The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

EC サイトにおける商品タイトルからの商品名抽出

Product Name Extraction from Product Entries on EC Pages

張 培楠
Peinan Zhang

株式会社サイバーエージェント
CyberAgent., Inc.

We propose a task to identify a product name from an EC page title. On EC pages, sellers need to design their posts to increase the visibility of their products in search results. One of the common techniques is including extra information to the title of their product page. However, adding many keywords can result in such a complicated page title that it is hard for buyers to distinguish a product name from the title. Therefore, extracting product names is important, yet has some challenges especially when titles are in Japanese. (1) Most titles do not have standard grammatical structures. (2) Diverse characters, such as Kanjis, Kanas, alphanumerics, and symbols often appear in a single title. These make models hardly handle the boundaries of words and lead to incorrect learning. In this work, we create a corpus and evaluate several conventional approaches for basic analysis. The results show that this task is still challenging; an existing approach for named entity recognition, which performs very well at some open datasets, can only achieve 23.0 of the F1 score with our dataset.

1. はじめに

インターネットの普及に伴い，電子商取引—いわゆる EC (Electronic Commerce) —が一般的になってきている．中でも多いのが提供された EC サイトのプラットフォーム内で各小売店が出店して自らの商品を売り出す形式である．この形式では，商品ページの閲覧が購買に直結するケースが多いため [1, 2]，商品に関連するキーワードを使用して検索される結果への露出を増やすための検索エンジン最適化 (SEO) 対策を各出品者側で施すことが多い．その中でも特によく行われる方法が商品タイトルへの情報付加である (図 1)．このような方法を多くの出品者が特定のルールや基準に従うわけでもなく行っているため，購入者側からは何を売っているのか咄嗟に判別しにくく，さらに EC サイトの運営者としても販売内容を把握するための障害になりうる．そのため雑多な商品タイトルから出品している商品名を判別することには高い需要がある．

商品タイトルから商品名を抽出することは，固有表現抽出のタスクとして考えることができるが，一般的な固有表現抽出タスクと異なる性質がいくつか存在する．ひとつは，固有表現抽出タスクで想定されている多くのデータは正しい文法で成り立っている自然文であるが，この場合商品タイトルに含まれるトークンのほとんどが名詞・名詞句である上に，通常の文にあるような文法性がなく，語順も考慮されていない．したがって固有表現抽出のような系列ラベリング問題では，ラベルと周囲の語の依存関係を仮定したモデルを採用しているが，そのような仮定が商品タイトルにおいては成立しない．ふたつめは，既存の系列ラベリングの手法ではある程度分かち書きが正しくされていることが前提になることが多いが，今回の場合ではブランド名や商品名，型番，サイズ情報，色情報といった数多くの未知語が頻出する上に，漢字かな，アルファベットや記号，スペース区切りなどの使用頻度や使用ケースも出品者に大きく依存していて統一性がない．そのため高精度の単語分割は困難である．これらのことから，本タスクは非常に難易度が高いタスクであると言える．

このような背景を元に，我々はこのタスクを商品名抽出と定義し，本研究ではそれに必要なコーパスの作成および複数のベースラインを本コーパスで評価する．

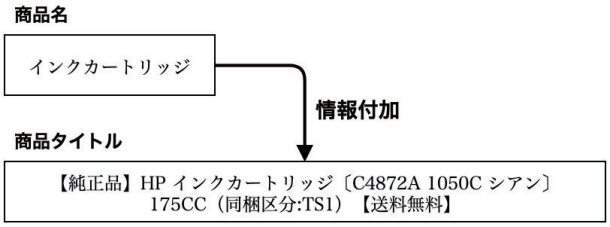

図 1　出品者が出品商品タイトルへ情報付加する例

2. 商品名抽出

2.1 タスクの定義とデータセットの作成

本タスクでは，EC サイトに掲載されているような商品のタイトルから商品そのものを指す商品の名称を抽出することを目標とする．そのためのデータセットを作成するにあたって，まずその特徴を分析する．データには Yahoo!ショッピング [*1] に掲載されている商品を用いた．特定のドメインに依存する言語的な特徴に偏りが生じないよう 18 カテゴリから合計で約 3 億件の商品タイトルを集め，そのうちの約 15,000 件を無作為に選んで分析に使用した．そのうちのいくつかの例を表 1 に列挙する．

