E-008

SO-PMI を用いた対話データに含まれる悪口対話の抽出システムの作成
Implementation and evaluation of a system to extract bad dialogue using SO-PMI

山本陽和† 鈴木海友† 井上啓‡ 松澤智史§
Hiyori Yamamoto Kaiyu Suzuki Kei Inoue Tomofumi Matsuzawa

1. はじめに

近年，コンピュータの普及により，書き言葉や話し言葉などを大量に集積したデータである「コーパス」の利用が拡大しており，大量に収集されたデータを機械学習することで，機械翻訳や対話生成技術に活用されている．代表的なものとしてチャットボットなどがある．しかし，多数の対話データの中に誹謗中傷や悪口が含まれた対話が存在した場合，それらも学習され，チャットボットが悪口を発話してしまう場合がある．それらが原因となり，使用しているユーザを不快にさせることや，社会問題に繋がることもありえる．実際にマイクロソフトで開発されたチャットボット「Tay」は学習する中で誹謗中傷に関する対話を学習し暴走を始め，問題となった[1]．チャットボットに悪口表現を出現させないためには，対話生成用学習コーパスから誹謗中傷につながる表現を取り除く必要がある．

本研究では，学習で必要となる対話コーパスから，先行研究の SO-PMI 判定の手法をもとに悪口度算出の改良，文のみの判定から対話文の発話応答を考慮した判定にすることで悪口対話のみを抽出・除去し，非悪口対話のみで構築される対話コーパスの設計・開発を行うことを目標とする．この際，悪口対話の抽出の精度を評価するため，対話コーパスとして非悪口対話と悪口対話のコーパスがそれぞれ必要となる．なお，非悪口対話にあたるコーパスは存在するが，悪口対話にあたるコーパスは存在しなかったため，悪口コーパスの作成も目標とする．

2. 関連研究

Guangwei Wang らの研究[2]では，SO-PMI (Semantic Orientation Using Pointwise Mutual Information)を用いた単語の悪口度の算出による悪口文を自動抽出する研究が行われている．SO-PMI とは事前に「悪口単語 (bad)」と「非悪口単語 ($good$)」の 2 つの基本単語を用意し，対象の単語 ($phrase$)がその 2 つのどちらと共起しやすいかを測ることで悪口度を数値として算出する．悪口度が負の値になれば悪口として，逆に正であれば非悪口として判定している．さらに石坂ら[3]によって，日本語における SO-PMI による単語悪口度の算出が行われ，高い評価が得られている．SO-PMI の式を示すと以下のようになる．

$$C(phrase) = \log_2\left(\frac{hit(phrase,"good")\cdot hit("bad")}{hit(phrase,"bad")\cdot hit("good")}\right) \quad (1)$$

$$f(phrase) = \alpha \cdot \log_2\left(\frac{hit("good")}{hit("bad")}\right) \quad (2)$$

$$SO\text{-}PMI(phrase) = C(phrase) + f(phrase) \quad (3)$$

式(1)は$phrase$が$good$とbadのどちらと共起しやすいかを求め，式(2)は$good$とbadの検索ヒット数の差による優位性を解消するために使用されており，SO-PMI は式(1)，式(2)の和からなる．なお，重み α は 0.0~1.0 のいずれかを与え，本研究では最も判定精度の高かった 0.7 を用いる．また，$hit()$は Web検索ヒット件数を表し，例えば，$hit(phrase)$は $phrase$ を検索した際の Web 検索ヒット件数となる．

† 東京理科大学　理工学研究科　情報科学専攻
‡ 山陽小野田市立山口東京理科大学　工学部　電気工学科
§ 東京理科大学　理工学部　情報科学科

3. 提案手法

本研究では，チャットボットに用いる対話コーパスに対して事前に SO-PMI を用いて対話内容の善悪判定を行い，悪口である対話のみを対話コーパスから抽出・除去するシステムを作成する．善悪判定を行うにあたって，非悪口対話と悪口対話の両方を含むコーパスを作成し，コーパスから SO-PMI で各対話文を分析することで，SO-PMI でどの程度の悪口対話の抽出が行えたかを精度評価できるようにする．(図 1)

