

ユーザ行動ログに基づく クエリ理解のための検索クエリ埋め込み

西川 莊介* 平子 潤* 鍛冶 伸裕 渡邊 幸暉
浅野 広樹 山城 颯太 佐野 峻平
LINEヤフー株式会社

{sonishik,jhirako,kokwatan,nkaji,hiroasan,soyamash,shsano}@lycorp.co.jp

概要

検索クエリの意図をよく反映した埋め込み表現は、検索エンジンの様々なクエリ理解 (Query Understanding, QU) タスクに活用できる。しかし、検索クエリは短く表層形の変化が生じやすいため、既存手法では意図を十分に捉えられない。本研究ではユーザ行動ログから抽出した、検索意図が類似するクエリペアを正例とする対照学習手法を提案する。このペアは表層形の類似性に依らず意図が類似するため、表層形の差異に頑健でクエリに内在する意図を的確に捉える埋め込みモデルを実現できる。提案モデルは複数の QU タスク評価で Ruri や SimCSE などの最先端テキスト埋め込みモデルを大幅に上回った。

1 はじめに

クエリ推薦やクエリ分類などのクエリ理解 (QU) タスクは、検索エンジンにおける検索体験向上の要となる技術である [1, 2]。この QU タスクに対して近年、表層形の不一致に頑健な埋め込みに基づく手法が盛んに適用されている [3]。しかし、個々の QU タスクに応じて埋め込みモデルを学習するのはコストが高く、多様な QU タスクに適用可能な汎用クエリ埋め込みの必要がある。

しかし、一般的に検索クエリは短く簡潔に表されるため、文脈情報が乏しくその意図を正確に捉えるのは難しい [4]。また、検索クエリは短いため同義語置換のような軽微な編集で表層形が大きく変化しやすい。例えば「ディズニー 最寄り」と「TDL アクセス」は類似する意図を持つが、表層形は大きく異なる。これらの性質を考慮し、検索クエリに適した埋め込みの学習法を検討する必要がある。

テキスト埋め込みの学習において対照学習に基づ

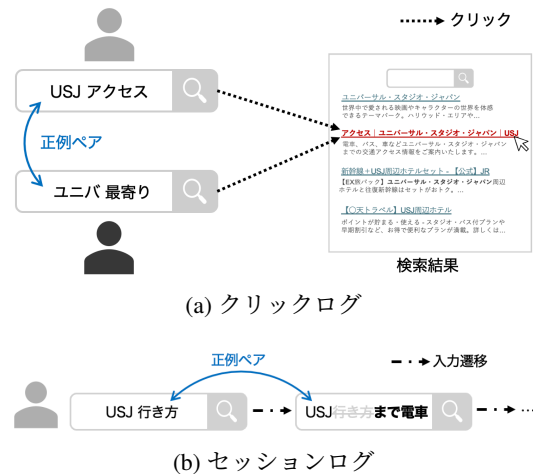


図 1: ユーザ行動ログから抽出される擬似正例ペア。

く手法が広く成功を収めている [5, 6, 7, 8, 9]。特に最先端の手法は、Q&A サイトの質問・回答のペアや Wikipedia のタイトル・節のペアなど、大規模なウェブ由来の弱教師ペアを活用する [6, 7, 8]。しかし、これらは主に長く文脈が豊富なテキストで構成され、短く文脈に乏しい検索クエリには十分に適用できない。また、検索クエリコーパスを使用し、教師なし対照学習を適用する方法がある。代表的かつ強力な手法である教師なし SimCSE [5] では同一文を異なるドロップアウトノイズを用いて 2 回符号化することで擬似正例ペアを生成する。しかし、このような正例ペアで学習されたモデルは表層形を重視して類似性判定する傾向にあり、表層形は異なるが意図は一致するペアを同定できないことが観測された (§5)。

以上より、既存手法に基づく正例ペアはクエリ埋め込み用の学習データとして不十分だと考えられる。本研究ではユーザ行動ログから抽出された意図が類似するクエリペアを対照学習の擬似正例として活用し、ユーザの意図に基づく埋め込みを学習する **UBIQUE** (User Behavior-driven contrastive learning with Intent alignment for search **Q**uery **E**mbeddings) を提案

* Equal contribution.