例を見て分かるように，いずれの商品タイトルも各単語・フレーズがほぼ名詞・名詞句になっており，「送料 300 円選択可能です」といった部分的に文法的なフレーズは存在するが，タイトル全体で一貫して文法構造は持っていない．各単語・フレーズも商品に関連する製造元やブランド名，商品名，型番，サイズ，色，内容量以外に，商品自体に関連しない「送料無料」といった配送情報や「代引き不可」といった取引情報が散見される．さらには漢字，かな，アルファベット，数字や記号といった表層が特定のルールに則ることなく書かれていることが見て取れる．

*1 https://shopping.yahoo.co.jp

表 1　商品タイトル例

最大 75 % off！ SALE 開催中★ [ロイネス]Roiness 犬用 いわし 150g /ウエット パウチ 国産 4571245858269 #w-148985
〔純正品〕 HP インクカートリッジ/トナーカートリッジ 〔C9363HJ HP134 3 色カラー〕
ＴＲＵＳＣＯ　ナベ頭組込ネジ　クロメートＰ－４　サイズＭ６Ｘ１０　５０本入【メール便 送料 300 円選択可能です。】
（送料無料）（代引き不可）アーノルドパーマー　ベビー サンダル 13.5cm AP4111
当店 1 年保証 カシオ Casio General Men's Watches Metal Fashion MTP-1183A-7ADF - WW

表 2　付加情報のカテゴリとその例

カテゴリ	例
商品名	「インクカートリッジ」
商品の別名	「トナーカートリッジ」
規格情報	「サイズＭ６Ｘ１０」「白」「90g」
製造元名	「ロイネス」「HP」
配送情報	「送料無料」「メール便」
その他情報	「最大 75 % off！」

また，ここで考慮しなければならない重要な事柄のひとつが，正解ラベルの付け方である．出品者は商品に対して数多くの情報を付与するが，それらを分類すると表 2 のように大別できる．その商品自体に関するものとしては，「商品名」「商品の別名」とそれらのサイズ，色などといった情報を補足する「規格情報」，さらに製造元やブランドを示す「製造元名」がある．配送に関しては「配送情報」があり，セール情報などは「その他情報」に属するように分類を行った．

本研究のゴールは商品タイトルから商品名を抽出することであるため，正解ラベルは「商品名」と「商品の別名」とすべきところだが，商品名の定義に情報の粒度からくるゆらぎが生じることが確認されている．例えば「ジャケット HOUSTON MA-1 ALL フライトジャケット ミリタリージャケット メンズ アメカジ 送料無料」というタイトルに対して，「ジャケット」「フライトジャケット」「ミリタリージャケット」「MA-1」のすべてが同じものを指しており，粒度が粗くカテゴリの意味合いが強い「ジャケット」から粒度が細かい型番に近い「MA-1」が存在する．この際にどれを商品名とすべきかは定量的な指標がないため，今回は広告制作者の感覚で選択してもらいながら約 3,800 件事例のコーパスを作成した．その内訳を表 3 に示す．

表 3　コーパスの内訳

商品タイトル数	3,841
商品タイトル延べ文字数	220,900
商品タイトル異なり文字数	1,634
商品タイトル平均文字数	57.5

また，コーパスを作成するに当たって設けた制約を以下に記す．

制約 1. 連続した単語を選択　例えば「[ロイネス]Roiness 犬用 いわし 150g /ウエット パウチ」という商品タイトルに対して，商品名に「犬用 いわし」や「Roiness 犬用 いわし」は使用可能だが，「ロイネス 犬用 いわし」は単語がタイトルで連続していないので不可とした．

