2U-06

# 不適切なテキストコンテンツの検出法 —Doc2Vecを橋梁とした多言語対応—

相川　和希†　　河合　新‡　　延原　肇‡
†筑波大学理工学群工学システム学類　　‡筑波大学 システム情報系

## 1　はじめに

本研究の目的は、web サービスにおける不適切なテキストコンテンツの自動検出である。現状、多くの web サービスにおいて不適切なコンテンツの検出や削除は、人手による作業に依存しており、web の発展に伴い従事者数が 10 万人を超えるという報告もある [1]。この不適切コンテンツの抽出作業は、単純労働かつ精神的に悪影響のあるコンテンツに直接触れることになるため、従事者の労働環境が社会問題となっていると同時に、AI 等による自動化が切望されている。

当該領域における関連研究として wiki-detox[2] が提案されている。この研究では、英語の wikipedia の議論ページとコメントに、攻撃的であるかのラベルを付与した約 10 万件のデータを用意し、Multi Layer Perceptron (MLP) およびロジスティック回帰による機械学習を行い、分類器を構成している [3]。

国内における当該分野の関連研究としては、肥合ら [4] による研究があり、他者への攻撃の一つである皮肉の自動検出が提案されている。この研究では、Twitter の投稿約 1000 件のデータセットを準備し、日本語における皮肉の検出に対する機械学習の有効性を示している。

これらの従来研究は高精度検出を実現している一方で、多言語には対応しておらず、例えば、wiki-detox[2] は英語のテキストコンテンツのみしか対応していない。今後、不適切なコンテンツの自動抽出が世界的な問題になってきた場合、すでに蓄積のある言語の資源を利用した多国語対応への拡張が必要となると考える。

本研究では、先行研究における巨大なデータセットに基づく高精度な分類器と機械翻訳を組み合わせることにより、データセットを有しない言語の環境においても、不適切なテキストコンテンツの自動検出を実現することを目的とする。これを実現するため、Doc2Vec[5] による類似文章の抽出を棟梁とし、テストデータの文章を用いて不適切度の算出を行う Neural Machine Translation - Doc2Vec(NMT-D2V) 手法を提案する

## 2　提案手法 (NMT-D2V)

本研究で提案する手法の概要を図 1 に示す。提案手法では日本語文章を入力とし、英語文章への変換処理、英語文章の評価処理、合成処理を行い日本語文章の不適切度を出力する。

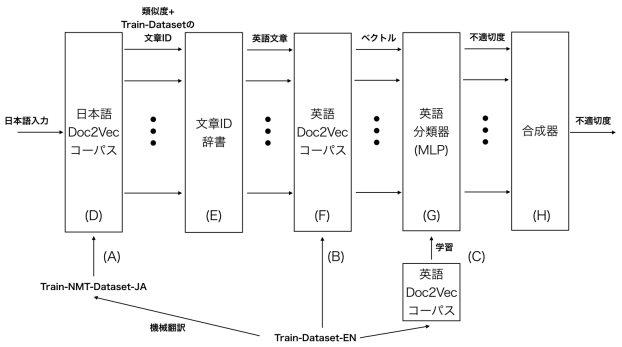


図 1: システム構成図

この手法では、以下の 4 つのデータセットを用いる。

- Train-Dataset-EN（[2] 学習データセット 約 2 万件）
- Train-NMT-Dataset-JA（[2] の学習データを機械翻訳したセット 約 7 万件）
- Test-NMT-Dataset-JA（[2] の学習データを機械翻訳したテストデータ 約 2 万件）
- Test-Wiki-Note-Dataset-JA（日本語 検証用 約 1000 件に手動でアノテーションを付与)

以上のデータセットを用いて、以下の学習処理を行いシステムを構成する。

(A) Train-Dataset-EN を機械翻訳した Train-NMT-Dataset-JA により日本語 Doc2Vec コーパスを構成
(B) Train-Dataset-EN により英語 Doc2Vec コーパスを構成
(C) Train-Dataset-EN を英語 Doc2Vec コーパスによりベクトル化。これによる学習で英語分類器（MLP）を構成

---

Detection method for inappropriate text contents Multilingual support with Doc2Vec as a bridge
†Kazuki AIKAWA　‡Shin KAWAI　‡Hajime NOBUHARA
†College of Engineering Systems, School of Science and Engineering, Tsukuba University
‡Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, Tsukuba University

構成されたシステムは入力に対し以下の処理を行い不適切度を算出する。

(D) 入力した日本語文章を日本語 Doc2Vec コーパスによりベクトル化。Train-Dataset との比較を行い、類似度の高い順に文章 ID と値を算出
(E) 文章 ID から英語文章を獲得
(F) 英語文章を英語 Doc2Vec コーパスに入力しベクトルを算出
(G) 英語分類機に入力し、不適切度の算出
(H) 合成器で各文章の不適切度を統合し、最終不適切度として出力

