

統計的尺度に基づく有害表現抽出手法の自動評価

Automatic Evaluation for Cyberbullying Detection Method based on Statistical Scale

荒田 真輝^{*1} 梶井 文人^{*2} プタシンスキ ミハウ^{*3}
Masaki Arata Fumito Masui Michal Ptaszynski

^{*1}北見工業大学大学院 工学研究科 情報システム工学専攻
Kitami Institute of Technology Department of Computer Science

^{*2,3}北見工業大学 地域未来デザイン工学科
Kitami Institute of Technology School of Regional Innovation and Social Design Engineering

Various methods have been proposed to detect harmful information automatically as a support of cyberbullying countermeasures. However, there remain various problems such as detection performance deterioration over time and detection of information including private information. In this paper, we propose an automatic evaluation method for harmful expression extraction method using BLEU, an automatic evaluation method typically used in evaluation of machine translation methods.. In this method, for as a detected as a harmful by the harmful expression extraction method, the evaluation is performed based on word N-gram Precision using the harmful entry gold standard dataset. As a result of the evaluation, the proposed allowed assigning a high score to harmful entries, which confirmed the effectiveness of the method.

1. はじめに

インターネットの普及に伴い、Web 掲示板や SNS 等に特定個人への誹謗中傷や個人情報等を書込む「ネットいじめ」が社会問題となっている [1]。しかし「ネットいじめ」の対処には膨大なコストがかかることや、監視自体が困難になってきているなど多くの課題があり、これらに対処するための有害書き込み自動検出手法が数多く提案されている。

新田ら [2] は松葉ら [3] が提案した PMI-IR を用いた有害極性判定手法を拡張し、カテゴリ別関連度最大化手法 (CRM 手法) を提案している。新田らは松葉らが用いた種単語と呼ばれる有害極性単語を 3 カテゴリに分類し、種単語を各カテゴリに頻出の 3 単語ずつ含めることにより、90%を超える精度で有害書き込みを判定できたと報告している。

しかし、後に畠山ら [4] が CRM 手法を再現した際、精度が低下していたことが報告されている。畠山らは原因として種単語が時代に適さないものになっていたことを挙げ、有害極性単語の組合せや規模を変化させた結果、人間の判断による有害極性単語の更新が性能に良い影響を及ぼしたと報告している。

また、Zhang ら [5] は学習素性に音素を用いたニューラルネットワークによる有害書き込み検出手法を英語のために提案している。この手法は有害情報に見られる意図的なスペルミスや音素を用いて訂正しており、98.9%の精度で有害情報を判定が可能であった。

上記のように有害表現抽出手法は数多く提案されているが、課題も多く残されている。そこで、本稿では各手法における課題の整理に向けた、有害表現抽出手法の自動評価手法を提案する。具体的には、Papineni らにより提案された、機械翻訳における自動評価手法である BLEU [6] を有害表現抽出手法の評価へと応用、自動評価を行う。

2. 基本的な考え方

本章では、機械翻訳の自動評価手法 BLEU を活用した有害表現抽出手法の自動評価について説明する。

2.1 機械翻訳自動評価手法 BLEU

BLEU は Papineni らにより提案された機械翻訳における自動評価手法である。

BLEU では、機械翻訳文とプロの翻訳者による翻訳文の似ている度合いを数値で表すため、対象文を 2 つの文章間の N-gram 一致度 P_n (式 1) と、機械翻訳文の長さに応じたペナルティ係数 BP (式 2) を用いて表される式 3 により評価する。また、BLEU は 0~1 の範囲になり、0 は類似度なし、1 は完全な類似度を表している。

$$P_n = \frac{\text{ある参照訳との単語 } n\text{-gram 共有数の最大値}}{\text{MT 訳中の単語 } n\text{-gram 数}} \quad (1)$$

$$BP = \begin{cases} 1 & (c > r) \\ e^{(1-r/c)} & (c \leq r) \end{cases} \quad (2)$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N \frac{1}{N} \log P_n\right) \quad (3)$$

2.2 BLEU の有害表現抽出手法評価への応用

本研究でも、BLEU の評価の根拠であるプロの翻訳者による翻訳文が似ているほど機械翻訳文の質が良いという考え方と同様に、プロとしてネットいじめ書き込みを手作業で検索し通報するインターネットパトロールメンバーが集めた有害な文書と似た文書は有害であるとの考え方を根拠に、本提案手法による有害表現抽出手法の評価を行う。有害表現抽出手法が有害と判断した文書に対して、人間により有害であると判断されたネット上から抽出された文書を参照し、対象の文と有害な文の間の N-gram 一致度 P_n (式 4) を算出する。

$$P_n = \frac{\text{ある有害文書との単語 } n\text{-gram 共有数の最大値}}{\text{対象の評価文中の単語 } n\text{-gram 数}} \quad (4)$$

連絡先: 北見工業大学, 〒 090-8507 北海道北見市公園町 165 番地, m1852400016@std.kitami-it.ac.jp

また、機械翻訳文の評価の場合とは異なり、有害な文章の評価では、とても短い文章の場合でも有害な文章である場合がある。そのため、有害な文章としての評価である本提案手法では、BLEU において評価対象の文の長さに応じてかけられるペナルティ係数 BP の適用は行わず、有害な文との近さを表す harmBLEU スコアは式 5 で求められる。

$$\text{harmBLEU} = \exp\left(\sum_{n=1}^N \frac{1}{N} \log P_n\right) \quad (5)$$

3. 提案手法

本章では、本提案手法の評価の妥当性について確認するため、松葉らが学校非公式サイト等のネット掲示板から抽出した書込み 2,998 件 (有害 1,509 件, 非有害 1,489 件) に対し、本提案手法による評価を行った。以下に評価の流れを示す。