する。UBIQUE では二種類のユーザ行動ログを活用する (図 1)。

- **クリックログ**: クエリ入力後に検索結果の URL をクリックした記録。同一 URL をクリックしたクエリペアは類似する検索意図を持つ。
- **セッションログ**: 同一ユーザが一定時間内に検索したクエリ列の記録。同じセッション内のクエリペアは類似する検索意図を持つ。

UBIQUE ではこれらのログから抽出した意図が類似するクエリペアで対照学習 [10] を行い、検索クエリに内在する意図を的確に捉える埋め込みモデルを構築する。また、正例ペアは表層形の類似性に依存せず構築されるため、モデルが表層形の差異に頑健になることが期待される。

実験では、4 種類の実践的な QU タスクで UBIQUE を多面的に評価した (§4)。クリックログに基づく提案モデル (UBIQUE_{click}) およびセッションログに基づく提案モデル (UBIQUE_{session}) は、Ruri_{large} や教師なし SimCSE などのベースラインを大きく上回る性能を示した。また、提案モデルを精査した結果、表層形の異同に対して頑健で、内在的な意図を適切に捉えられている点が確認された (§5)。

2 関連研究

クエリ理解 (QU) QU は検索エンジンが結果を返す前に行うクエリ処理の仕組みを指し、検索体験の向上を促す [1, 2]。クエリは短く意図解釈が難しいため、深層学習登場以前からユーザ行動ログが QU に活用されてきた。例えばセッションログに基づくクエリ推薦 [11] やクリックログに基づくクエリ分類 [12]、クラスタリング [13] がある。近年、事前学習モデルの登場により QU の研究が進展した。例えば Jiang らは文脈不足を拡張トークン分類により緩和し [14]、Li らはクリックロググラフに基づく事前学習を提案した [15]。本研究ではユーザ行動ログによる事前学習モデルのファインチューニングにより、固定長の汎用クエリ埋め込みを構築するため、既存の事前学習法と併用できる。

類似した先行研究として、Zhang らの GEN Encoder [3] がある。これはクリックログとクエリペア関係性アノテーション付きデータを用いて学習され、クエリペア類似度計算に適した埋め込みを生成する GRU モデルである。一方で本研究は自動収集可能なユーザ行動ログのみ利用し、様々な QU タスク

に適用可能な汎用埋め込みモデルを提案する。

対照学習によるテキスト埋め込み学習 対照学習 [16] によるテキスト埋め込み学習では良質な正例ペアの構築法が一つの重要なテーマである。初期の研究では NLI データセットのようなアノテーション付きデータが利用されたが [5, 17]、最近ではウェブ由来の大規模弱教師ラベルの活用により最先端の性能を達成している [6, 7, 8, 9]。これらの研究は長いテキストを中心に扱う一方で、本研究では短い検索クエリの処理に焦点を当て、ユーザ行動に基づく弱教師ラベルを用いた対照学習手法を提案する。

教師なしの手法も広く研究されており、代表的な手法として教師なし SimCSE がある [5, 18]。Wu ら [19] は教師なし SimCSE で生じる文長に関するバイアスを議論しているが、本研究では教師なし SimCSE が表層形の異同に敏感な点を扱う。

3 UBIQUE

3.1 全体像

UBIQUE は検索意図が類似するクエリペア (q, q^+) を正例とした対照学習を行う。これらのクエリペアは、後述する 2 種類のユーザ行動ログから抽出される。UBIQUE モデルは、クエリペア集合 $D = \{(q_i, q_i^+)\}_{i=1}^m$ を用いた in-batch negatives [20] による InfoNCE 損失により学習される。

$$L_i = -\log \frac{e^{\text{sim}(\mathbf{q}_i, \mathbf{q}_i^+)/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{q}_i, \mathbf{q}_j^+)/\tau}} \quad (1)$$

