BERT の事前学習では大量の旅行ドメインのテキストを用いており、旅行ドメインに関する知識が埋め込みとして学習されていることが期待できる。モデルの知識量を評価するため、常識的知識のみならず、特定分野（特に検索タスク）の知識量の多さが知りたい。そのため、クエリに特定分野の知識に関する質問を加えると、もしモデルがその知識を知らない場合、検索結果が間違ってしまう。

我々は日本にある宿の検索タスクに注目するため、以下の三つの知識に対してモデルの知識量を評価する。

- 宿関連の知識
- 地理・文化の知識
- 日本独自の知識

3.3 尺度 3：逸脱表現に対する受容性

顧客の意図を汲んだ検索結果を得るためには、単なる文字列ベースの検索では不十分であり、クエリの意味を解釈することが必要不可欠である。そのためには、日本語の様々な文章表現（特にレアケース）に対する受容性・頑健性が求められている。大規模言語モデルは事前学習時に大量の日本語テキストから、基本的な語彙や表現の解釈は可能になっている。難しいのはインフォーマルな表現を受け入れるかどうかである。それを評価するために、逸脱表現をクエリに入れる必要がある。具体的に、三つの

2) https://openai.com/blog/new-and-improved-embedding-model

逸脱表現が用いられている。

- イレギュラーな日本語表現
- 顔文字と絵文字
- 英単語や横文字が散らばっている表現

4 実験

4.1 実験設定

宿検索用データ 検索対象はじゃらん[3]に掲載されている日本全国の宿である。それぞれの宿に対して、検索する際に使われたデータは口コミのみである。宿の口コミは文単位に分割され、それぞれの文を検索対象とする。このデータは J-BERT と S-BERT の学習時にも使われた。

比較用クエリ 第 3 節にクエリの分類体系を説明した。その分類体系のぞれぞれのカテゴリに、クエリを 2 つずつ作成した。使われたクエリの一覧は表 2 にある。

効率的検索用ツール 検索する際、埋め込みの類似度を効率的に計算するために、vecscan[4]が使われている。

4.2 実験結果

クエリの分類体系にあるそれぞれの尺度に対し、代表的なクエリを一つずつ選び、上位 3 個の口コミ検索結果を表 1 に記載する。[5]

4.2.1 視点の広さに関する評価

視点の広さを評価するために使われたクエリの例は「寝る前に美しい夜景を楽しめる高層階に部屋をアップグレードできるホテル」である。このクエリの核心的要望は「高層階」であり、それ以外の「夜景」や「アップグレード」などは修飾用単語であるため、検索結果に「高層階」があることが望ましい。

J-BERT と S-BERT の検索結果にある全てのホテルの口コミに、「高層階」や「最上階」といった表現があり、良い結果と言える。一方、ada-002 の検索結果に、例えば、1 番目と 2 番目の口コミは共に「高層階」とは関係なく、修飾用の「夜景」や「アップグレード」だけがある。よって、このクエリに対して、ada-002 の検索結果は他の二つのモデルに劣っていることが分かった。

4.2.2 知識の多さに関する評価

視点の広さを評価するために使われたクエリの例は「延泊料金が安い宿」である。「延泊」は旅行領域の専門用語だが、特に「宿泊代」との違いについて区別されることが重要である。

検索結果が良好だったのは S-BERT である。S-BERT の 1 番目は「延泊」という表現が含まれ、2 番目にも「延泊」とそれに近い「連泊」といった表現が含まれ、良い結果である。J-BERT の検索結果は少々劣っているが、1 番目の口コミにも「延泊」という単語があり、クエリの要望に満たす。一方、ada-002 の検索結果にある三軒の宿の口コミは全部「延泊」に無関係であり、宿泊料金は安いが、延泊料金が安いかどうかは結局不明である。

4.2.3 逸脱表現に対する受容性に関する評価

逸脱表現に対する受容性を評価するために使われたクエリの例は「拙者、潔き室と夜の静寂、また、舌鼓を打つ食事を願う。」である。「拙者」や「潔き」のような表現は現代日本語に（特に旅行領域のデータに）滅多に見ないため、モデルがこのような表現を解釈できない恐れがある。

J-BERT も S-BERT もクエリと無関係な検索結果を出した。全ての口コミはクエリのどの部分に対応するかは判別が難しく、クエリはモデルによってうまく解釈されていないことが推測できる。一方、ada-002 の検索結果の方が遥かに優れている。「潔き室」は「清潔感」という意味で、「舌鼓を打つ」は「美味しい」という意味であることが知られているようで、口コミは全てクエリに一致している。

4.3 まとめ

本研究の実験結果をまとめる。

- 視点の広さにおいて、ada-002 が劣っている。クエリの核心的要望や打ち消し表現を無視し、修飾・理由・無意味な単語にミスリードされる傾向がある。
- 知識量の多さにおいて、ada-002 が劣っている。特定分野の知識が備えていないため、クエリにある特定分野の専門用語を正確に解釈できないことがある。
- 逸脱表現に対する受容性において、ada-002 が優れている。古文・外国語・絵文字といった一

3） https://www.jalan.net
4） https://github.com/megagonlabs/vecscan
5） 実際に検索結果は宿のリストとなる。宿のスコアには該当する口コミ文の上位 3 件のスコアの重み付き平均を使用している。

