E コマースにおける商品用途表現の抽出とグルーピング

梶原奨　井上翔太　岡林遥平　稲田和明　張信鵬
株式会社 MonotaRO
{sho.kajihara,shota.inoue,yohei.okabayashi,kazuaki.inada,xinpeng_zhang}@monotaro.com

概要

商品の利用用途などを説明した文は、商品検索などの特定のサービスへの利用を前提として作成された構造化データとは異なり、E コマースのサービスでの活用が難しい。しかし構造化されていないデータ内にも、ユーザーが商品を探す際に有用な商品の使い道を記した表現が多数含まれているため、そのような表現を抽出・利用することで商品検索の質を向上させることができると考えられる。そこで本稿では、ユーザの商品検索に有用な 8 種の商品用途表現ラベルを定義し、実際の商品情報に含まれる構造化されていない文にアノテーションした上で、系列ラベリングモデルを作成することで、商品検索に有用な商品の用途表現を抽出する。さらに実際の E コマースのサービスで活用することを想定し、類似する商品用途表現のグループ化を実施する。

1　はじめに

E コマースに利用されるデータは、商品検索などの特定のサービスに応じて事前定義されたスキーマに沿って、意味の通じる最小単位に分割された構造化データで表現されることが多い。しかし、中には商品の説明文のような構造化されていないデータも存在する。たとえば「剪定ハサミ」の商品には、(材質, ステンレス)、(大きさ, 20 インチ) のような key-value 形式の構造化された商品情報だけでなく、「庭木の剪定作業、ガーデニング・盆栽・植木のお手入れに。」のような構造化されていない商品情報も存在する。

上述のような構造化されていない商品情報の中には、「庭木の剪定作業」といった使い道、「ガーデニング」といった環境、「盆栽」「植木」といった商品の使用対象などの、商品検索に有用な情報が記述されていることがある。これらの有用な情報を抽出することができれば、構造化されたデータと同様に検索条件のフィルタリングなど E コマースのサービスに活用でき、ユーザーの利便性を高められると考えられる。

表 1　商品用途表現ラベル

ラベル名	説明（例）
使用者	誰向けの商品か（乳児, 製造業）
使用環境	どこで使うか（工場, 屋外）
使用対象	何に使うか（枝, アルミ）
使用目的	何のために使うのか（研磨, 固定）
使用 タイミング	自然的に起きる出来事に使う （季節, 地震）
使用 イベント	人為的に起きる出来事に使う （運動会, 放電加工）
特徴	商品固有の特性（防水, 省エネ）
機能	商品が果たす役割（防音材, 留め具）

Alicoco[1] では、ユーザーのニーズとして Time, Location, Object, Function, Incident, Cate/Brand, Style, Intellectual Property の 8 種類の関係を定義・活用することで、商品間の知識グラフを作成した。この知識グラフは Alibaba グループの商品検索や推薦に応用されており、E コマースで広く利用されている階層的なカテゴリによる商品分類よりも、ユーザーのニーズをより深く汲み取ることに成功している。

そこで本稿では、商品に記述されている文から、意味の通じる最小粒度の表現（以降、**商品用途表現**と呼ぶ）を抽出し、商品と商品用途表現間の関係を整理する。まず、商品用途表現の役割として 8 種のラベルを定義し、実際の E コマースの商品情報として記述される文に対してアノテーションする。次にラベル付きデータを用いて教師ありの系列ラベリングモデルを作成し、ラベル付けしていないデータから商品用途表現を抽出する。最後に、実際の商品検索サービスへの活用を想定し、類似する商品用途表現のグループ化を実施する。

2　商品用途表現のラベリング

Alicoco[1] におけるニーズを参考に、商品用途表現を分類するラベルを表 1 に定める。

表 1 で定めたラベルのアノテーションの例を図 1

に示す。アノテーション対象として、monotaro.com[1)]の商品の中で「用途」として定義された属性に含まれる約 10 万文の内、商品に付与されているカテゴリを網羅できるように層化抽出した 10%を利用した。アノテーションは 2 人の作業者によって実施し、作業中のアノテーション対象の文だけでなく掲載元の web ページの閲覧や Google などの検索エンジンの利用を可能とした。またアノテーションの精度を確保するため、100 件のデータに対して正しくラベル付けできるようトレーニングした。さらにアノテータ 2 人で相互にレビューする体制を取ることで、ラベリング結果の一貫性とアノテーションの精度を高めた。

