E コマースにおける検索クエリの整形と属性値抽出への適用

中山 祐輝　Chen Zhao　Erick Mendieta　村上 浩司　新里 圭司
楽天グループ株式会社 楽天技術研究所
{yuki.b.nakayama,chen.a.zhao,erick.mendieta}@rakuten.com
{koji.murakami,keiji.shinzato}@rakuten.com

概要

E コマースでの商品検索の質を向上させるために，クエリを意味のあるまとまりに整形する研究が盛んである．しかし，既存研究のほとんどは，一つの操作に絞っており，かつ英語を対象とした研究である。また，クエリ整形手法を E コマースドメインのタスクに応用する試みは，ほぼなされていない．本稿では，三つの操作を考慮する日本語のクエリ整形手法を提案し，実験によってその有効性を示す．さらに，本手法を E コマースの属性値抽出に適用し，クエリ整形の重要性を示す．

1　はじめに

EC サイトでの購買活動が高まり，商品検索のクエリが膨大に蓄積されている．蓄積されたクエリは，ユーザの需要把握や，検索の質向上を目的としたクエリ補完や属性値抽出などのタスクに有効利用できる．しかし，クエリは語の境界が不適切な場合があり，検索結果やタスクに悪影響を及ぼす．例えば，「アディダスマスク」というクエリは，「アディダス」というブランドと「マスク」という分野相当の語が混在しているため，ゼロヒット [1] を引き起こす可能性がある．一方，「肉■の■とみや」というクエリ（■は空白）は，「肉のとみや」という店名を誤って分割しているため，当該肉店に関連のない商品が表示される可能性がある．

このような問題を軽減するために，E コマースドメインにおいてクエリの語に表 1 のような操作を行い，クエリを意味のあるまとまりに整形するするクエリ整形の研究が盛んである [2, 3, 4]. しかし、既存研究の手法には，(1) 一つの操作しか対象としない研究がほとんど，(2) 日本語を対象とした研究はない，(3) 改良手法を NLP タスクに適用した研究はない という三つの課題がある．そこで，本稿では日本語を対象とした三つの操作を考慮するクエリ改良手法を提案する．本稿の貢献は，E コマースドメインにおける日本語に対応した全く新しいクエリ改良の手法を提案し，ベースライン手法と比較して優れた成果が得られたこと，クエリ改良手法を NLP タスクに応用し，形態素解析を事前処理に用いた手法と比べて精度の向上を達成したことである．我々は，提案手法を NLP タスクとして代表的な属性値抽出に適用する．

表 1　本研究で対象とするクエリ整形

操作	オリジナル	整形後
分割	アディダスマスク	アディダス■マスク
結合	肉■の■とみや	肉のとみや
セグメント	new balance shoes	“new balance” shoes

2　関連研究

E コマースにおけるクエリ整形手法として，Salehi ら [5] は，英語を対象としたセグメント化を対象としたクエリ改良の手法を提案した．Li ら [6] は，中国語を対象としたクエリ改良の手法を提案した．しかし，英語や中国語とは異なり，日本語は分かち書きされておらず，異なる三つの字種を持つことから，既存手法を日本語のクエリに適用することは困難である．Guo ら [7] は，表 1 の三つの操作を網羅したクエリ改良の手法を提案し，適合判定の IR タスクに適用した．それに対して，我々は属性値抽出という NLP の代表的なタスクに焦点を当てる．

E コーマスの検索クエリを対象とした様々な属性値抽出手法がある．Kozareva ら [8] と Zhai ら [9] は，それぞれ LSTM-CRF に基づく手法と文法規則を用いて，ブランドと商品ジャンルを認識する手法を提案した．Jiang ら [10] は，人手による少数の学習データに品質改善を施した擬似学習データを追加することで，属性値抽出の改善を図った．Cheng ら

[11] は，人手による少数の学習データで学習されたモデルを用いて，品質の高い擬似データの生成を繰り返して学習を行うフレームワークを提案した．しかし，クエリ整形という事前処理の観点から精度向上を図った研究はない．

