

# オントロジー拡充のための 固有表現の共起情報を用いた語彙獲得

稲葉 真純 長野 伸一 川村 隆浩 服部 正典

Masumi INABA, Shinichi NAGANO, Takahiro KAWAMURA, and Masanori HATTORI

株式会社 東芝  
Toshiba Corporation

**Abstract:** オントロジー活用アプリケーションの運用において、継続的なオントロジーの拡充は重要な課題である。我々は、商品情報に関するオントロジーを活用した評判分析技術の研究開発を行っている。本稿では、大量の商品データから未知の商品や新商品の名称を抽出し、オントロジーの語彙獲得を効率化する手法を提案する。DVDの商品情報を用いた評価実験では、適合率 87.1%、再現率 87.4%の精度で商品名を抽出できた。

## 1. はじめに

近年、Weblog や Wiki, SNS (Social Networking Service) などを利用して、誰もが簡単に Web コンテンツを作成できるようになった。消費者によって発信される情報は CGM (Consumer Generated Media) と呼ばれ、商品やサービスに関するクチコミも多い。クチコミは消費者の購買行動に大きな影響を与えており、マーケティングの面からも注目されている。そこで、我々は、CGM からクチコミを収集し、商品の評判情報を分析するサービス「ユビ de コミミハサnder」を開発してきた [1]。

ユビ de コミミハサnder の特徴は、商品に関する知識体系として、商品オントロジーを用いることにある。商品オントロジーを用いることで、クチコミで言及されている商品の特定や、商品ジャンルの判定が可能になる。しかし、情報伝達速度が速い CGM で言及される新商品を分析するには、定期的にオントロジーを更新する必要がある [2]。オントロジーを用いたアプリケーションの構築運用において、語彙獲得の効率化は、十分な情報量を確保するために不可欠な課題である。

本稿では、商品オントロジーの更新コストを低減するため、大量の商品データから効率的に商品名を抽出し、商品オントロジーの既存のクラスに属するインスタンスを拡充する手法を提案する。2 章では、オントロジーを活用した評判分析技術について述べる。3 章では、固有表現の共起情報を用いた語彙獲得の手法を提案する。次に 4 章で、提案手法の評価実験について報告する。最後に、5 章で今後の課題について述べてまとめとする。

## 2. 評判分析技術

ユビ de コミミハサnder は、CGM から商品に関するクチコミを抽出し、要約を提示するサービスである。ユビ de コミミハサnder の構成を図 1 に示す。例えば、ユーザがある本のタイトルを入力すると、著者や出版社などの情報を取得し、本の感想を述べたブログを収集する。ポジティブ・ネガティブ判定機能では、収集したブログの内容を解析し、評判情報を集計する。関連トピック抽出機能では、話題の関連商品を提示する。また、ソート&フィルタリング機能では、有用なブログを選出する。これらの機能により、多数のクチコミ情報を誰でも簡単に調べることができる。要約の抽出には、商品の感想に関する表現をまとめた感性表現オントロジー [3] と、個別の商品について製造メーカーや出演者などの詳細

---

連絡先：稲葉 真純, masumi.inaba@toshiba.co.jp  
〒212-8582 川崎市幸区小向東芝町 1  
株式会社 東芝 研究開発センター  
知識メディアラボラトリー

情報を記した商品メタデータ、および商品情報をまとめた商品オントロジーを利用する。

商品オントロジーを図 2 に示す。商品オントロジーは、商品ジャンルで分類したクラス階層と、商品名を表す膨大なインスタンスを関連付けたライトウェイト・オントロジーである。

商品オントロジーの拡充には、商品ジャンルを表すクラスや、個別の商品名を表すインスタンス、商品の特徴を表す属性のノード数を増やすアプローチや、それらのノード間の関係を蜜にするアプローチが考えられる。このうち、提案手法では、指定のクラスに属するインスタンスのノード数を増やすために、未知の商品や新商品の名称を獲得することをターゲットとする。