ここで、 N はバッチサイズ、 τ は温度、 $\text{sim}(\cdot)$ はコサイン類似度であり、 \mathbf{q}_i および \mathbf{q}_i^+ はそれぞれ q_i および q_i^+ の埋め込み表現を表す。以降の節では正例クエリペア (q, q^+) の構築法を述べる。¹

3.2 クリックログ

クリックログはユーザが検索クエリを入力した後にクリックした URL の記録であり、同一 URL のクリックにつながったクエリ群は、類似する検索意図を持つ場合が多い [13, 21]。しかし、単に同一 URL をクリックするクエリペアを正例とすると、EC サイトのように様々な文脈でクリックされる URL や低頻度のミスクリックなどに起因する異意図クエリペアが生じる。このノイズを軽減するため、クエリに対してクリックされた全 URL を抽出し、その集合が類似

¹ なお、クリックログに基づくハード負例抽出を試行したが、明確な改善は確認されなかった (付録 A)。

するペアを正例として抽出した。これにより複数のクリック URL に基づき、より適切にユーザ意図を反映した信頼性の高いクエリペアを抽出できる。先行研究 [13, 22] に基づき、集合類似度にはジャッカード係数を用いた。

$$\text{Sim}_{\text{click}}(q_1, q_2) = \frac{U(q_1) \cap U(q_2)}{U(q_1) \cup U(q_2)} \quad (2)$$

ここで、 q_1 および q_2 は検索クエリ、 $U(q_i)$ はクエリ q_i 入力後にクリックした URL 集合を示す。類似度が閾値 θ を上回るクエリペアを正例とした。

3.3 セッションログ

セッションログは、時間 t 内に同一ユーザが検索したクエリ列である。同一セッション中のクエリ列は、前のクエリの編集や情報の追加など類似の意図を保持する場合が多い [11]。しかし、単に同一セッションに出現したクエリペアを正例とすると、目的のないネットサーフィンのように意図変動が激しい場合に異意図ペアが生じやすい。そこで典型的で信頼性のあるクエリ遷移ほど高頻度で共起するという仮定に基づき、複数のセッションから計算される共起度を使用して正例ペアを抽出する [23]。共起度には、様々な文脈で共起するクエリ (例：「YouTube」) によるバイアスを抑制するため、以下のジャッカード係数を用いる [11]。

$$\text{Sim}_{\text{session}}(q_1, q_2) = \frac{c(q_1, q_2)}{f(q_1) + f(q_2) - c(q_1, q_2)} \quad (3)$$

ここで、 $c(q_1, q_2)$ はクエリ q_1 と q_2 がセッション中で隣接して共起する頻度、 $f(q_i)$ は q_i の出現頻度を示す。閾値 ϕ を上回るクエリペアを正例とした。

実際のクエリペア例は付録の表 3 に示す。また、ジャッカード係数の有効性を付録 B で確かめた。

4 実験

4.1 評価

埋め込みの汎用性の評価には複数タスクでの多面的評価が重要である。本研究では公開データを含む 4 つの QU タスクで構成された日本語 QU ベンチマークを構築した (詳細は付録の表 4 を参照)。検索タスクでの類似度指標はコサイン類似度を用いる。

- **同義クエリ抽出 (Query-Synonym Retrieval, QR)**：検索意図が一致するクエリペアに対し、片方のクエリから対となるクエリを埋め込み上で引き当てる。評価指標は MRR。
- **クエリ推薦 (Query Suggestion, QS)**：ユーザが次