3 提案手法

3.1 概要

図 1 に提案手法の概観を示す．空白によって分割されたクエリの語 x_i からなる入力系列 $\boldsymbol{x} = x_1 ... x_l$ が与えられた時，我々の目標は整形された語の系列 $\boldsymbol{y} = y_1 ... y_L$ を予測することである．l と L はそれぞれ，入力系列の語数と出力系列の語数である．本研究では，表 1 にある分割，結合，セグメント化の三つの操作に焦点を当てる．具体的には，整形操作の系列 $\boldsymbol{o} = o_1 ... o_l$, ただし $o_i \in \{Split, Merge, Segment, Single\}$ を予測するタスクを解く．ここで，*Single* は操作を必要としないクエリの語に割り当てられる．整形操作の系列が決定されれば，$\boldsymbol{o}$ を用いて，$\boldsymbol{x}$ を整形後の系列に変換できる．

Wikipedia のタイトルは，製品名やブランド名など E コーマスに関わる固有表現を多く含み，それらは整形後のクエリ語となる可能性が高い．また，二つのクエリ語が頻繁に隣接して共起すれば、それらからなる語も整形後のクエリ語となる可能性が高いと考える．以上の二つの考えに基づいて，まずキーワードマッチング（3.2 節）によって，x_i の操作 o_i を *Merge*,*Segment*,*Single* のいずれかに分類する．もしクエリ語が，いずれのキーワードにもマッチしなかった場合は，*Uncetain* label を暫定的に付与し，整形後のクエリ $\boldsymbol{x_r}$ を入力とした機械学習（3.3 節）によって *Split*，*Merge*，*Segment*，*Single* のいずれかに分類される．

3.2 キーワードマッチング手法

まずは，以下の手順でクエリ語の系列と収集されたキーワードとのマッチングをとる．

1 $\boldsymbol{o}$ を [*Uncetarin*, ..., *Uncertain*] に初期化
2 *Uncertain* ラベルを持つ全ての系列 $x_i ... x_j$ を取得
3 各系列とキーワードとの最左最長マッチをとる
4 $x_i ... x_j$ がキーワードにマッチすれば，$x_i ... x_j$ に対応する *Uncertain* ラベルを *Merge*，*Segment*，*Single* のいずれかに更新する

マッチングは，三つのキーワードリストを用いて，信頼性を考慮し以下の順番で行われる．

1. E コーマスドメインの属性値リスト (ORD) ブランド名，シリーズ名，サイズ，色など E マースドメインに関連する約 7 万 8 千語からなる属性値のリストを構築する．

2. ウィキペディアタイトルと本文のキーワード (Wiki) 2021 年 3 月時点最新の日本語版ウィキペディアダンプと，2021 年 2 月時点最新の英語版ウィキペディアダンプから，それぞれ約 190 万のタイトルと約 1273 万のタイトルを収集する．また，2021 年 2 月時点最新のからを収集する．さらに，日本語版ウィキペディアの本文から形態素解析によるトークン N-gram（$1 \leq N \leq 4$）を抽出し，頻繁に隣接して共起する N-gram ペアをキーワードとして登録した．二つの n-gram トークン w_1，w_2 の共起度合いを測る尺度として，式 (1) の相互情報量を用いる．

$$\mathrm{PMI}(w_1, w_2) = \log_2 \frac{N * \mathrm{Freq}(w_1, w_2)}{\mathrm{Freq}(w_1) * \mathrm{Freq}(w_2)} \quad (1)$$

N は語の総数であり，freq(w_1,w_2) は，w_1，w_2 の順で両者が隣接して出現する頻度である．

3. クエリ語のトークンリストと接頭辞/接尾辞リスト（Query） クエリ語内の語の共起性を考慮する．クエリ語の形態素解析で得られたトークンに対して，隣接トークンペアの PMI を式 (1) で求め，閾値を上回るペアを収集する．

3.3 マッチング結果を考慮した機械学習に基づく手法

3.2 節で整形された *Uncertain* ラベルを持つクエリの系列 $\boldsymbol{x_r} = x_1 ... x_m$ が与えられた時，*Uncertain* ラベルの分類を文字単位の BIOES チャンキングとして定式化する．具体的には，空白を含む $\boldsymbol{x_r}$ の文字系列 $\boldsymbol{c} = c_{1,1} ... c_{m,n}$ に対して，系列 $\boldsymbol{z} = z_{1,1} ... z_{m,n}$, $z_{i,j} \in \{B, I, O, E, S\}$ を予測するタスクを解く．ここで，$c_{i,j}$ は，i 番目のクエリ語内の j 番目の文字である．最後に，出力系列 $\boldsymbol{z}$ から *Uncertain* ラベルに対応するクエリ語の部分を抜き出し，ラベルを *Split*,*Merge*,*Segment*,*Single* のいずれかに更新する．チャンキングの手順は，文字系列の埋め込み表現の獲得，マッチング手法の結果を考慮したクエリ語の埋め込み表現の結合，全結合，CRF の 4 層からなる．本稿では前半の二つのステップについて説明する．まずは，周辺文字を意識するために，