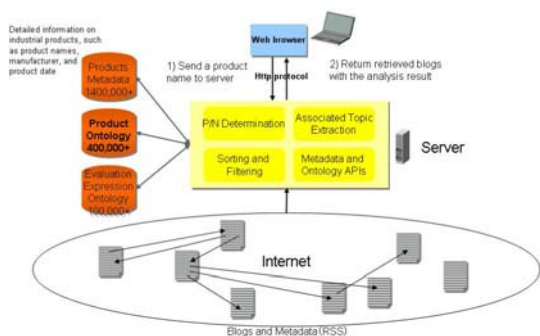


図 1 ユビ de コミミハサンダーの構成

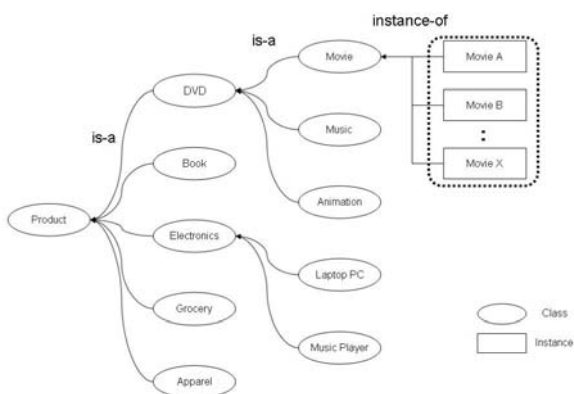


図 2 商品オントロジー

### 3. 提案手法

新製品の発売サイクルに追従するため、定期的に Web から収集する大量の商品名をコーパスにする。

商品情報コーパスの一部を図 3 に示す。コーパスには、以下の特徴がある。

- 商品名は単数または複数の固有表現で構成される
- 固有表現の並び順には規則性がある
- コーパス中で繰り返し用いられる固有表現がある

そこで、固有表現の共起情報を用いたインスタンス拡充システムを開発した。システム構成を図 4 に示す。システムは、商品名の集合を入力とし、インスタンスを出力する。システム内部は、初期化部、形態素解析部、意味タグ付与部、インスタンス生成部の 4 つの処理で構成される。意味タグとは、固有表現の意味を分類したものである。意味タグ付与には、固有表現と意味タグのマッチング履歴と、意味タグの並び順を共起情報として記録したデータベースを用いる。

まず、ステップ 1 の初期化部で、コーパスの全角半角を正規化する。次に、ステップ 2 で、正規化されたコーパスを形態素解析する。次に、ステップ 3 で、コーパスに意味タグを付与する。意味タグは、IREX [4] や拡張固有表現階層 [5] の固有表現分類を元に定義する。最後に、ステップ 4 で、商品名のインスタンスを生成する。