1. データセットから評価対象の書込みを抽出する
2. 対象の書込み以外の有害書込み全てを参照文とし、対象の harmBLEU スコアを算出
3. データセット中の全書込みに対し、harmBLEU スコアを算出し、スコアの高い順に並べ替える
4. 順位毎にその順位以上の書込みを有害な書込み、順位未満の書込みを非有害な書込みとした際の適合率、再現率を算出

各順位における適合率、再現率を図 1 に示す。図 1 より、順位が 1000 位程度の閾値まで適合率が 0.8 以上と高い値が見られる。これより、有害な書込みは相対的にみて非有害な書込みより高い harmBLEU スコアを付与できており、本提案手法が有害書込みの評価に一定の有効性が有ることを示している。

4. 適用実験

次に、既存手法への評価の有効性について調査を行うため、同様のデータセットを用いた畠山らがアンケートを行い人手により有害と判断された単語を使用し改良を行った CRM 手法による有害書込みの判定実験の結果に対し、本提案手法を用いて評価を行った。

CRM 手法では、各書込みに対し有害度合いを表すスコア (SO 値) を付し、一定の閾値を用いて判定を行う手法であるため、本実験では各閾値毎の harmBLEU スコア平均を算出した。閾値による適合率、再現率と harmBLEU スコア平均の推移を図 2 示す。

図 2 より、harmBLEU スコア平均は閾値である SO 値の低下に伴い、緩やかに低下していることが読みとれる。これは改良 CRM 手法における SO 値が高いほど harmBLEU スコアも高い傾向にあることを示している。

5. 結果と考察

本適用実験の結果、本提案手法による改良 CRM 手法の評価は閾値が高い場合に高く、閾値を低くしていくにつれて評価も低くなっていることが分かった。これは、畠山らの改良 CRM 手法において、閾値が厳しい場合は適合率が約 8 割、閾値を緩くしていくと適合率が 5 割に近づき判定が機能なくなっていくという結果と一致しており、本提案手法による評価が有効である可能性を示している。

一方、閾値を低くした場合に着目してみると SO 値の上位約 1500 件を境に改良 CRM 手法の適合率の値が上昇を始めてい

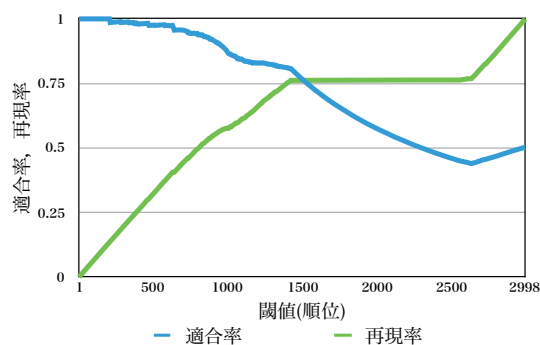


図 1: 各閾値における適合率と再現率

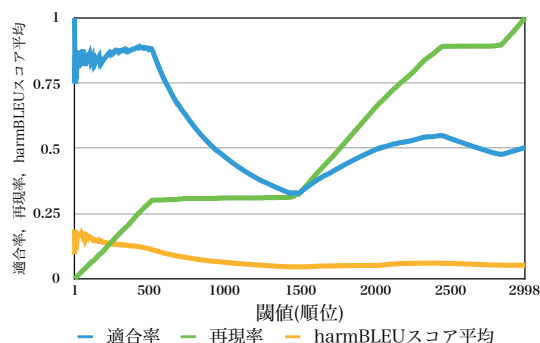


図 2: 各閾値における適合率、再現率と harmBLEU スコア平均

るが、harmBLEU スコア平均はほぼ一定の値を取っていることが読み取れ、本提案手法は有害である可能性が低い文書に対しては上手く評価が出来ていないことが分かった。

6. おわりに

今回、機械翻訳における自動評価手法である BLEU を有害表現抽出手法へと応用し、その手法を自動評価する手法を提案した。本提案手法による掲示板書込みの評価により、本提案手法が有害表現抽出手法の評価について一定の有効性を確認できた。しかし、今回の実験では参照文、対象文共に同じ掲示板の書き込みを利用しており、N-gram 一致度の算出の際において、個人名による影響が大きくなっていることが考えられるため、今後さらに異なるデータを用いた性能も確認する必要がある。

参考文献

- [1] 文部科学省:「「学校ネットパトロールに関する取組事例・資料集」(教育委員会等向け) 資料編第 2 章・巻末資料」, 文部科学省, (2012.9).
- [2] 新田大征, 榊井文人, ブタシンスキ・ミハウ, 木村泰知, ジェブカ・ラファウ, 荒木健治:「カテゴリ別関連度最大化手法に基づく学校非公式サイト上の有害書き込み検出」, 第 27 回人工知能学会全国大会発表論文集, (2013.6).
- [3] 松葉達明, 榊井文人, 河合敦夫, 井須尚紀:「学校非公式サイトにおける有害情報検出を目的とした極性判定モデルに関する研究」, 言語処理学会第 17 回年次大会発表論文集, P2-26(2011.3).
- [4] 畠山鈴生, 榊井文人, ブタシンスキ・ミハウ, 山本和英:「有害表現抽出に対する種単語の影響に関する一考察」, 第 30 回人工知能学会全国大会, (2016.6).
- [5] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whittaker, Joseph P.Mazer, Robin Kowalski, Hongxin Hu, Feng Luo:「Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network」, 15th IEEE International Conference on Machine Learning and Applications, (2016).
- [6] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu:「BLEU: a Method for Automatic Evaluation of Machine Translation」, 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp. 311-318, (2002.7).