表1: QUベンチマークを用いた比較実験の結果。

モデル	Params	QR	QS	QC	SR	平均
<i>General</i>						
最先端モデル						
教師あり SimCSE _{large}	337M	40.9	81.3	85.4	88.4	74.0
Ruri _{large}	337M	67.7	86.3	88.0	90.5	83.1
mE5 _{large}	560M	63.1	87.3	82.4	91.1	81.0
Sarashina _{1.1b}	1.2B	<u>73.9</u>	89.2	84.5	<u>91.3</u>	84.7
OpenAI _{3-large}	-	65.9	89.9	80.8	91.4	82.0
同サイズモデル						
LINE DistilBERT	68M	20.3	79.7	83.4	87.3	67.7
Ruri _{small}	68M	54.5	87.6	84.1	90.8	79.3
mE5 _{small}	118M	59.5	87.7	71.5	90.8	77.4
<i>Search Logs</i>						
教師なしモデル						
fastText	-	22.9	84.8	82.5	87.7	69.5
教師なし SimCSE	68M	28.5	84.8	83.2	88.2	71.2
提案モデル						
UBIQUE _{click}	68M	91.4	<u>91.2</u>	85.8	90.5	89.7
UBIQUE _{session}	68M	71.9	92.6	<u>86.9</u>	90.3	<u>85.4</u>

に検索する関連検索キーワードを埋め込み上で引き当てる。評価指標は NDCG@10。

- **クエリ分類 (Query Classification, QC)**：クエリをカテゴリ別に分類する。埋め込みを入力とする線形分類器を用い、地図検索関連クエリをランドマーク名、チェーン店名、住所、駅名の 4 つに分類。評価指標は Macro-F1。
- **短文リランキング (Short-Text Reranking, SR)**：クエリとの埋め込み上での関連性に応じて短いテキスト (商品名) を並び替える。公開データである ESCI [24] を使用。評価指標は NDCG。

4.2 実験設定

2024 年 4 月の Yahoo! 検索のログデータを用いて、5,000 万クエリペアからなる学習データを構築した。クリックログでは θ を 0.4 とし、セッションログでは ϕ を 0.2、 t を 300 秒とした。ベースモデルには軽量で推論コストが低い LINE DistilBERT [25] を利用し、出力には [CLS] ベクトルを利用した。

ベースラインには JMTEB [26] で高いスコアを記録したテキスト埋め込みモデル (SimCSE_{large} [27], Ruri_{large} [28], mE5_{large} [7], Sarashina_{1.1b} [9], OpenAI_{3-large} [29]) と、提案モデルと同サイズのモデル (Ruri_{small}, mE5_{small}) を採用した。また、検索クエリで学習する教師なしモデルとして fastText [30] と教師なし SimCSE を採用した。教師な

表2: 「ロス 旅費」に対する埋め込み上での近傍クエリ.

モデル	1st Query	2nd Query	3rd Query
UBIQUE _{click}	旅行ロス	ロサンゼルス 旅行 費用	ロサンゼルス物価
UBIQUE _{session}	ロサンゼルス 旅行 費用	ロス 羽田	ロス現地時間
Ruri _{large}	バックパッカー費用	カリフォルニア ディズニー 旅費	ロサンゼルス 旅行 費用 4泊6日
教師なしSimCSE	ケアンズ 旅費	スイス 旅費	シンガポール 旅費

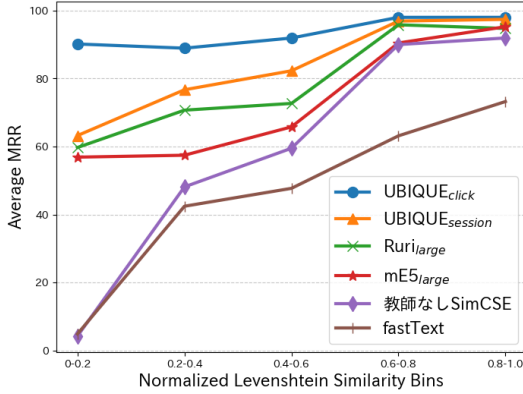


図 2: Levenshtein 類似度ごとの QU タスク性能.

しモデルは 5,000 万件の検索クエリで学習した. 実験設定の詳細は付録 C に示す.