制約 2. 複数単語は使用不可　「アーノルドパーマー ベビー サンダル」という商品タイトルの場合，「アーノルドパーマー サンダル」や「ベビー サンダル」ではなく「サンダル」もしくは「アーノルドパーマー」を使用する．

制約 3. より認識される最小単位の単語を選択　これは前述の通り，「ジャケット HOUSTON MA-1 ALL フライトジャケット ミリタリージャケット メンズ アメカジ 送料無料」といった，同じ商品に複数の名称が存在する時，一般的に最もよく用いられる名称を選択する．この場合だと「MA-1」が選択される．

制約 1 と 2 はタスクをよりシンプルにするために設けた制約である．この制約を解除することで，現実的にはより需要のある，しかし研究的により難しいタスクとして派生させることも可能である．また，現時点では制約 3 が最も定性的だが，定量的な指標が利用または発見できていないため，一時的な妥協点として設けている．より定量的な指標を設計することを今後の課題とする．

2.2　手法

2.1 節で作成した商品コーパスを使用して，いくつかの手法を使って商品タイトルから商品名を抽出する実験を行っていく．実験に使用したアプローチは大きく Term Weighting として解く方向と系列ラベリング問題として解く方向のふたつに分けられる．

2.2.1　Term Weighting として解くアプローチ

Term Weighting は，抽出対象のドキュメントに含まれる単語に対して重要度を示すスコアを付与していく伝統的な情報抽出のいちタスクである．スコアの付与には多くの手法が存在するが，中でもよく使用されるのは TF-IDF (Term Frequency Inversed Document Frequency) と呼ばれる手法である．TF-IDF は単語の出現頻度（TF）と単語逆文書頻度（IDF）に分かれており，以下のように定式化される．

$$\begin{aligned}\mathrm{tfidf}(t,d,D) &= \mathrm{tf}(t,d)\cdot\mathrm{idf}(t,D)\\ &= \frac{\mathrm{freq}(t,d)}{\sum_{t_i\in d}\mathrm{freq}(t,d)}\cdot\log\left(\frac{|D|}{1+n_t}\right)\end{aligned}$$

t は各単語を表し，$d \in D$ は各ドキュメント，$\mathrm{freq}(*)$ は頻度を返す関数であり，n_t は単語 t が出現する文書数である．

TF-IDF の根幹にある考え方は，多くのドキュメントで出現する単語は重要でない単語である可能性が高く，反対に特定のドキュメントにしか多く出現しない単語は重要な単語である可能性が高い，というものである．本タスクにおいて各商品タイトルをドキュメントとして見立て，「送料無料」や「代引不可」といった商品以外の情報は特定の商品タイトルに限らず広く出現し，反対に商品名といったその商品自体を表している単語は限定的な商品タイトルにしか存在しない，と仮定することができる．そうすることで TF-IDF を本タスクに適用できると考え，この手法による商品名抽出を試みた．

2.2.2　系列ラベリング問題として解くアプローチ

本タスクに近しい分野として固有表現抽出があり，一般的な固有表現抽出タスクは系列ラベリング問題として解かれることがほとんどである．系列ラベリング問題とは，入力文の単語列 $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ からラベル列 $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$

表 4　人手によって設計された素性

素性分類	説明・例
品詞素性	名詞や接続詞など
表層素性	漢字かな，数字やアルファベットなど
位置素性	タイトルでの単語が存在している位置
辞書素性	形態素解析器の辞書に含まれているかどうか

を予測するタスクで定式化でき，y_i にあたるラベルには固有表現のタグが付与される．このタグ付与の一例として BIO タグ (Begin, Inside, Other) があり，例えば $\mathbf{X}$ = (太郎, は, 明日, 東京, ディズニーランド, に, 行く) という系列に対して $\boldsymbol{y}$ = (B-PER, O, B-DAT, B-LOC, I-LOC, O, O) という正解のラベル列が付けられる．ハイフン後の表記は，人物であれば PER，日付や場所であれば DAT と LOC というように，そのタグが示す固有表現の種類である．今回は BIO タグに商品名を示すタグである PRD タグの 1 種類のみを使用した．つまり $\mathbf{X}$ = (Steve, Madden, メンズ, Taslyn, ロー, ファー, Grey) という系列に対して，$\boldsymbol{y}$ = (O, O, O, O, B-PRD, I-PRD, O) というラベル列を付与する．