アノテーションしていない残りの 90%のデータに対する商品用途表現ラベルの同定には、図 2 に示す BERT のネットワークの最終層に CRF 層を追加した BERT+CRF モデルを用いる [2]。まず、入力された文を WordPiece トークナイザでトークンに分割した後、各トークンに対して BERT モデルで表 1 に定めた各ラベルの予測確率を計算する。その後、CRF 層で入力全体のラベル遷移を考慮して最終的なラベルを予測する。なお予測対象のラベルには、各商品用途表現ラベルに B および I を組み合わせたラベルと O の 17 種類が存在し、B は各商品用途表現ラベルが付与された表現の先頭、I は各商品用途表現ラベルが付与された表現の先頭以外、O は商品用途表現ラベルが付与されていないトークンを意味する。

また本稿ではモデルの精度改善のために、Least Confidence を用いた能動学習を導入する [3]。アノテーションの半分が終わった時点で学習したモデルを用いて、アノテーション対象外のデータをラベル付けし、各トークンの予測確率に対して Least Confidence を計算する。そして、文内の最も小さな Least Confidence をその文のスコアとし、スコアの低い 2000 文をアノテーション対象として追加する。

3 商品用途表現のグルーピング

商品用途表現には異なる表記で同じ意味を持つものが多数存在する。商品検索における絞り込み条件などの実際の E コマースのサービスを想定すると、たとえば、「穴や破損個所を修復」と「つなぎ合わせ」のような表現が大きく異なる商品用途表現も、同じ意味として 1 つにグループ化しておく必要がある。

1) https://www.monotaro.com/

単語やフレーズなどの意味的な類似度を計算する手法の 1 つとして、計算対象の表現をそれぞれベクトル化しそれらの距離を比較する手法が近年よく利用されるが、その類似度は主に含まれる文字列や単語の種類が大きく影響することが報告されている [4]。

そこで本稿では、以下に示す 3 段階のルールベースのグルーピングを適用する。1 段階目として、文字列の部分一致により商品用途表現をグループ化する。次に 1 段階目で形成されたグループを、商品用途表現に紐づいている商品のカテゴリ情報によって分割する。最後に日本語 WordNet[5] の概念と部分関係の情報を用いて細分化する。

3.1 部分一致と同義語によるグルーピング

1 段階目のグルーピングとして、文字列情報に着目した以下に示す手順を適用する。

1. 他の商品用途表現と文字列が部分一致しない商品用途表現を抽出する。
2. (1) で抽出した商品用途表現と文字列が部分一致する商品用途表現でグループを形成する。
3. 同義語辞書を用いて (1) で抽出された商品用途表現同士をマージする。

たとえば、(1) で「研磨」「研ぐ」のような商品用途表現が抽出され、(2) で「研磨」と部分一致する「金属研磨」「プラスチック研磨」、「研ぐ」と「彫刻刀を研ぐ, 整えて研ぐ」などを集約してグループを形成し、(3) で「研磨」と「研ぐ」が同義語として判定されて、「研磨、研ぐ、金属研磨、プラスチック研磨、彫刻刀を研ぐ, 整えて研ぐ」というグループが出力される。ただし (1) では、使用目的・機能・使用イベントの 3 つのラベルに属する商品用途表現が述語によって構成されることを考慮し、用言の有無を確認することでグルーピングの精度を高める。たとえば、「切断」と「切断を防ぐ」は部分一致しているが、用言を比較すると「切断」と「防ぐ」と異なるため、グループ化させない。

3.2 商品カテゴリによるグルーピング

商品カテゴリは末端ほど細分化された意味を持つ木構造で表現されており、たとえば「鋏」のカテゴリでは、親に「農具」、子に「剪定鋏」「収穫鋏」「替え刃」などを持つ。このとき商品カテゴリの根はすべてのカテゴリの祖先に相当するが、2 段階目のグ