4.3 結果

評価結果を表 1 に示す. UBIQUE_{click} と UBIQUE_{session} は共に同サイズのモデルを多くのタスクで大幅に上回り, 5 倍から 18 倍程度のサイズを持つ最先端モデルと比較しても平均スコアで上回った. 特に Ruri_{large} と比較すると, UBIQUE_{click} は平均スコアで 6.6 ポイント, UBIQUE_{session} は QS タスクで 6.3 ポイント上回った. Ruri や mE5 はリランカを利用した二段階の学習で構築されているにもかかわらず, 単一の対照学習のみで構築した提案モデルが性能を上回っており, クエリ埋め込みに特化した学習データ構築の重要性が確認された.

提案モデルは教師なしモデルと比較しても大幅に性能を上回った. 特に QR タスクでは, UBIQUE_{click} は教師なし SimCSE より 60 ポイント以上高い (91.4 vs. 28.5). この極端な性能差は, 教師なしモデルが表層形の異同に敏感で, 表層形が異なるペアの類似性を適切に捉えられないためだと考えられる (§5).

5 分析

UBIQUE による埋め込みの改善理由を分析するため, 各モデルにおける埋め込み空間での近傍クエリを調べた (表 2).²

² Faiss [31] を用いて約 1,000 万件のランダムクエリに対して近似最近傍探索を実施した.

ベースラインでは「ケアンズ 旅費」のように表層形の類似性のみを引きづられている異意図クエリや, 「カリフォルニア ディズニー 旅費」のように多少関連性はあるが意図は大きく離れたクエリが近傍に分布する傾向にあった. 一方, 提案モデルは「ロサンゼルス 旅行 費用」のように表層形の類似性に依らずより類似した意図を持つクエリが分布する傾向にあり, 内在的な意図を捉えられていることが分かった.

また, UBIQUE_{session} は UBIQUE_{click} と比べると「ロス現地時間」のように, より広範囲で遷移的な意図を持つクエリが分布する傾向にあった. この観察結果は, 学習データの特性が埋め込みモデルに反映されていることを示唆する.

また, 提案モデルがクエリの表層形に依らず内在的な意図を捉えている傾向を定量的に確認するため, QR タスクにおいてクエリペアの編集距離ごとの性能評価を実施した (図 2). 表層形が類似したペア (例: 「ユニバ 最寄り」と「ユニバ 最寄り駅」) では全モデルで比較的良好な性能を示す一方で, 表層形の差異が増えるにつれて (例: 「ユニバ 最寄り」と「USJ アクセス」), ベースラインの性能は大きく低下した. 特に教師なし SimCSE や fastText で著しく性能劣化しており, これらの教師なしモデルが表層形の異同に対して非常に敏感であることが分かった. 一方で提案モデルは表層形の差異に頑健でクエリの内在的な意図を適切に捉えられる点が確認された.

6 結論と今後の展望

本研究では, ユーザ行動ログに基づく対象学習により汎用クエリ埋め込みを構築する UBIQUE を提案した. 提案モデルは 4 つの QU タスクにおける比較実験で最先端手法を上回る性能を発揮し, さらにクエリの表層形の変化に対し高い頑健性を示した. 本研究は日本語検索システムに焦点を当てたが, 他言語での有効性評価は今後の課題となる. また, クリックログとセッションログのそれぞれの強みの統合によりさらなる性能向上が期待できる.

