4W-03

機械翻訳と多言語 RoBERTa を利用した日本語での攻撃的な言葉の自動検出

藤原 知樹 †　伊藤 彰則 ‡　能勢 隆 §
東北大学大学院工学研究科 †　東北大学大学院工学研究科 ‡　東北大学大学院工学研究科 §

1 はじめに

ハラスメントやヘイトスピーチなどの攻撃的な言葉を自動的に検出する研究が盛んに行われており, 公開されているコーパスが数多く存在する. しかしながら, これらのコーパスのほとんどは英語であり, 公開されている日本語のコーパスは存在しない. そこで本研究では, 機械翻訳や多言語対応の言語モデルを用いて外国語コーパスを利用し, 日本語の高性能な検出モデルの構築を目指す.

2 機械翻訳と多言語 RoBERTa を用いた外国語コーパスの利用

他の自然言語処理分野の研究と同様, 近年の攻撃的な言葉の自動検出手法の多くは事前学習済みの言語モデル (Pre-trained Language Model; PLM) を使用している. PLM を用いて攻撃的な言葉を検出するには, 攻撃的か否かのラベルが付与されたデータセットを用いて PLM を Fine-tuning する必要があるが, このようなラベルが付与された公開されている日本語のコーパスは存在しない.

このようなデータ不足の問題は英語以外の外国語でも生じており, 機械翻訳 [1] や多言語版の PLM[2] を用いて外国語コーパスを利用する手法が提案されている. そのような研究では評価対象言語のコーパスが学習に使用されるが, 外国語コーパスを利用する効果が不明であった. そこで本研究では, 機械翻訳や多言語版 PLM を用いて外国語コーパスを利用することで, 十分な日本語データがなくとも高性能な検出モデルが構築可能か検証する.

3 実験

3.1 検出モデル

PLM については英語 BERT (bert-base-uncased), 日本語 BERT (cl-tohoku/bert-base-Japanese-v2), 多言語 RoBERTa (xlm-roberta) の 3 種類を比較する. いずれも, HuggingFace 社が提供している Transformers にて公開されている PyTorch 版の事前学習済みモデルを使用する. 検出モデルは PLM の最終層から得られる [CLS] トークンを入力として, 1 層の Dropout 層と線形層からなる構造である. エポック数は 3 エポック, 言語モデルの最大トークン長は 256, 最適化手法は Adam, 損失関数は MSE Loss, Dropout 率は 0.1 とする. また, 本研究では Google 社の Jigsaw が提供している文の有害度を測定する無料 API の Perspective API[*1]の日本語版の判定性能とも比較を行う.

3.2 データセット

英語コーパス Offensive Language Identification Dataset (OLID)[3], および, アラビア語, デンマーク語, トルコ語からなる多言語コーパス (Multi) を使用する. また, OLID を機械翻訳ツール DeepL[*2] で日本語に翻訳したデータセット (Japanese OLID; JOLID) を作成した. さらに, 機械翻訳を使用しない日本語のデータセットとして Noisy を作成した. Noisy は, 攻撃的な文 (OFFensive) をネット上の悪口やハラスメントの例文から, 攻撃的でない文 (NOT offensive) を対話コーパスからランダムに収集した, ラベルの信頼性が低いデータセットである. OLID 及び JOLID については学習用データの 1 割を検証データとして使用し, Noisy については学習データに加える場合は 5 分割交差検証を行う.

評価データには JOLID, Noisy, Fujihara の 3 種類のデータセットを使用する. Fujihara データセットは, 複数の対話コーパスから収集した文に対し, 第 1 著者が攻撃的か否かのアノテーションを行ったデータセットである. ただし, 英語 BERT を用いた検出モデルを評価するときのみ, JOLID は翻訳前の英語原文を, Noisy 及び Fujihara データセットについては DeepL で英語に翻訳した文を入力する. 学習, 評価データセットの内訳を表 1 に示す.

表 1　各データセットのサンプル数

クラス	Multi	JOLID		Noisy		Fujihara
	学習	学習	評価	学習	評価	評価
OFF	8,104	3,960	240	653	163	104
NOT	34,613	7,956	620	2,000	500	280

Japanese offensive language detection using machine translation and XLM-RoBERTa

† Tomoki Fujihara, Graduate School of Engineering, Tohoku University
‡ Akinori Ito, Graduate School of Engineering, Tohoku University
§ Takashi Nose, Graduate School of Engineering, Tohoku University

[*1] Perspective: https://perspectiveapi.com/
[*2] DeepL: https://www.deepl.com/translator

