商品タイトルから商品名を自動抽出するための効率的な教師データ作成手法

佐々木 稔　　新納 浩幸
茨城大学 工学部 情報工学科
{msasaki, shinnou}@mx.ibaraki.ac.jp

1　はじめに

大手のインターネットショッピングモールでは、それぞれのモールが膨大な数の商品を扱っている。そのため、利用者が目的の商品を素早く探し出せるように、モールの運営者はなんらかの情報を利用して商品ページの整理を行っている。このように整理された膨大な商品情報は検索システムの中で保存され、キーワードを指定することで素早く取り出される。他にも商品情報を検索する方法は存在するが、利用者の多くはこのキーワード検索を利用して商品情報を探すことが多い。

商品情報を素早く見つける環境の整備は進んでいるが、検索結果や商品説明の表示において不便な点も多い。そのため、利用者に分かりやすい検索結果を表示するための課題がいくつか存在する。そのひとつとして、商品ページのタイトル中に、商品名や型番という情報だけではなく、「メーカー在庫限り」や「送料無料」などといった付加的な情報を含んでいることが挙げられる。この場合、検索結果リストの商品タイトルには付加的な情報だけが表示され、商品名は省略されてしまう事例もある。早く商品を見つけるなど目的を明確に持つ利用者にとっては、探している商品の名称以外にこのような付加的な情報の多いことが負荷を感じる要因となりやすい。このような現状において、利用者がより理解しやすい商品情報を提供することが必要だと考えられる。検索結果の中に商品タイトルを掲載するだけではなく商品名も表示することができれば、利用者は多くの商品情報について、本文を確認することなく知りたい商品を見つけることが可能となる。

そこで、本稿ではショッピングモールの商品タイトルから商品名を抽出する課題を設定し、商品名を自動的に抽出にするための方法について検討を行う。この課題では商品タイトルに含まれる商品名の設定が重要なポイントとなる。一般的に商品名は商品タイトルの中にある製品名、材料名、食品名などの製品名 [5] となるが、その直前にブランド名が存在することも多い。ここでは、ブランド名も含めた製品名を抽出するべき商品名と定義して、その自動抽出を考える。

この課題を考える場合、最も効果的な解法は与えられた単語列に対して商品名とそれ以外の部分にそれぞれ対応するラベルを付与する系列ラベリング問題として定式化することであると考えられる。その場合、あらかじめラベルが付与された学習データを用意することで、そのラベルを付与するためのモデルを学習することが可能である。しかし、これまでに存在する膨大な商品集合の一部に対し、その商品名にラベルを付けて学習データを作成することは時間のかかる大変な作業となる。そこで、商品集合の中に型番だけが異なるなど類似した商品名が存在することに注目し、類似した商品名の系列については1度のラベル付けで済むことができるようにフィルタリングを行い、効率的に教師データを作成する手法を提案する。提案手法に対して、ラベル付け作業を行う時間と得られた教師データによる自動商品名抽出の精度について分析を行う。

2　関連研究

文書から商品名を抽出する簡単な方法は商品タイトル内に存在する商品名の抽出規則を生成し、それを商品タイトルに適用して抽出することだと考えられる [4]。名詞が連続したフレーズを商品名とした場合 [1]、「甘辛」や「めちゃうま」などが含まれる商品名を抽出することができず、また、「シーズン特価」といった付加的な情報も抽出されてしまう。そのため、商品名を抽出するためには数多くの複雑な抽出規則が必要となるため、機械学習手法を利用することが一般的である。

機械学習手法による商品名抽出研究はこれまでにいくつか存在する。中国語の Web ページを対象として製品名を抽出するもの [2] や、音響機器のレビューを

※ ハツモール薬用スカルプシャンプー300ML1800 4975446394180 【YDKG-k】【W3】

図 1: 商品タイトルの例、下線部分が抽出する商品名 (楽天市場商品データより引用)