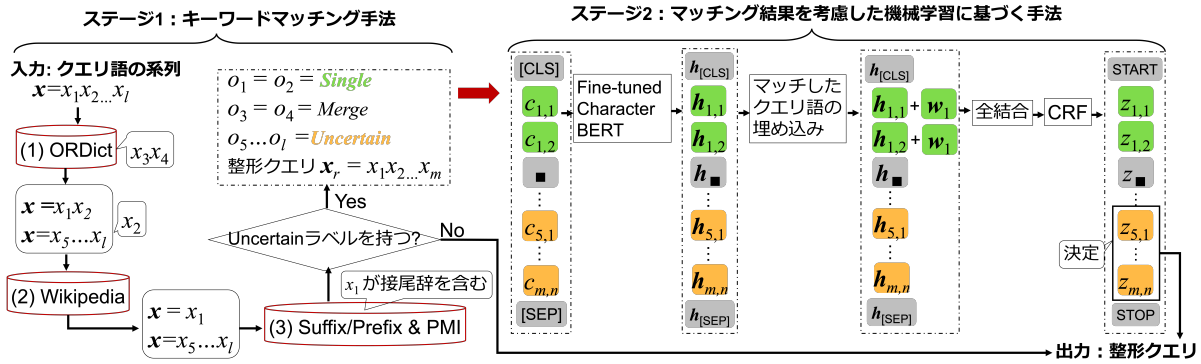

図 1　提案手法の概観 (■ は空白を表す)

Character-BERT を用いて文字系列の埋め込み表現 $\boldsymbol{H} = [\boldsymbol{h}_{[\mathrm{CLS}]}, \boldsymbol{h}_{1,1}, ..., \boldsymbol{h}_{m,l}, \boldsymbol{h}_{[\mathrm{SEP}]}]$ を獲得する．ここで，$\boldsymbol{h}_{i,j}$ は，文字 $c_{i,j}$ に対応する Character-BERT による埋め込み表現である．次に，クエリ語 x_i の埋め込み表現を，$c_{i,j}$ の埋め込み表現に結合させる．提案手法では，以下の二つの理由でキーワードマッチングでマッチしたクエリ語のみの埋め込み表現を考える．一つ目は，マッチング情報をモデルの素性として加えるためである．二つ目は，不適切な埋め込み表現を避けるためである．一般的に，文字レベルの系列ラベリングタスクにおいて，単語の埋め込み表現は Watson ら [12] のように対応する全ての文字に考慮される．しかし，「アディダスマスク」のような語の境界が不明な問題を扱っているため，不必要なクエリ語の埋め込みを文字の埋め込みに結合する可能性がある．クエリ語 x_i の埋め込み表現は，構成される文字における埋め込み表現の平均によって得る．

4　評価実験

4.1　クエリ整形

日本語のクエリ整形における公開データセットは存在しないため，楽天市場の商品検索で使用された 10,239 件のクエリに対して整形操作のスパンを人手で特定した．10,239 件のクエリのうち，8,240 件（学習 : 7325 件，開発 : 825 件）は Character-BERT のファインチューニングに用いた．Character-BERT は，東北大学が提供している事前学習モデル[1] を用いた．また，999 件のクエリを 3.2 節の PMI の閾値を決定するために用いた．残り 1000 件のクエリを評価セット $|Q|$ として用いた（表 2）．Guo ら [7] にしたがい，(2) 式の正解率を評価尺度とした．

$$\frac{1}{|Q|}\sum_{q\in Q}\frac{q\text{ 内で整形操作が正解である語の数}}{\text{クエリ } q \text{ の語数}} \qquad (2)$$