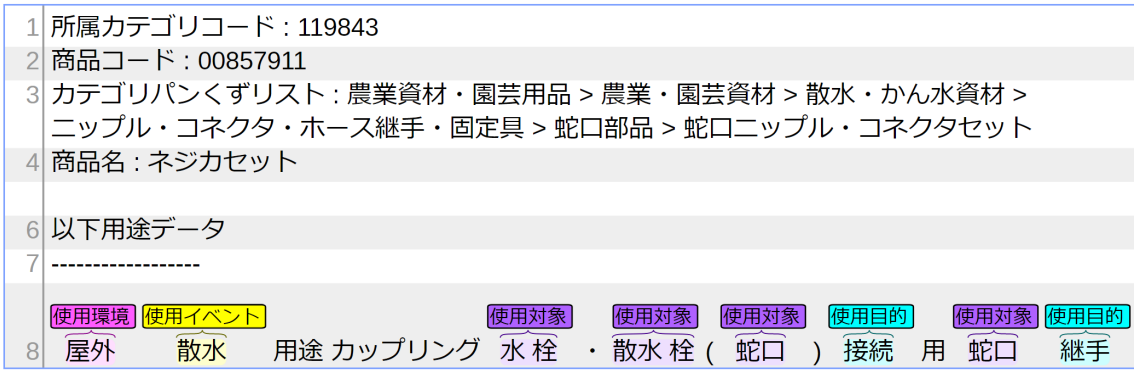

図 1　brat によるアノテーションの様子

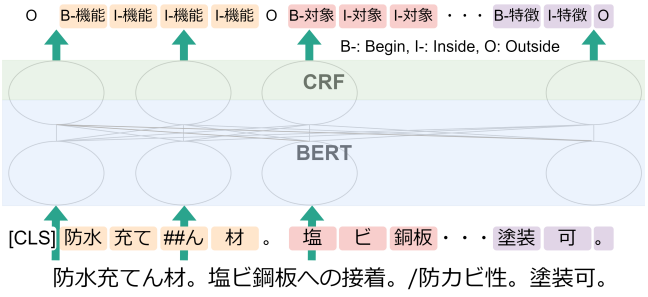

図 2　BERT+CRF モデルによるラベル付け

ルーピングとして、商品用途表現の抽出元である商品と商品カテゴリの根の子に相当するカテゴリ (以降、モールカテゴリと呼ぶ) の関係を用いる。

まず、各商品用途表現に対応するモールカテゴリを決める。商品用途表現は複数の商品から同じ表現が獲得されるため、1 つの商品用途表現は複数の商品カテゴリとそれらのモールカテゴリを持つが、最多のモールカテゴリを各商品用途表現の代表とする。その後、3.1 で形成された各グループの中の商品用途表現をモールカテゴリごとに分割する。たとえば、「金属研磨」「プラスチック研磨」が「切削工具・研磨材」、「電解研磨」が「スプレー・オイル・グリス/塗料/接着・補修/溶接」のモールカテゴリに属し、3.1 節で「金属研磨、プラスチック研磨、電解研磨」というグループが形成されていた場合、「金属研磨、プラスチック研磨」と「電解研磨」の 2 つのグループに分割される。

3.3　日本語 WordNet によるグルーピング

3 段階目のグルーピングに使用するスコアとして、日本語 WordNet 内の 2 つの単語 w_i, w_j 間の意味の距離を以下のように定義する。

1. w_i と w_j が同じ概念 $\Rightarrow 0$
2. w_i と w_j が部分関係 $\Rightarrow 0.5$
3. w_i の 1 つ上の上位概念と w_j の概念が同じ $\Rightarrow 1$
4. w_i と w_j それぞれ 1 つ上の上位概念が同じ $\Rightarrow 1.5$
5. 上記以外 $\Rightarrow \infty$

しかしながら、商品用途表現は複数の単語で構成されていることがあるため、構成単語の中から最もその意味を表す単語を抽出したい。一方で、日本語 WordNet に登録されている概念の中には「熱」と「柔軟性」の 2 つの単語が 1 つ上の概念で紐づくような、グルーピングに活用しにくい抽象度の高い概念が存在する。そこで、式 1 で概念 i の抽象度 α を定義し、商品用途表現を構成する単語の中である抽象度以下かつ最も抽象度の高い単語を用いて、日本語 WordNet による商品用途表現間のスコアを求める。

$$\alpha = \frac{1}{|l(i)|} \sum_{j \in l(i)} d_{i,j} \tag{1}$$

なお、$l(i)$ は概念 i が持つ下位概念のうち末端にある概念を、$d_{i,j}$ はツリー上で概念 i から j までに辿る概念数である。2 段階目で作成した各グループ内で上述の商品用途表現間のスコアを求め、最長距離法による階層的クラスタリングを適用し、閾値 T 以下の商品用途表現間を最終的なグループとした。