日本語 Doc2Vec コーパスにより類似度の高い順に文章を複数得ることができる。類似度の高い文章の選択方法として、上位 3 種類、上位 10 種類、0 から 1 の範囲で得られる類似度が 0.9 以上の 3 種類を提案する。それぞれの文章の不適切度の合成方法として、最大値を取る手法と、類似度により重み付けした平均を取る手法の 2 種類を提案する。

従来手法では各言語においてデータセットが必要であるが、提案手法は機械翻訳を組み合わせることで、各言語のデータセットを手動で作成する必要がない。すなわち、不適切コンテンツの抽出に従事する人の作業を削減することに連動する。また、テキストによる検出であるため、他のメディアをテキスト化するシステムと組み合わせることで画像や動画等にも拡張可能であり、現在社会問題となっている不適切コンテンツ除去に関する労働環境を改善する、本質的なソリューションになっていると言える。

## 3 不適切なテキストコンテンツの分類実験

提案手法の有効性を確認するため、2 種類のテストデータを用いた実験を行う。ここで、テストデータは、機械翻訳による Test-NMT-Dataset-JA、手動による Test-Wiki-Note-Dataset-JA、とする。提案手法の調整可能な部分、具体的には図 1 の (D) および (H) の類似文章の選択方法 3 種類 (上位 3 種類、上位 10 種類、類似度が 0.9 以上) と合成方法 2 種類 (最大値、類似度による重み付け平均) を考え、それぞれ個人攻撃文章の検出の性能評価を行う。性能評価の指標として、0 から 1 の範囲での出力される不適切度が閾値よりも大きい場合に不適切と判定し、Receiver Operating Characteristic(ROC) 曲線の Area under the curve(AUC) の値を算出する。

表 1 から表 4 に示す結果のように、最適値はどちらのデータセットにおいても、閾値 0.10, 重み付け平均、上位 3 の組み合わせである。Test-Wiki-Note-Dataset-JA よりも Test-NMT-Dataset-JA の方が全体的に高い精度で検出されている。本手法では学習データに機械翻訳を行なった日本語を用いているため、機械翻訳の影響が大きく、この点については今後の課題として残る。一方で、現状、高い検出精度は得られていないが、データセットの存在しない言語において攻撃的な文章の検出が可能であること示唆できる。

表 1: Test-NMT-Dataset-JA 最大値

| 閾値 | 上位 3 | 上位 10 | 類似度 0.9 以上 |
|---|---|---|---|
| 0.10 | 0.722 | 0.714 | 0.721 |
| 0.20 | 0.743 | 0.737 | 0.740 |
| 0.50 | 0.713 | 0.710 | 0.715 |

表 2: Test-NMT-Dataset-JA 重み付け平均

| 閾値 | 上位 3 | 上位 10 | 類似度 0.9 以上 |
|---|---|---|---|
| 0.10 | **0.753** | 0.745 | **0.753** |
| 0.20 | 0.742 | 0.737 | 0.740 |
| 0.50 | 0.626 | 0.631 | 0.630 |

表 3: Test-Wiki-Note-Dataset-JA 最大値

| 閾値 | 上位 3 | 上位 10 | 類似度 0.9 以上 |
|---|---|---|---|
| 0.10 | 0.637 | 0.613 | 0.635 |
| 0.20 | 0.591 | 0.582 | 0.561 |
| 0.50 | 0.570 | 0.571 | 0.559 |

表 4: Test-Wiki-Note-Dataset-JA 重み付け平均

| 閾値 | 上位 3 | 上位 10 | 類似度 0.9 以上 |
|---|---|---|---|
| 0.10 | **0.656** | 0.612 | 0.617 |
| 0.20 | 0.625 | 0.567 | 0.591 |
| 0.50 | 0.518 | 0.516 | 0.555 |

## 参考文献

[1] WIRED コンテンツモデレーション SNS のダークサイドを見つめる仕事 https://wired.jp/special/2016/unseen (2016)

[2] wiki-detox https://github.com/ewulczyn/wiki-detox

[3] Ex Machina: Personal Attacks Seen at Scale, Ellery Wulczyn, Nithum Thain, Lucas Dixon arXiv (2016)

[4] 関係ベクトルを利用した皮肉の検出, 肥合 智史, 嶋田 和孝, 言語処理学会 第 24 回年次大会 発表論文集 (2018)

[5] Distributed Representations of Sentences and Documents, Quoc Le, Tomas Mikolov ICML (2014)