フォトフレーム　クローバー　L 判★総額 5250 円以上で送料無料★激安祭
フォトフレームクローバキャビネ版★総額 5250 円以上で送料無料★激安祭
【送料無料】フォトフレーム　クローバー L 判 W ★総額 5250 円以上で送料無料★激安祭【smtb-s】

図 2: 商品タイトルの例、下線部分が抽出する商品名 (楽天市場商品データより引用)

書いた Web ページから製品名を抽出するものがある [3]。これらの研究はデータ集合全体に対して、ランダムに選んだ文書を学習データとして利用している。

また、系列ラベリングにおける学習データの効率的な選択手法についてもいくつか研究が行われている。商品についてのブログ記事に対し、ブログに登録された商品カテゴリをラベルとして記事中の商品名に付与する方法がある [8]。商品カテゴリ情報を利用してブログ検索を行い、教師データとして使用する記事をフィルタリングする。商品カテゴリを利用することで、商品名の知識がない場合でも教師データを作成することが可能となる。教師データとしてラベル付けするデータを効果的に収集する手法も提案されている [6]。この手法は、はじめに初期的な教師データとしていくつかの文に対してラベル付けを行う。次に CRF による自動的なラベル付けの結果から教師データに追加するデータを選択する。そのデータ集合に対して、ラベル付けを行って教師データに追加、再度 CRF により追加すべきデータ集合の選択を繰り返す。さらに、外部知識を利用して効果的に教師データを作成する手法も提案されている [7]。ラベル付けの対象となる固有表現についての参照情報を利用することで正解率を向上させる学習モデルを構築することができる。

3　課題設定

本研究では、商品ページのタイトルから商品名を抽出する課題を考える。この課題では、製品名だけでは具体的に商品を特定することが難しいことから、その直前にあるブランド名も合わせて抽出対象と設定する。例えば、図 1 にある商品タイトルでは、商品名として「ハツモール薬用スカルプシャンプー」が商品名であると設定するが、「300ML」といった内容量についての具体的な値や、「1800」、「4975446394180」などの型番を表したフレーズは、ユーザにとって意味が分かりにくいため抽出の対象外として扱う。

4　提案手法

4.1　商品タイトル集合の特徴

商品タイトル集合には、図 2 に示すように同じ商品名でも型番の異なるものや、「送料無料」などの文字列が追加されているものが数多く存在する。このような色や形状、対象機種などの違いによって商品タイトルが微妙に異なる商品データは、ショッピングモール内において数多く存在する。また、各ショップ毎に商品情報の記述が似ているため、商品名を含めた商品情報の系列も類似するものが多く存在する。商品の総数は膨大に存在しているため、大量の商品情報に対して記述されたデータを整理するのは非常に困難ではある。しかし、このような類似性を持つ特徴を利用することにより、整理すべき情報を少しでも少なくすることが重要な課題だと考えられる。

4.2　教師データ作成手法

そこで、本稿では類似した商品タイトルをフィルタリングし、教師データ作成のためのラベル付け作業の負荷を軽減する手法を提案する。商品タイトル集合から教師データを作成する場合、なるべく商品タイトルは 1 種類に対して 1 回のラベル付けをすれば済むようにフィルタリングを行いたい。大規模なデータに対する手作業によるラベル付け作業は非常に面倒であるため、フィルタリングにより教師データの作成時間を軽

減することが目的となる。提案手法は教師データ中の類似した商品名の系列によるデータの偏りを少なくする効果があり、より効果的なラベル付けを行うモデルを構築することができる。

4.3 商品タイトルのフィルタリング

商品タイトル集合から類似したタイトルをフィルタリングする場合、処理時間はできるだけ短く抑える必要がある。商品タイトル集合に対して厳密な類似性判定を行うと、そのデータ数が増えるごとに比較する回数が増加する。そのため、フィルタリング処理と処理後のラベル付け時間の合計が、元のタイトル集合へのラベル付け時間よりも長くなる場合もある。教師データ作成時間を短縮するために、類似した商品タイトルが連続して出現するタイトル集合の特徴を考慮し、対象の商品タイトルを直前に残った商品タイトルと類似性を比較し、フィルタリングの判定を行う。