次節以降では、ステップ 3 の意味タグ付与部について説明する。

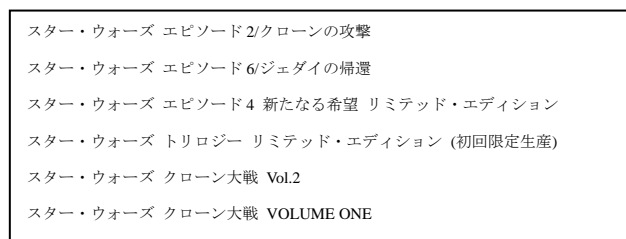


図 3 商品情報コーパスの一部

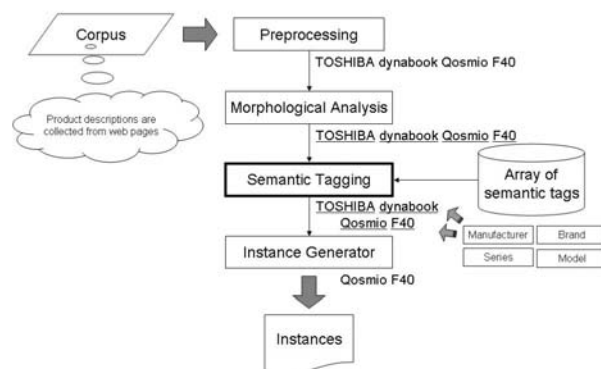


図 4 インスタンス拡充システム

### 3.1 未知語の獲得

新しいコーパスには未知語が多く含まれる。これらの未知語に意味タグを付与するために、意味タグの並び順を利用する。未知語の獲得手順を図 5 に示す。まず、コーパス中の既知の語彙に意味タグを付与する。未知語には仮に「不明」の意味タグを付与し、意味タグの並びを完成させる。次に、意味タグの並びの履歴を参照し、意味タグの並び数が同数の意味タグの並びを抽出する。既知の意味タグについて、出現位置のマッチングを行う。出現位置が一致する意味タグの並びのうち、出現頻度が最も高い意味タグの並びから順に候補とする。続いて、「不明」の意味タグを仮付与した語彙と、既知語の部分マッチングを行う。最も文字列が一致する順に、既知語に与えられた意味タグを未知語の意味タグ候補とする。ここで、前述の出現頻度が高い意味タグの並びの候補との一致を判定することで、「不明」の意味タグが解決し、未知語を獲得することができる。

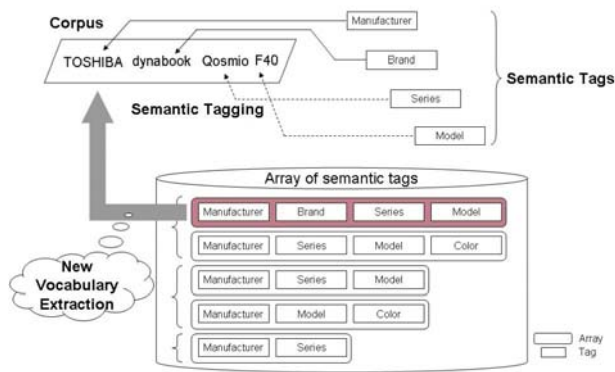


図 5 未知語の獲得手順

### 3.2 意味タグの並びの獲得

意味タグの並びの獲得手順を図 6 に示す。はじめに、コーパス中の既知の語彙に意味タグを付与する。全ての語彙に意味タグが付与され、且つ意味タグの並びの履歴に同じ構成の意味タグの並びがない場合は、新たな意味タグの並びとして抽出し、出現頻度が最も低い意味タグの並びとして履歴に記録する。また、コーパス中の語彙に意味タグを付与し、未知語を含む意味タグの並びを完成させたとき、履歴に同じ意味タグの並び数がない場合は、未知語と既知語の部分マッチングを行う。最も文字列が一致する既知語に与えられた意味タグを、未知語の意味タグとする。これにより、新たな意味タグの並びを獲得できる。

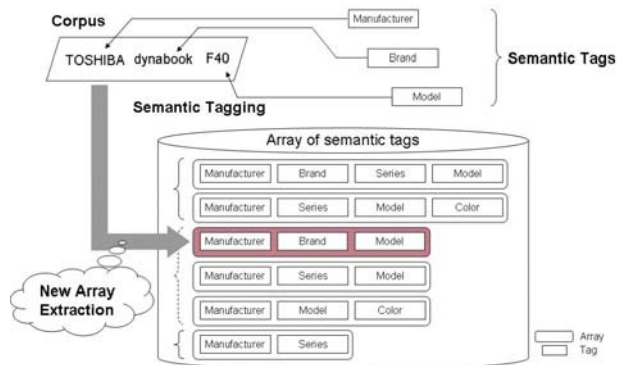


図 6 意味タグの並びの獲得手順

## 4. 評価実験

### 4.1 実験概要

提案手法の有効性を確認するため、評価実験を行った。本実験では、DVD オントロジーのインスタンス（商品名）の作成に関して、人手で作成した正解データと、提案手法により生成した実験データを比較し、抽出精度を算出した。

コーパスとして、DVD 発売情報のポータルサイトから、邦画、洋画、音楽、アニメ、TV ドラマの 5 分野 1 万件の商品名を人手で収集した。これにより、実験の前提条件として、コーパスは DVD 以外のノイズ情報を取り除いたものとみなすことができる。

実験データの作成方法を説明する。まず、人名、商品ジャンル、タイトル、サブタイトル、ボリューム、エディションの 6 つの意味タグを定義する。次に、提案手法を用いて固有表現を抽出し、<タイトル>または<タイトルとサブタイトル>の組み合わせを商品名と定義して商品名を生成した。