3.3 評価方法

評価指標には攻撃的なサンプルを陽性とした F 値を使用する．また, いずれの学習条件も 5 個のランダムシードで初期化し, 学習及び評価を行う．評価方法としては, まず一つの評価データセットに対して各学習条件で得られる 5 個の検出モデルの F 値を算出し, それらの算術平均を取ることでその評価データセットに対する F 値を求める．同様にして前述の 3 種類の評価データセットそれぞれに対する F 値を算出し, それらの調和平均をとることで 3 種類の評価データセットに対する平均的な検出性能を求める．

3.4 検出モデルの比較

検出モデルの比較結果を図 1 に示す．検出モデルごとに 3 種類の評価データそれぞれに対する性能を棒グラフで, それらの調和平均を折線で示している．英語 BERT は OLID のみ, 日本語 BERT は JOLID のみで学習しており, 多言語 RoBERTa は, OLID のみ, JOLID のみ, または, それらを混ぜたデータ OLID + JOLID で学習した場合の結果を示している．結果としては, 日本語 BERT もしくは多言語 RoBERTa を用いた検出モデルが同程度に高性能で, 英語 BERT を用いた検出モデルはやや性能が劣っており, Perspective API を用いた検出モデルはかなり性能が低くなった．続く学習データの比較では, 全て多言語 RoBERTa を使用する．

3.5 学習データの組み合わせの効果

図 2 では Noisy, OLID, JOLID の様々な組み合わせを比較しており, 横軸が学習データの組み合わせを表している．また図中の NoisyOFF は Noisy データのうち攻撃的なサンプルのみを使用していることを, × 20 は 20 倍にオーバーサンプリングしていることを表している．ただし, Noisy データの OFF サンプル数を 20 倍すると, JOLID のサンプル数の総数と概ね同数となる．

まず, Noisy のみに比べて Noisy + JOLID の方が調和平均が高い．次に, Noisy や NoisyOFF × 20 に対して OLID または JOLID を追加した場合を比較すると, JOLID を追加した場合の方が調和平均が高い．このことから, Noisy のみでは学習データとして不十分であり, そこに英語コーパスを追加する場合, 日本語に機械翻訳して追加する方が検出性能が高くなることが示された．この要因としては, 機械翻訳に伴う文の自然性劣化や意味の変化よりも, 言語が異なることの方が学習効果の減少を招いている可能性が考えられる．また, Noisy と NoisyOFF を比較すると, NoisyOFF を使用する方が調和平均が高くなっており, 対話コーパスからランダムに収集した文を攻撃的でないと仮定して学習データに加えるのは, 学習に悪影響がある可能性が示唆された．

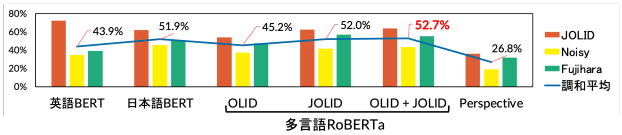

図 1　検出モデルの比較

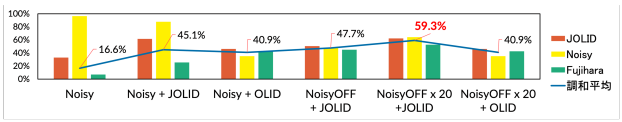

図 2　翻訳データの効果

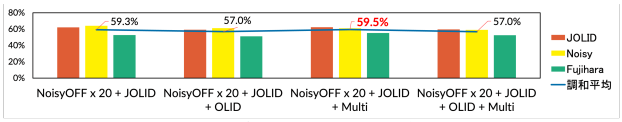

図 3　多言語データの効果

さらに, 図 3 では図 2 において最も性能の高かった条件に英語コーパス OLID や多言語コーパス Multi を追加した結果を示している．結果としては, Multi を追加した場合は僅かに調和平均が大きく, OLID を追加した場合はやや調和平均が小さくなっており, いずれの場合でも検出性能の大幅な向上は生じなかった．また, 本研究で得られた F 値は最高で 59.5% であるが, 最新の英語の検出手法の多くは F 値が約 7〜8 割であるため, 十分な検出性能に達したとは言えない．

4 まとめ

本研究では, 外国語コーパスを利用することで高性能な日本語の検出モデルが構築可能か検証した．結果としては, 機械翻訳を用いることでデータ不足に伴う検出性能の低さを改善可能であることが示された．一方で, 多言語 RoBERTa を用いて外国語コーパスを追加する手法については検出性能の大幅な向上は生じなかった．今後はまず小規模な日本語コーパスを構築する予定であり, その際に本研究で得られた検出モデルを用いて攻撃的な文が効率よく収集可能か検証する予定である．

謝辞 本研究の一部は, JSPS 科研費 20K20799 の支援を受けた．

参考文献

[1] F.Z. El-Alami et al. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University-Computer and Information Sciences*, Vol. 34, No. 8, pp. 6048–6056, 2022.

[2] S Wang et al. Galileo at semeval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. *arXiv preprint arXiv:2010.03542*, 2020.

[3] M Zampieri et al. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*, 2019.