フィルタリング処理はすべての商品タイトルに対して以下の手順により行われる。

1. 最初のタイトル文は残し、比較元タイトルと設定
2. 比較元タイトルを形態素解析し、全単語の頻度を要素とするベクトルを作成
3. 次のタイトル文を判定対象として設定し、形態素解析し、同様にベクトル化
4. 比較元タイトルと判定対象タイトルの Dice 係数を計算
5. 閾値以上の場合はそのタイトルをフィルタリングし、処理 3 に戻る
6. 閾値以下の場合はそのタイトルを残し、それを比較元タイトルと設定し、処理 3 に戻る

ここで、Dice 係数の閾値は 0.7 と設定した。これは、系列の構造は類似するものの、型番や短い製品名などが異なる商品タイトルをフィルタリングすることが可能な数値として設定した。

4.4 商品タイトルのラベル付け

教師データを作成するために、商品タイトルに含まれる商品名の系列にラベルを付与する。ラベルは系列ラベリングで一般的に使用される IOB タグを使う。これは、商品名が始まる単語にラベル ‘B’、商品名の途中にはラベル ‘I’ を付与し、それ以外の部分はラベル ‘O’ を付与する。フィルタリングをした商品タイトル集合に対し、IOB タグでラベル付けを手作業で行い、それを教師データとして利用する。

4.5 商品名の自動ラベル付け

得られた教師データに対して、新しい商品タイトルに商品名を自動的にラベル付けするためのモデルを作成する。本稿では、系列ラベリングの一般的な手法である Conditional Random Field (CRF) を利用する[1]。このとき、学習する際に扱う素性は前後 2 単語の表記と品詞と設定する。

教師データから作成したモデルを使い、新しい商品タイトルについて最適な系列を計算により求める。このとき、商品タイトルには必ず商品名が含まれると定め、条件付き確率が最大で、必ずラベル ‘B’ が存在する系列を最適な系列として出力する。

5 実験

提案手法の効果を評価するために、実際の商品タイトル集合に対してラベル付け実験を行う。本節では、使用したデータと実験結果を説明する。

5.1 データ

実験で使用するデータは、楽天技術研究所が公開した「楽天市場の全商品データ」の商品名を利用する[2]。この商品データには約 5000 万件の商品情報が存在し、各商品情報は店舗コード、商品 ID、商品名、商品説明文、商品 URL、商品画像、商品価格、および、ジャンル ID の属性から成り立っている。ひとつの商品に対して、属性値が割り当てられることで、その商品を特定する情報が整理されている。

実験ではこのデータを 1,000 件ずつ分割し、そのうちの 10 個をそれぞれ教師データとして利用する。各データに対して、すべての商品タイトルを抽出し、形態素解析を行うことによって単語と品詞の組の系列を求める。また、提案手法により商品タイトルをフィルタリングしたものについても同様に単語と品詞の組の系列を用意する。各データ集合には偏りがあるため、10 回の実験結果を平均することでより信頼性の高い結果が得られると考えられる。また、テストデータは教師データとして利用したデータ集合以外から商品タイトル 100 件分を偏りのないように選択し、教師データと同様に形態素解析を行い、正解となるラベル付けを行って準備した。

1,000 件の教師データでモデルを学習し、テストデータを自動ラベル付けした精度をベースラインとし、フィ

[1] http://crfpp.sourceforge.net/
[2] http://rit.rakuten.co.jp/rdr/index.html

表 1: 元データの正解数とフィルタリング後の文書数、および正解数

データ	元データの正解数	選別後のタイトル数	選別後の正解数
1	21	261	23
2	39	388	48
3	23	163	38
4	20	321	21
5	24	202	22
6	22	331	24
7	23	514	23
8	22	460	22
9	23	210	25
10	18	183	20
平均	23.5	303.3	26.6