表 3 に実験結果を示す．ベースライン手法として，固有表現抽出でよく利用される Character-BERT CRF を用いた．提案手法の正解率は 88.12%となり，ベースラインより 4.54 ポイント優れた正解率を達成し，両側 t 検定により両者の間には 1%の統計的有意差があった．また，アブレーションテストを通してキーワードマッチング手法における各ステップの有効性を確認した．さらに，全てのクエリ語の埋め込み表現を追加した場合と比べて，クエリ語の埋め込み表現の追加をキーワードリストに含まれる語に限定した場合，正解率が 87.18%から 88.12%に向上し，埋め込み対象のクエリ語を限定したとことの有効性を確認した．

提案手法によって誤った整形操作に分類された 272 のクエリ語に対して分析を行った．272 語中，227 語がステージ 1 にで誤りが発生した．そこで，STAGE1 においてキーワードリストのアブレーションテストを行うことで，どのリストが誤りを引き起こしているかを調査した．227 語中，119 語が単一のリストによって引き起こされた誤りとなり，それぞれ ORD が 19 語，Wiki が 37 語，63 語がクエリからのキーワードに起因する誤りだった．ORD と Wiki については，キーワードが網羅されておらず，キーワードの部分一致を引き起こしたことが原因である．例えば，二つのクエリ語「出産」と「祝い」が一つのクエリ語「出産祝い」に結合されるべきところが，「出産祝い」がリストに登録されておらず，「出産」と「祝い」というキーワードにマッチし，*Single* ラベルが付与されたため，結合されなかった．クエリからのキーワードに起因する 63 語

1)　https://huggingface.co/cl-tohoku/bert-base-japanese-char-v2

の誤りのうち，61 語は隣接するトークン間の PMI が意図せず閾値を上まわった結果，両者の結びつきが強くなり，複数のクエリ語に分割されなかった．ステージ 2 では，45 件の誤りが発生した．誤りの傾向として，「リバーシブル D-86」や「かみつき！ T-レックス」のようなアルファベット，数字，カタカナなど様々な字種で構成される商品名が複数のクエリ語に誤って分割されていた．このような複雑なケースに対応することが今後の課題である．

表 2　クエリ整形における評価データの基礎情報

整形操作	サンプル数	
	語レベル	クエリレベル
結合/セグメント	238	97
分割	150	144
必要なし	1,995	763
総数	2,383	1,000

表 3　クエリ整形の実験結果

手法		正解率
ステージ 1	ステージ 2	
CBERT-CRF	N/A	83.58%**
マッチング	N/A	81.87%**
w/o ORD	N/A	71.70%**
w/o Wiki	N/A	79.37%**
w/o Query	N/A	70.94%**
マッチング	CBERT-CRF	87.36%*
マッチング	w/ 全てのクエリ語	87.18%**
マッチング	w/ マッチしたクエリ語のみ	**88.12%**

**:1%有意差あり，*：5%有意差あり

4.2　ブランド抽出

クエリ整形手法をブランド抽出に適用し，E コマースドメインにおける提案手法の有効性を検証した．ブランド抽出を文字レベルの系列タギング問題として定式化し，再現性を担保するために FLAIR [13] で公に利用できるモデル [2] を用いた．モデル内の埋め込み層では，以下の Flair 埋め込みと単語埋め込みを用いた．

単語埋め込み　楽天市場の商品検索で入力された 2018 年分のクエリをランダムに 240 万件抽出し，word2vec によって次元数 300，最小頻度 100 で事前学習した．

Flair 埋め込み　Flair は，文脈の前後の関係を意識した強力な文字列の埋め込み手法である [14]．楽天市場の商品検索で入力された 2018 年分のクエリをランダムに 100 万件抽出し，事前学習を行う．文字列ベクトルの次元数は，前方からのモデル 2048 次元と後方からのモデル 2048 次元を結合した 4096 である．

ファインチューニングには，人手による 9000 件のラベル付きクエリと辞書ベースによって付与された擬似ラベル付き 560K のクエリを学習に用い，別途 2,248 件の人手によるラベル付きクエリを開発用セットとした．表 4 に評価データの詳細を示す．表 5 に実験結果を示す．MeCab と Unidic 辞書による形態素解析結果をモデルの入力とした場合と比較して，提案手法を適用したモデルは F1 スコアで 8.9 ポイントの精度向上を達成した．