系列ラベリング問題として解く際にふたつの手法を使用した．ひとつめは人手によって素性を設計し，その素性をもとに学習する手法．ふたつめはニューラルネットワークによる End-to-End に学習してタグを予測する手法である．

素性設計による手法　今回設計した素性を表 4 に示す．これらの素性を用いて条件付き確率場 (CRF) でラベリングを行う [3]．CRF は識別モデルの一種で，入力系列 $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ に対して系列 $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ を予測系列として出力する．系列の確率の予測は以下のように定式化できる．

$$P(\boldsymbol{y}|\mathbf{X}) = \frac{1}{Z} \exp \sum_{t=1}^{n} \sum_{k=1}^{K} \lambda_k f_k(x, y_{\{t-i,\ldots,t\}}, t)$$

$$Z = \sum_{y \in \boldsymbol{y}} \exp \sum_{t=1}^{n} \sum_{k=1}^{K} \lambda_k f_k(x, y_{\{t-i,\ldots,t\}}, t)$$

k は素性の数で，$f_k(*)$ はウィンドウ幅 i の素性関数，λ_k は素性ベクトルの重みを表している．計算されるラベル列 $\boldsymbol{y}$ の確率 $P(\boldsymbol{y}|\mathbf{X})$ を最大化するための重み λ_k を最急降下法などの最適化手法によって最適化する．

ニューラルネットワークによる手法　Lample ら [4] によって提案された，単語・文字を Bidirectional LSTM (BiLSTM) を用いて得られた内部表現を CRF でラベルを予測する手法（図 2）．具体的には単語列を $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ としたときに，t 番目の単語の分散表現を $\boldsymbol{w}_t$ とする．また単語 x_t は文字列 $\mathbf{C}_t = (c_1, \ldots, c_m)$ から構成されており，内部で使用する t 番目の単語は $\boldsymbol{h}_t$ で表現し，以下のように算出する．なお $\oplus$ はベクトルの連結を表す．

$$\boldsymbol{h}_t = \mathrm{BiLSTM}(\boldsymbol{w}_t \oplus \boldsymbol{c}_t)$$

$$\boldsymbol{c}_t = \mathrm{BiLSTM}(\mathbf{C}_t)$$

これにより得られた $\boldsymbol{h}$ を $\boldsymbol{z} = \mathbf{W}\boldsymbol{h} + \boldsymbol{b}$ のように重み $\mathbf{W}$ でラベル数次元に非線形変換したのち，CRF を用いてラベル列 $\boldsymbol{y}' = (y'_1, \ldots, y'_n)$ を予測値として出力する．
なお、この手法は CoNLL2003 データセット [5] において F 値 90.93 を達成しており，当時の固有表現抽出において最高精度を持つ手法であった．

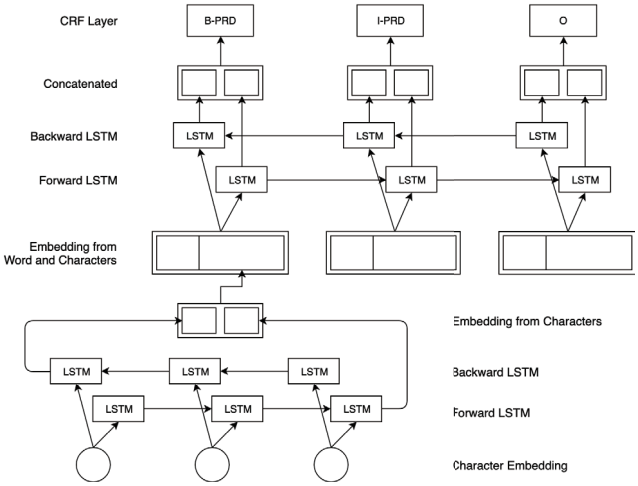

図 2　BiLSTM + CRF の構造図

3. 実験

3.1 実験設定

2.1 節で作成した商品コーパスを用いて評価実験を行っていく．全ての手法において共通する 10 個のシード値でコーパスを (train : valid : test) = (8.5 : 0.5 : 1) の割合で分割し，評価は予測ラベルと正解ラベルの F 値，適合率，再現率で行う．