作成される商品名の正解数については、作品のセット販売などを考慮し、1 つの元データから、複数の正解が作成することがある。

### 4.2 結果と考察

コーパスから抽出した意味タグの並び順の一部を表 1 に示す。実験の結果、タイトルのみで構成される商品名はコーパス全体に対して少なく、商品名を抽出する手法の必要性が確かめられた。また、並び数が増加するほど、並び順のバリエーションも増加することがわかった。意味タグの共起情報については、機械学習を適用できると考えられる。

表 1 意味タグの並び順の一部

並び数	意味タグの並び順
1	タイトル
2	タイトル ポリウム
	タイトル サブタイトル
	タイトル エディション
	商品ジャンル タイトル
3	タイトル サブタイトル エディション
	タイトル サブタイトル エディション
	タイトル ポリウム エディション
	ジャンル タイトル サブタイトル
	人名 タイトル サブタイトル
	商品ジャンル タイトル サブタイトル

次に、生成した商品名の精度を算出した。適合率 (Precision) の算出方法を式(1)に、再現率 (Recall) の算出方法を式(2)に示す。コーパスに対する精度の算出結果を表 2 に示す。適合率は 87.1%であった。精度が低くなった原因には、複数の語彙で構成される複合語が抽出できなかったことが挙げられる。しかし、生成した商品名の出力結果については、正解データと完全一致はしないまでも、部分的には有効であり、ユーザへの推薦データとしては十分な結果が得られたと考えられる。一方、再現率は 87.4%であった。精度低下の要因としては、1 つのデータから複数の正解を生成すべき場合に、生成漏れがあったことが挙げられる。今後は、複数の商品名を判別するためのパターンを整備する必要がある。

$$Precision = \frac{correct}{all_{method}} \quad (1)$$

$$Recall = \frac{correct}{all_{manual}} \quad (2)$$

表 2 適合率と再現率

	適合率 (%)	再現率 (%)
提案手法	87.1	87.4

タイトルとサブタイトル以外で、コーパスに含まれる意味タグの割合を表 3 に示す。DVD のコーパスには、エディションやポリウムの意味タグに分類できる固有表現が多く含まれることがわかった。これらの意味タグは、分野に関わらず DVD 全般に広く分布していた。この結果から、出現頻度が高い意味タグに分類される固有表現の抽出精度を高めることで、インスタンスの抽出精度を高めることがで

きると考えられる。一方、意味タグの並び順については、分野ごとに特徴があり、異なる分野への転用は困難であると考えられる。しかし、同ジャンルのインスタンスを拡充する用途においては、共起情報を用いる提案手法の有効性が確かめられた。この結果から、一度システムでインスタンスを生成した分野については、定期的なオントロジーの更新コストを低減できると考えられる。

表 3 意味タグが含まれる割合

意味タグ	割合 (%)
エディション	51.8
ポリウム	30.9
人名	10.1
商品ジャンル	7.2

## 5. まとめ

Web リソースを用いることによって、膨大かつ新鮮なデータを入手することは魅力的である。しかし、取得した語彙の正確性を自動的に判定することは困難である。最終的な判定は人間が行うことを前提としているが、判定規則を学習することで抽出精度を高めることもできる。今後は、共起情報を機械学習に適用し、判定処理を効果的に行うことも検討したい。また、構築したオントロジーを元に、CGM などの口語表現が含まれるデータから、商品名の略語や愛称などの同義語を抽出する技術にも取り組んでいきたい。

## 参考文献

- [1] T. Kawamura, S. Nagano, M. Inaba, Y. Mizoguchi: Mobile Service for Reputation Extraction from Weblogs - Public Experiment and Evaluation, Proceedings of Twenty-Second Conference on Artificial Intelligence (AAAI-07), 2007.
- [2] 鈴木健之, 丸山広, 中村太一: 競合製品や競合するサービスに関するオントロジーの構築とその利用, 第 16 回セマンティックウェブとオントロジー研究会, 2007.
- [3] 飯田貴之, 長野伸一, 山崎智弘, 服部正典, 川村隆浩: 評判分析のための感性表現オントロジーメンテナンスツールの開発, 第 18 回セマンティックウェブとオントロジー研究会, 2008.
- [4] IREX: Information Retrieval and Extraction Exercise, <http://nlp.cs.nyu.edu/irex/>.
- [5] S. Sekine, K. Sudo, C. Nobata: Extended Named Entity Hierarchy, 2002.