表 2: 元データとフィルタリング後のデータにおけるラベル付け時間

データ	元データの作業時間 (分)	選別データの作業時間 (分)
1	150	30
2	153	64
3	109	18
4	146	38
5	115	30
6	76	30
7	91	60
8	80	37
9	63	23
10	60	20
平均	104.3	35

ルタリング後のデータでモデルを学習し、同様に自動ラベル付けした精度と比較を行う。

5.2 実験結果

表 1 にデータ集合の偏りを考慮して 10 個のデータ集合に対して行った実験結果を示す。表の各列には、元データを教師データとしてテストデータをラベル付けした時の正解数、提案手法でフィルタリングを行った際に残るタイトル数、および、それを教師データとしてテストデータをラベル付けした正解数を示している。

この表より、フィルタリングした商品タイトル集合を教師データとして利用することで、一部のデータでは正解数が下がっているが、ほとんどのデータで正解数が上昇する結果が得られた。これにより提案手法がラベル付けの精度に対して有効であることが分かる。商品タイトルのデータ集合の特徴として、類似した系列を持つ商品タイトルが数多く存在するため、データが偏りやすいことがある。そのため、フィルタリングを行うことで教師データにおける商品タイトルの系列に偏りがなくなり、精度が向上したと考えられる。

表 2 は、大学生が元のデータとフィルタリング後のデータにラベルを手作業で付与した時の時間を表す。ラベル付け作業への慣れもあり、判断が早くなったことから徐々に元データの作業時間が短くなっているが、フィルタリングを行うことで作業時間は平均して 1/3 に軽減している。教師データを作成するための負荷が軽減され、自動的なラベル付けの精度も向上することから、提案手法は教師データ作成の作業効率に対して有効であると考えられる。

6 おわりに

本稿ではショッピングモールの商品タイトルから商品名を抽出する課題を設定し、系列ラベリングモデルを作成するための教師データの効率的な作成手法を提案した。本手法の有効性を評価するために、実際の商品タイトル集合に対してラベル付け実験を行った結果、ラベル付けの精度が向上し、教師データを作成するための負荷も軽減できることが分かった。このことから、提案手法は商品名を抽出する課題に対して有効であると考えられる。

今後の課題としては、効果的なフィルタリング処理に向けた改善が挙げられる。本稿では類似した商品タイトルが連続した場合に類似性の比較を行う簡単な処理でフィルタリングした。より高速で効果的なフィルタリング処理を考えることが重要である。

参考文献

[1] Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 21–26, 2010.

[2] Feifan Liu, Jun Zhao, Bibo Lv, Bo Xu, and Hao Yu. Product named entity recognition based on hierarchical hidden markov model. In *Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, IJCNLP2005, pp. 40–47, 2005.

[3] John M. Pierre. Mining knowledge from text collections using automatically generated metadata. In *Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management*, PAKM'02, 2002.

[4] 林華, 佐々木稔, 新納裕幸. 商品説明文から商品名の直接的な説明、言い換え表現の自動抽出. 第 3 回楽天研究開発シンポジウム, 2010.

[5] 関根聡, 竹内康介. 拡張固有表現オントロジー. 言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」, pp. 23–26, 2007.

[6] Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. Accelerating the annotation of sparse named entities by dynamic sentence selection. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pp. 30–37, 2008.

[7] Yu Usami, Han-Cheol Cho, Naoaki Okazaki, and Jun'ichi Tsujii. Automatic acquisition of huge training data for biomedical named entity recognition. In *Proceedings of BioNLP 2011 Workshop*, BioNLP'11, pp. 65–73, 2011.

[8] 渡辺尚吾, 乾孝司, 山本幹雄. 商品カテゴリ情報に着目した教師データ収集による商品名抽出手法. 第 25 回人工知能学会全国大会, 2F3-1, 2011.