4　評価

4.1　商品用途表現の抽出

表 2 に、表 1 のラベルを付与した商品用途表現数、図 2 のモデルによる各ラベルの F_1 値、およびラベル付けしていないデータから取得できた商品用途表現数を示す。モデルの訓練および評価には層化グループ付き 5 分割交差検定を用い、各商品用途表現に対して予測された B と I のラベルがアノテーション結果と完全に一致しているかで正誤判定をした。なお、O ラベルの F_1 値は 0.91 であった。

表 2 より、すべてのラベルにおいて F_1 値が 0.8 未

表 2　商品用途表現ラベルのアノテーション数と予測精度

	アノテーション数	F_1	抽出数
使用者	1,092	0.66	8,061
使用環境	3,938	0.78	31,600
使用対象	10,806	0.79	279,702
使用目的	21,800	0.71	97,710
使用タイミング	262	0.51	1,324
使用イベント	5,191	0.60	41,836
特徴	4,304	0.42	22,986
機能	1,207	0.37	7,983

表 3　商品用途表現のグルーピング結果

P1	チタニウム部品研磨, 金属面研磨, アルミ製品研磨
P2	ステンレス鋼研磨, 工具鋼研磨, ステンレス研磨
P3	プラスチック面研磨, プラスチック研磨, 合成樹脂研磨
P4	切刃研磨, 刃研磨, 刃物研磨
P5	難削材研磨, 硬い素材研磨, ステンレス材研磨
T	コンテナ運搬, 計測器運搬, 低温容器運搬

満であり O ラベルの F_1 値が 0.91 であることから、商品用途表現の位置は同定できているが、その種類を特定できていないことがわかる。この問題点の 1 つとして、同じ単語でも文脈によって正解ラベルが変わることが挙げられる。学習データとして使用した約 13,000 種類の商品用途表現のうち、同じ文字列かつ複数のラベルを持つものが約 36%存在した。たとえば、「研磨」という商品用途表現では、サンドペーパーのような商品の場合には「使用目的」のラベルが、研磨液のような商品であれば「使用イベント」のラベルが付与されていた。すなわち、1 つの同じ文字列の商品用途表現内に複数の意味が存在する可能性を考慮しなければならないと言える。

4.2　同義語のグルーピング

表 3 に、3 章で提案した手法で商品用途表現をグルーピングした結果の例を示す。P1 から P5 は「研磨」に関するグループ、T は「運搬」に関するグループである。なお、3.3 節における抽象度 α と階層的クラスタリングの閾値 T はそれぞれ 1.75、1.51 とした。

表 3 より、「研磨」の文字列が含まれる商品用途表現が「金属」や「プラスチック」など使用対象の素材に応じて分割できていることが分かる。各グループの商品用途表現がどのタイミングでグループピングされているかを詳しく分析すると、P1 から P3 は 3.2 節の商品カテゴリによるグルーピングが、P4 は 3.3 節の日本語 Wordnet の概念がそれぞれ貢献していた。

一方、P5 の「ステンレス材研磨」は P2 に所属することが望ましい。この原因は、3.3 節の日本語 Wordnet によるスコア付与時に「材」という抽象度の高い単語が選択されていたためであった。また T1 は、それぞれ異なる意味を持つ 3 つの商品用途表現が同じグループに割り当てられた例である。これは、日本語 Wordnet を参照する単語がそれぞれ「コンテナ」、「器」、「容器」であり、これらが同じ概念に属していることが原因である。改善策として、「計測器」や「低温容器」のような句単位の情報を利用することで、3 つの表現が同じグループになることを防ぐことができると期待される。

5　おわりに

本稿では、商品の使い道などを説明した文などの構造化されていないデータから有用な商品用途表現を抽出して活用するための方法として、商品用途表現を分類する 8 種類のラベルを提案し、そのラベルを用いた商品用途表現の推定とグルーピングについて示した。今後の課題として、提案したグルーピング手法の精度の改善や、今回作成したデータやモデルを実際の E コマースサービスに適用しての検証が求められる。

参考文献

[1] Xusheng Luo, Yonghua Yang, Kenny Qili Zhu, Yu Gong, and Keping Yang. Conceptualize and Infer User Needs in E-Commerce. In **Proceedings of the 28th ACM International Conference on Information and Knowledge Management**, 2019.

[2] Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. Portuguese Named Entity Recognition using BERT-CRF. **CoRR**, 2019.

[3] David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In **Machine Learning Proceedings 1994**, 1994.

[4] Shufan Wang, Laure Thompson, and Mohit Iyyer. Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration. **CoRR**, 2021.

[5] Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. Development of the Japanese WordNet. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation**, 2008.