謝辞

本稿の執筆にあたり、LINEヤフー研究所の赤崎智さんには貴重なご意見をいただきました。深く感謝致します。

参考文献

- [1] Ben Shneiderman, Don Byrd, and W. B Croft. Clarifying Search: A User-Interface Framework for Text Searches. Technical report, 1997.
- [2] Tessa Lau and Eric Horvitz. Patterns of Search: Analyzing and Modeling Web Query Refinement. In **UM 1999**, pp. 119–128, 1999.
- [3] Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N. Bennett, Nick Craswell, and Saurabh Tiwary. Generic Intent Representation in Web Search. In **SIGIR 2019**, p. 65–74, 2019.
- [4] Homa Baradaran Hashemi. Query intent detection using convolutional neural networks. 2016.
- [5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In **EMNLP 2021**, pp. 6894–6910, 2021.
- [6] 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, 2024.
- [7] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 Text Embeddings: A Technical Report, 2024.
- [8] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards General Text Embeddings with Multi-stage Contrastive Learning, 2023.
- [9] SB Intuitions. Sarashina-embedding-v1-1b, 2024.
- [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In **ICML 2020**, Vol. 119, pp. 1597–1607, 2020.
- [11] Chien-Kang Huang, Lee-Feng Chien, and Yen-Jen Oyang. Relevant term suggestion in interactive web search based on contextual information in query session logs. **J. Am. Soc. Inf. Sci. Technol.**, Vol. 54, No. 7, p. 638–649, 2003.
- [12] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang. Context-aware query classification. In **SIGIR 2009**, p. 3–10, 2009.
- [13] Doug Beeferman and Adam Berger. Agglomerative clustering of a search engine query log. In **KDD 2000**, p. 407–416, 2000.
- [14] Haoming Jiang, Tianyu Cao, Zheng Li, Chen Luo, Xianfeng Tang, Qingyu Yin, Danqing Zhang, Rahul Goutam, and Bing Yin. Short Text Pre-training with Extended Token Classification for E-commerce Query Understanding, 2022.
- [15] Juanhui Li, Wei Zeng, Suqi Cheng, Yao Ma, Jiliang Tang, Shuaiqiang Wang, and Dawei Yin. Graph Enhanced BERT for Query Understanding. In **SIGIR 2023**, p. 3315–3319, 2023.
- [16] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In **CVPR 2006**, Vol. 2, pp. 1735–1742, 2006.
- [17] Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. Pairwise Supervised Contrastive Learning of Sentence Representations. In **EMNLP 2021**, 2021.
- [18] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Fast, Effective, and Self-Supervised: Transforming Masked Language Models into Universal Lexical and Sentence Encoders. In **EMNLP 2021**, 2021.
- [19] Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. ESIMCSE: Enhanced Sample Building Method for Contrastive Learning of Unsupervised Sentence Embedding. In **COLING 2022**, pp. 3898–3907, 2022.
- [20] Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. On Sampling Strategies for Neural Network-based Collaborative Filtering. In **KDD 2017**, p. 767–776, 2017.
- [21] Bruce Croft, Donald Metzler, and Trevor Strohman. **Search Engines: Information Retrieval in Practice**. 1st edition, 2009.
- [22] Yupin Huang, Jiri Gesi, Xinyu Hong, Han Cheng, Kai Zhong, Vivek Mittal, Qingjun Cui, and Vamsi Salaka. Behavior-driven query similarity prediction based on pre-trained language models for e-commerce search. In **SIGIR 2023 Workshop on eCommerce**, 2023.
- [23] Bruno M. Fonseca, Paulo Golgher, Bruno Pôssas, Berthier Ribeiro-Neto, and Nivio Ziviani. Concept-based interactive query expansion. In **CIKM 2005**, p. 696–703, 2005.
- [24] Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search, 2022.
- [25] Kobayashi Koga, Shengzhe Li, Akifumi Nakamachi, and Toshinori Sato. LINE DistilBERT Japanese. 2023.
- [26] Shengzhe Li, Masaya Ohagi, and Ryokan Ri. JMTEB: Japanese Massive Text Embedding Benchmark, 2024.
- [27] Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. Japanese SimCSE Technical Report, 2023.
- [28] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese General Text Embeddings, 2024.
- [29] OpenAI. New embedding models and API updates, 2024.
- [30] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. **TACL**, Vol. 5, pp. 135–146, 2017.
- [31] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library, 2024.
- [32] 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 **EMNLP 2020**, pp. 6769–6781, 2020.

表 3: UBIQUE で使用した学習クエリペアの実例.

クエリ 1	クエリ 2
<i>Click</i>	
海外旅行 クレカ	海外に強い クレジットカード
最も長い 蛇	10m 蛇
<i>Session</i>	
コンビニ大根サラダ	コンビニ 大根サラダ アレンジ
アメリカ 80 万 旅行	アメリカ 1 週間 旅費

表 4: 評価ベンチマークの統計量. 人手でのアノテーションにより構築されている.