表 6 に，クエリ整形の提案手法がブランドの抽出に成功し，形態素解析（MeCab+Unidic）による事前処理が失敗した例を示す．提案手法は，クエリ「カネテツ■デリカ■フーズ」を意味のある一語に結合させたことで，「カネテツデリカフーズ」をブランドとして認識できた．この結果から，提案手法のクエリの整形化によって文脈情報を保持し，かつ不必要な単語埋め込みを避けることができていると考える．

表 4　ブランド抽出における評価データの基礎情報

(a) ブランドの有無

有無	クエリ件数
ある	901
ない	1324
Total	2225

(b) ブランドの含有数

ブランド数	クエリ件数
一つ	811
二つ	83
三つ	7
総クエリ件数	901
総ブランド数	998

表 5　ブランド抽出の実験結果

Method	P	R	F1
MeCab+Unidic	51.5%	65.0%	57.5%
クエリ整形の提案手法	59.2%	75.6%	**66.4%**

表 6　ブランド抽出て提案手法が成功した例（カネテツデリカフーズ）

Query words	オリジナル	カネテツ■デリカ■フーズ
	MeCab	カネテツ■デリカ■フーズ
	本手法	カネテツデリカフーズ
Results	MeCab+Unidic	カネテツ
	本手法	カネテツデリカフーズ

5　おわりに

本稿では，E コマースにおける三つの操作を考慮した日本語のクエリ整形手法を提案し，実験によってその有効性を示した．また，提案手法をブランド抽出に適用し，形態素解析器による事前処理と比較実験を行い，クエリ整形の重要性を示した．誤り事例の解決方策を考案することが今後の課題である．

2）https://github.com/flairNLP/flair/blob/master/flair/models/sequence_tagger_model.py#L26

参考文献

[1] Yuki Amemiya, Tomohiro Manabe, Sumio Fujita, and Tetsuya Sakai. How do users revise zero-hit product search queries? In *European Conference on Information Retrieval*, pp. 185–192. Springer, 2021.

[2] Saša Hasan, Carmen Heger, and Saab Mansour. Spelling correction of user search queries through statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 451–460, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[3] Saurav Manchanda, Mohit Sharma, and George Karypis. Intent term selection and refinement in e-commerce queries. *arXiv preprint arXiv:1908.08564*, 2019.

[4] Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. Query reformulation in e-commerce search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, p. 1319–1328, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Bahar Salehi, Fei Liu, Timothy Baldwin, and Wilson Wong. Multitask learning for query segmentation in job search. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '18, p. 179–182, New York, NY, USA, 2018. Association for Computing Machinery.

[6] Zhao Li, Donghui Ding, Pengcheng Zou, Yu Gong, Xi Chen, Ji Zhang, Jianliang Gao, Youxi Wu, and Yucong Duan. Distant supervision for e-commerce query segmentation via attention network. In *Intelligent Processing Practices and Tools for E-Commerce Data, Information, and Knowledge*, pp. 3–19. Springer, 2022.

[7] Jiafeng Guo, Gu Xu, Hang Li, and Xueqi Cheng. A unified and discriminative model for query refinement. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 379–386, 2008.

[8] Zornitsa Kozareva, Qi Li, Ke Zhai, and Weiwei Guo. Recognizing salient entities in shopping queries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 107–111, Berlin, Germany, August 2016. Association for Computational Linguistics.

[9] Ke Zhai, Zornitsa Kozareva, Yuening Hu, Qi Li, and Weiwei Guo. Query to knowledge: Unsupervised entity extraction from shopping queries using adaptor grammars. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, p. 255–264, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Haoming Jiang, Danqing Zhang, Tianyu Cao, Bing Yin, and Tuo Zhao. Named entity recognition with small strongly labeled and large weakly labeled data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1775–1789, Online, August 2021. Association for Computational Linguistics.

[11] Xiang Cheng, Mitchell Bowden, Bhushan Ramesh Bhange, Priyanka Goyal, Thomas L. Packer, and Faizan Javed. An end-to-end solution for named entity recognition in ecommerce search. In *AAAI*, 2021.

[12] Daniel Watson, Nasser Zalmout, and Nizar Habash. Utilizing character and word embeddings for text normalization with sequence-to-sequence models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 837–843, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[13] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, 2019.

[14] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.