TF-IDF　TF-IDF の場合は出力が各単語に対するスカラー値となるので，タイトル内の単語で最も高い TF-IDF 値を持つ単語を B タグとし，他を O タグとして評価する．また，TF-IDF は分かち書きされた単語がそのまま固有表現のチャンクになるため，分かち書きに使用する辞書の比較と前処理を施した．

CRF　前後 2 単語のウィンドウ幅での素性を作成し，L1/L2 正則化と準ニュートン法 (L-BFGS) [6] で最適化を行う．L1/L2 の重みはそれぞれ 1.0 と 0.001 で early stopping を考慮した最大 50 回のイテレーションを実行する．

BiLSTM + CRF　最大 100 epoch の early stopping で学習し，単語と文字の分散表現の次元はそれぞれ 100 と 25，Dropout は 0.5 に設定して Adam を用いて最適化を行った．

3.2 実験結果と分析

実験結果を表 6 に示す．結果より，全ての指標において BiLSTM+CRF が他手法を大きく上回った．しかし最も高い F 値でも 23.0 であることから，本タスクの難易度の高さが分かる．

TF-IDF 系列ラベリング問題とは異なるアプローチをしているので厳密には同じ基準での比較はできないが，正解ラベルを必要としない教師なし手法であり，それと比較すると正解ラベルを使用した教師あり手法のほうが F 値において 2〜3 倍以上の性能を達成していることが見て取れる．

人手による素性設計を元に学習する CRF は適合率が 20.3 と再現率に比べて高いことが見て取れる．これはハンドクラフトな素性では正しい商品名を正確に抽出できていないことに起因すると考えられる．品詞，表層，位置以外に考慮できていない特徴が商品名には潜んでおり，意味的な観点でも辞書を一般的な文書コーパスから作成されたものではなく，EC サイトに特化した辞書を使用することで単語の適用範囲が限定され，より

表 5　正答例および誤答例

1 正答例	Steve Madden (スティーブマッデン) メンズ Taslyn ロー ファー Grey
2 正答例	防 ダニ ・ 抗菌 ・ 防臭 フランネルラグマット / 絨毯 〔 ボリューム タイプ / 約 185 × 280 cm ピンク 〕
3 両方	(業務 用 10 個 セット) 三甲 (サンコー) ベタ 目 コンテナ ボックス / サンボックス 11 - 2 (代 引 不可)
4 誤答例	メンズ ブーツ チャッカブーツ カジュアル Timberland Men ' s Groveton Leather Fabric Boot 正規 輸入 品
5 誤答例	日本 製 【 ■ nano cafe 】 ベビー 手 マグ54294 4522202542943
6 誤答例	HiKOKI (旧 日立工機) フロア 用 タッカ 電源 電圧 V N 5004 MF (送料 無料)(代 引き 発送 不可)
7 誤答例	ジョインテックス 応接 センター テーブル KE - 1260 W

表 6　商品コーパスに対する各手法の実験結果

手法	適合率	再現率	F 値
TF-IDF	5.7	11.1	7.0
CRF	20.3	11.4	14.5
BiLSTM + CRF	**25.4**	**21.2**	**23.0**

効果を発揮する素性になりうる．また，ただ辞書に存在するかどうかという 0/1 のバイナリ値だけでなく，単語が属するドメイン名そのものを利用することでより情報に富んだ素性として使える可能性がある．