タスク	データ数	平均紐づき数
同義クエリ抽出	5,000	1
クエリ推薦	951	8.2
クエリ分類	1,456	-
短文リランキング	4,667	25.5

表 5: ハード負例導入の結果.

モデル	QR	QS	QC	SR
UBIQUE _{click}	91.4	91.2	85.8	90.6
w/ ハード負例	90.2	90.1	86.2	90.1

A ハード負例の検証

UBIQUE モデルのさらなる性能向上を目指し, ハード負例の導入を検証した. 先行研究 [32] に従い, アンカークエリと編集距離は小さいが, 紐づくクリック URL 集合は重複しないクエリをハード負例として抽出し, これを追加の負例とした対照学習を UBIQUE_{click} に対し適用した.

しかしその結果, 平均的なタスク性能はわずかに低下した (表 5). QR タスクでの定性評価の結果, アンカーと表層形が類似した負例が存在する困難なケースで性能が向上している一方で, それ以外の多くのケースでは抽出性能が下がる傾向にあった. これは表層形が類似したクエリを常に負例として用いることで学習が困難になったためだと推察される. 今後の展望として編集距離以外の指標を使用したハード負例抽出の検討が必要となる.

B ジャックカード係数の有効性検証

ジャックカード係数を用いたクエリペア抽出の有効性を検証するため, ジャックカード係数を用いずに抽出した擬似正例ペアで学習したモデルとの比較実験を行った. その結果, 表 6 に示すようにジャックカード係数の導入でモデルの平均性能の向上が確認され,

表 6: ジャックカード係数の有無による性能比較.

モデル	QR	QS	QC	SR
UBIQUE _{click}	91.4	91.2	85.8	90.5
w/o ジャックカード係数	89.4	90.8	85.6	90.4
UBIQUE _{session}	71.9	92.6	86.9	90.3
w/o ジャックカード係数	58.7	92.1	87.1	90.1

表 7: ハイパーパラメータの一覧.

種類	UBIQUE	教師なし SimCSE
ベースモデル	DistilBERT	DistilBERT
ミニバッチサイズ	1,024	1,024
最適化アルゴリズム	AdamW	AdamW
学習率	2e-4	3e-5
エポック数	5	5
学習率スケジューリング	Linear	Linear
Warmup ステップ数	1%	1%
ドロップアウト確率	0.1	0.2

ジャックカード係数導入によるノイズペア抑制の重要性が確かめられた.

C 実験設定の詳細

学習データでは過学習を防ぐため重複排除を行った. また, リーク防止のため, 評価データに含まれるペアを削除した. アダルトコンテンツに関連するクエリは多様な URL へのクリックや短時間での意図変化を生じやすく, ノイズペアを多く生成する傾向にあったため, 事前に定義したアダルト用語を含むペアを削除した.

UBIQUE と教師なし SimCSE のハイパーパラメータ一覧を表 7 に示す. 学習率は 5,000 件の QR タスク開発セットを用い {2e-4, 3e-4, 3e-5} の範囲で探索した. また, 4,000 ステップごとに開発セットで評価を行い最良性能のチェックポイントを選択した. 教師なし SimCSE では学習率に加え, ドロップアウト率を {0.05, 0.1, 0.2} の範囲で探索した. 学習は 4 枚の NVIDIA V100 で実施し, 16 時間要した.

ベースラインにおいて, SimCSE_{large} では [CLS] プーリング, LINE DistilBERT と Ruri, mE5 では平均プーリングを利用した. 全てのモデルで最大系列長は 512 とした. また, Ruri と mE5 では論文 [7, 28] に倣いタスクの対称性に依りて検索元か検索対象かを区別する文字列を入力テキストの先頭に付与した. fastText では, MeCab³でクエリを分割し, 各トークンに対するベクトルの平均を使用した. 学習では 300 次元の Skip-gram モデルを利用した.

3 <https://taku910.github.io/mecab/>