表 1　検索結果の比較。口コミだけが記されている。角括弧にある数字は口コミスコアである。

クエリ	J-BERT	S-BERT	ada-002
【尺度 1：視点の広さ】 寝る前に美しい夜景を楽しめる高層階に部屋をアップグレードできるホテル	[0.9705] 高層階に宿泊できたときは東京の夜景が楽しめます。 [0.9695] 部屋は高層階駅側指定だったので東京タワーや遠くにスカイツリーも見えて非日常を味わえました。 [0.9680] 高層階ですと小さな豊橋の夜景を楽しむことができます。	[0.7942] 一泊でサクッと泊まる予定でシングルを予約したら、ホテル側の都合で最上階の夜景ルームと言う素敵なお部屋に変更になっていました！ [0.7941] ３６階の高層部屋さすがに夜景はきれいでした。 [0.7930] 非常に夜景がきれいな高層階の広いお部屋にサービスで変えていただき感激しました。	[0.9340] ホテルのご好意で夜景の綺麗な部屋にアップグレードしていただきました。 [0.9317] 予約していたプランより高い部屋にグレードアップして頂いており、夜景も楽しめました。 [0.9311] 高層階の宿泊だと、夜景も楽しめます。
【尺度 2：知識量の多さ】 延泊料金が安い宿	[0.9691] 急な延泊で当日宿を取りました。 [0.9666] 料金からするとコスパの高い宿です。 [0.9635] 連休初日に安価に宿泊できてありがたい宿です。	[0.6526] また翌日延泊対応も感謝です。 [0.6485] 連泊プランも有り安く宿泊できる。 [0.6222] 平日宿泊費が安価。	[0.9233] 宿泊費が安い。 [0.9229] 宿泊料が安い。 [0.9191] 宿泊代が安い。
【尺度 3：逸脱表現に対する受容性】 拙者、潔き室と夜の静寂、また、舌鼓を打つ食事を願う。	[0.9756] また、早朝に神職様のご祈祷を受け、心が清められる思いでした。 [0.9744] また、幸いにして普通の日本間で「荘」に相応しいお部屋でした。 [0.9711] また、境内には松尾芭蕉の芭蕉像を安置する芭蕉堂のあります。	[0.7073] 夕食は、また、小僧達もグレイトですって、確かに、食事処は、最高です。 [0.6762] 珍しい休暇は、不愉快な宿泊に換えて経験して、民宿主人に稼いだ良心があることを希望します。 [0.6760] 宿も素敵なお部屋と静けさに包まれ、食事も美味しい！	[0.8949] 清潔かつ朝食が美味しく大浴場があり夜泣きそばまで。 [0.8934] 又部屋は、清潔感があり夜の食事は、魚好きには、満足するとおもいます。 [0.8933] 濁り湯、食事良しこじんまり静かな宿好みの自分には相性が良いです。

般的ではない表現を正確に解釈できている。

となった。

ada-002 の強みは逸脱表現に対する受容性だが、宿検索タスクにおいて、古文や絵文字など一般的ではない表現は稀にしか見えないため、その強みを発揮する場があまりない。一方、宿検索タスクにおいては、長いクエリや旅行領域の専門用語が含まれたクエリは非常に重要であるが、このようなクエリに対する ada-002 による文埋め込みの性能は相対的に低かった。

5　おわりに

本研究は宿検索を題材として、三つの文埋め込み計算手法を比較した。宿検索タスクで重要な「視点の広さ」や「知識の多さ」については J-BERT および S-BERT の文埋め込みが ada-002 のものより優れていた。他方、「逸脱表現に対する受容性」については ada-002 の方が優れていた。総じて、宿検索タスクにおいては従来型の BERT をベースとした文埋め込み手法は ada-002 より実用性に勝ると考えられる。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

A　クエリ一覧

表 2 には、モデルを評価するために使われたクエリの一覧が記されている。

表 2　モデルを評価するために使われたクエリの一覧。

尺度	種類	クエリ
視点の広さ	長い修飾をつける	a) 寝る前に美しい夜景を楽しめる高層階に部屋をアップグレードできるホテル b) 瑞々しい旬の京野菜が食べられて食通でも満足できる宿
	理由をつける	a) 書類を整理したり、作業をするのが好きなので、静かで作業がしやすい部屋があるホテルを探しています。 b) 初めての結婚記念日を迎えますので、特別なホテルを探しています。私たちは美しい夕日が見える海沿いの部屋がいいと思っています。
	明示的否定を加える	a) 冷蔵庫が無くてもいい、とにかく安い宿がいい。 b) 駅から遠くない宿
	婉曲な打ち消し表現を加える	a) 子連れの家族には厳しいかも。 b) 部屋は綺麗ですが、フロントは…
	無意味な単語を入れる	a) ChatGPT のせいで、研究が超大変になった。静かな旅館に夕食を食べてゆったりと休みたい。来週また仕事頑張るぞ。 b) 彼女と付き合って 1 周年、すごいでしょう？ところで、今回は一人で出張、カプセルでいい。
知識量の多さ	宿関連の知識	a) フロントのスタッフに目覚ましのコールをお願いしたい。 b) 延泊料金が安い宿
	地理・文化の知識	a) うどんを満喫したい。 b) ラフテーが食べられる
	日本独自の知識	a) 花見の季節に泊まりたい b) アイドルのライブを見るのに便利な場所
逸脱表現に対する受容性	イレギュラーな日本語表現	a) 拙者、潔き室と夜の静寂、また、舌鼓を打つ食事を願う。 b) 月見の間で静寂を楽しむ場所を設けてくださると幸いでござる
	顔文字・絵文字	a) 😭😭初めての一人旅😭😭、おすすめな宿あります？😭😭😭😭 b) (T ． T)(T ． T) 初めての一人旅 (T ． T)(T ． T)、おすすめな宿あります？ (T ． T)(T ． T)(T ． T)(T ． T)
	英単語や横文字が散らばっている	a) 日本の history が好きで、very interesting と感じているので、歴史がある hotel をお願いします b) ジャパンのヒストリーがとてもインプレッシブと感じているので、サムライやニンジャが楽しめるところをお願いします。