違法・有害情報検出のための 固有表現辞書の生成についての研究

Study on Generation of Named Entity Dictionary for Identifying Illegal and Harmful Websites

前原 洸貴† 関 哲朗† 池辺 正典†
Kouki Maehara† Tetsuro Seki† Msanori Ikebe†

†文教大学大学院 情報学研究科
†Graduate School of Information and Communications, Bunkyo University

要旨

本研究の目的は，違法・有害情報を検出することである．本稿では，違法・有害情報の一部である薬物関連情報の検出手法を提案する．提案手法では，薬物関連の商品を取り扱う EC サイトの各商品説明ページの HTML ソースコードの差分から薬物関連の EC サイトにのみ現れる特徴や商品属性を観察した．それらを元に抽出ルールを構築し，商品情報を固有表現として抽出する．抽出ルールについての考察を報告する．

1. はじめに

近年のスマートフォン等の携帯情報端末の普及は，インターネット上の児童ポルノや規制薬物等の違法情報，出会い系サイト等の有害情報に接することを容易にし，結果として犯罪に巻き込まれることが増加している．警察庁より業務委託された一般財団法人インターネット協会が運用を行うインターネット・ホットラインセンター(以下「IHC」と略す)では，違法・有害情報該当性の判断ができるものについて，警察へ通報したりプロバイダや電子掲示板の管理者等に対して適切な措置を依頼したりしている．IHC では，公序良俗に反する情報を有害情報にする一方，違法情報を，「わいせつ関連情報」，「薬物関連情報」，「振り込め詐欺等関連情報」，「不正アクセス関連情報」に分類している［1］．

本研究では，指定薬物を取り扱う EC サイトを抽出するための固有表現辞書を生成する手法を提案する．本研究により作成された固有表現辞書を用いることで，指定薬物を取り扱う EC サイトを効率良く検出でき，IHC の違法判定業務や各県警が運営するボランティアの活動を補助することが可能となる．

2. 固有表現の抽出について

固有表現とは，人名，地名，造語などの固有名詞や，日付，時間に関する表現の総称である．固有表現を抽出するために，小山ら[2]は，用語定義を抽出し，分類・体系化する仕組みを提案している．この手法の評価実験では，再現率が約 50.8%であるときに，最も良い結果として約 82.1%の適合率を得ている．また，上位語を含まない定義文の抽出の失敗と，抽出結果のノイズが原因による用語抽出の失敗が再現率を上げるための課題とされたことが指摘されている．

本研究では，薬物関連の EC サイトから，商品情報を抽出することを目標とする．飯村ら[3]は，EC サイトごとの商品説明ページの定型性を利用し，複数のページ間で共通部分と変化する部分から商品情報抽出ルールを自動生成することを提案している．この手法では商品名を約 90%の精度で抽出できている．本研究で対象とする薬物関連の商品を取り扱う EC サイトは，飯村らの扱う一般的な EC サイトとはいくつかの違いが存在する．例えば，「エキサイト 3g」のように商品名に重量(ロット)が併記されていることが多い． 本研究では，このようないくつかの特徴の違いに対応した抽出ルールを検討する．

3. 提案の概要

商品説明ページの HTML ソースコードから，商品情報の記載箇所の定型性を推測し，薬物販売 EC サイトからの固有表現抽出のためのルールを構築する．

検討の手順は以下の通りである.

1) 同一 EC サイトに存在する各商品説明ページの HTML ソースコードの差分を抽出
2) 差分箇所同士を比較し，ページの特徴や商品属性を観察
3) 商品属性の該当箇所については形態素解析を行い，品詞の割合を観察
4) (2)と(3)から抽出ルールの構築

4. 抽出ルールの構築に向けて

薬物を取り扱う EC サイトにはどういった商品属性が存在するのかを観察するため，サンプルとして 11 件の合法ハーブを取り扱う EC サイトごとに商品説明ページのみを 10 ページずつ収集した. 手順については，前述したとおり進めた．商品説明ページの共通部分を切り捨て，差分箇所から商品属性を抽出するために薬物に関わる固有表現の特徴を観察し，まとめたものを表 1 に示す.

表 1 薬物に関わる固有表現の特徴

商品名	title要素に含まれている場合が多い.
重量	商品名と一緒に記載されていることが多い．例)エキサイト 3g
商品説明	商品についての説明や使用方法が記載されている．比較的長いテキストノードであり，div要素に含まれている場合が多い.
価格	周辺に「価格」や「円」または通貨記号が現れる.
商品ID	input要素のValueに記載されている場合が多い

表 1 で示した固有表現の特徴を別の視点で観察するために，HTML ソースコード中の商品属性に該当する箇所について形態素解析を行った．形態素解析は MeCab を用い，辞書にない単語は「未知語」として出力させた．同一 EC サイトの各商品説明ページ 59 件を対象とし，その結果を表 2 に示す．商品名については複合名詞が 30 件と未知語が 29 件出力された．重量や価格，商品 ID は未知語となった.

表 2 商品属性ごとの品詞の割合

品詞 / 属性	記号	助詞	助動詞	動詞	副詞	未知語	名詞
商品名		3.3%	1.1%	3.3%	3.3%	29.7%	59.3%
重量						100.0%	
商品説明	17.3%	9.1%	0.1%	0.4%	0.4%	24.1%	51.2%
価格						100.0%	
商品ID						100.0%	

5. おわりに

今回は，商品情報を抽出するため，固有表現をまとめたり重量や商品 ID などの特徴のある表現を分離し，未知語としたりすることができた．今後の課題は，抽出ルールを実装し，固有表現を自動抽出することである．同一 EC サイトの各商品説明ページでは差分に固有表現が存在しているということがわかっているので，この性質を用いて抽出可能か否かを検証していく．また，固有表現として薬物関連の商品名などの抽出が可能になれば，厚生労働省が公開している指定薬物の製品リストと自動的に比較ができ，違法薬物を扱う通販サイトを検出することが可能となる.

参考文献

[1] インターネット・ホットラインセンター：ホットライン運用ガイドライン，http://www.iajapan.org/hotline/center/20150401guide.pdf 参照 2015-10-23).
[2] 小山誠，酒井哲也，真鍋俊彦：新聞記事からの用語定義の抽出と固有表現クラスに基づく分類，自然言語処理研究報告，情報処理学会，Vol.2004，No.93，pp.45-51，2004.
[3] 飯村結香子，真鍋知博,塩原寿子，内山匠：EC サイトからの商品情報抽出ルールの自動生成，情報処理学会研究報告，情報処理学会，Vol.2012，No.13，pp.1-7，2012.