ニューラルネットワークによる手法では適合率と再現率のバランスが他手法と比べて取れており，F 値も他より大きく上回る．それでも全体的にスコアが低いのはデータに依るものと考察できる．まず今回使用したデータのサイズは大きくなく，性能を発揮するにはより大きなコーパスが必要となる．また，データの性質上，ひとつの商品タイトルにはひとつの商品名しか正解ラベルが付与されないため，たとえ類似した単語 — 2.1 節の分類で言うところの「商品の別名」など — があったとしてもそれにはネガティブなラベルが付与されている．したがって，意味的に類似している単語の表現を捉える傾向があるニューラルネットワーク手法ではそれに引っ張られ，商品名を当てることが難しくなっている可能性が考えられる．

より詳細な分析をするため，BiLSTM+CRF の出力と正解データを照らし合わせて正答例と誤答例を表 5 に示した．緑ハイライトが正解箇所，赤ハイライトが予測箇所，青ハイライトが正解と予測が被っている箇所になる．(1)(2) はふたつとも正答できている例で，(2) のような長いチャンクでも正しくラベリングできていることが見て取れる．(3) はシステムが複数の商品名を認識できた例で，意味表現の近い「コンテナボックス」と「サンボックス」が識別されており人間が見ても正しいが，制約上正解ラベルはひとつしか存在しないため正解と不正解両方の判断になっている．(4) は (3) に類似しているが，意味的に近い「メンズ ブーツ」を認識したが，正解ラベルである「チャッカブーツ」を認識できなかった誤答例である．(5) はブランド名を商品名と誤認識した例であり，(6) では「フロア用タッカ」を予測しており，それは人間が見ても正しく認識できているが，教師データ作成時のゆらぎで正解ラベルが型番に近い「N5004MF」になっているため不正解となっている．(7) はシステムの出力が「センターテーブル KW-1260W」となっているが，正解ラベルは「応接センターテーブル」と範囲を誤っている．

これらのことから，実験結果においての F 値の低さは全てがシステムの性能が低いことに起因するだけではなく，人間が見ても理解できる予測をしているがデータの基準や制約条件に依るものであるものも多いと考えられる．したがって，性能向上のためには手法の改善に伴ってより基準に一貫性がある直感的なデータ作成が今後必要になる．

4.　まとめと今後の課題

本研究では EC サイトにおける商品タイトルからの商品名抽出というタスクを提案し，そのためのコーパスの作成およびそれを使った実験・分析を行った．実験に使用した手法は Term Weighting としてのアプローチである TF-IDF と，系列ラベリング問題として解くアプローチである CRF があり，CRF を用いる場合は素性設計による手法と End-to-End なニューラルネットワークによる手法で比較実験を行った．その結果，ラベル列を F 値で評価した場合はニューラルネットワーク手法である BiLSTM+CRF が最も良い性能である 23.0 を達成した．しかしこの数字は CoNLL 2003 データセットと比較して大きく下回ることから，本タスクの難易度の高さを裏付けることができた．

これからの展望として，データに関しては商品名を定義づけるより定量的な指標を設計することやタグの種類の拡張や複数単語対応すること，手法に関しては，ノイジーな入力に頑健な単語分割に依存しないような文字ベース・サブワードベースの固有表現抽出手法を使用することなどが挙げられる．また，今回の方向性以外のアプローチも存在し得るので探っていきたい．

参考文献

[1] Eric Enge, Stephan Spencer, Jessie Stricchiola, and Rand Fishkin. *The art of SEO*. O'Reilly Media, Inc., 2012.

[2] StephenO' Neill, Kevin Curran. The core aspects of search engine optimisation necessary to move up the ranking. *International Journal of Ambient Computing and Intelligence (IJACI)*, Vol. 3, No. 4, pp. 62–70, 2011.

[3] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 282–289, 2001.

[4] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pp. 260–270, 2016.

[5] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of NAACL-HLT*, pp. 142–147, 2003.

[6] Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, Vol. 35, No. 151, pp. 773–782, 1980.