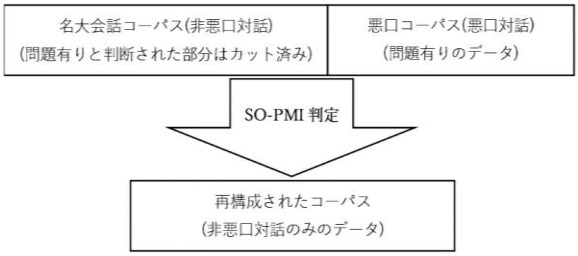

図 1　データコーパスの作成手順

4. 実験

4.1. 非悪口対話コーパスおよび悪口対話コーパスの作成

本研究では，非悪口コーパスに名大会話コーパスを用いた．名大会話コーパスには，プライバシーの問題，その他公開に問題ありと判断された部分はカットされており，コーパス内には明らかに悪口といえる語も見受けられないため，非悪口コーパスとして扱った．悪口コーパス作成には，Twitter より悪口単語が含まれる 1989 対話を収集，さらに悪口の極性が強いものを悪口対話コーパスに付け加えることとした．含まれる悪口単語の一例として「きも」，「馬鹿」，「殺す」，「アホ」など言葉として明らかに悪口と判断できるものを対象としている．また，高村らによって公開されている「単語感情極性対応表」[4]を用いて，発話・応答文に-1.0~1.0 のそれぞれ数値を与え感情分析を行うことで，ネガティブ感情の強い対話を抽出することにした．(図 2)

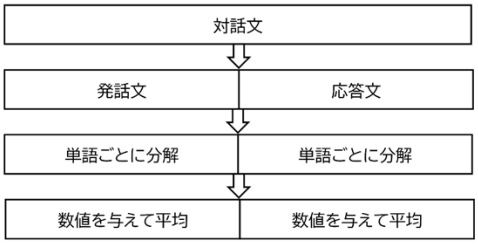

図 2　対話データの判定方法

4.2. SO-PMI による対話コーパスの悪口極性判定の設計

前述の名大会話コーパス 32675 対話の非悪口対話および作成した 305 対話の悪口コーパスを対象に SO-PMI による善悪判定を行い悪口対話の抽出精度を確認した．石坂らの研究[3]では SO-PMI に用いる基本単語「$good$」「bad」に

は「振替」「消えろ」が使用された．しかし，非悪口極性の基本単語"$good$"は単独での出現頻度が高い単語を選んでいるため，$hit(\text{"}good\text{"})$は$hit(\text{"}bad\text{"})$よりも高い値になりやすく，$phrase$が非悪口単語でも SO-PMI 判定では悪口単語と判定されてしまうことがある．そこで，SO-PMI の悪口極性の基本単語「bad」には，悪口単語と共起しやすい単語 2 つを使用し，非悪口単語との関係性を弱くし，悪口単語との関係性を維持した．なお，この際の基本単語「bad」には「消えろ」と悪口単語との自己相互情報量の数値が高く，多くの悪口単語と共起している単語の 1 つであった「死ね」を用いることにした．

5. 実験結果

5.1. 悪口コーパスの作成結果

発話文数値が 0 以下，応答文数値が-0.78 以下で悪口対話が多く出現していたため，この範囲を悪口コーパスとして扱うことにした．作成した悪口コーパスがどれ程の精度で悪口対話を含んでいるか精度の確認をしたところ，抽出した 305 対話の中で悪口対話にあたるであろう対話は 214 対話であり，悪口コーパスの精度としては約 70%となった．

5.2. 再構成された対話コーパスの精度結果

名大会話コーパス 32675 対話の非悪口コーパスおよび 305 対話の悪口コーパスを，SO-PMI を用いて対話の善悪判定を行ったところ，発話文の単語に 0.5 以下の数値があり，かつ応答文の単語に-0.03 以下の数値が存在したとき，悪口対話が多く出現したため，この部分を悪口対話として抽出した．表 1 のように非悪口対話として再構成された対話コーパスは 30,538 対話で，その中で悪口コーパスの 87 対話が誤って判定されていた．混同行列を用いて正解率，適合率，再現率，F 値を確認すると表 2 のような結果となった．

表 1　SO-PMI 判定の結果

混同行列		判定結果	
		非悪口対話と判定	悪口対話と判定
実際の結果	実際に非悪口対話	30451	2224
	実際に悪口対話	87	218

表 2　SO-PMI 判定の精度

正解率	0.930
適合率	0.997
再現率	0.932
真陰性率	0.715
F値	0.963

6. 評価と考察

6.1. 悪口コーパスの作成結果の考察

感情極性分析による悪口コーパスの精度が約 70%であったことから，感情極性分析は，悪口単語を含む文から相手を傷つける攻撃的な悪口文の抽出にも概ね使用できるといえる．また，課題として悪口コーパスの精度の向上が挙げられる．非悪口対話が悪口対話として誤って悪口コーパスとして抽出された原因は，係り受け分析をしなかったためであり，相手ではなく自身を卑下する場合や相手ではなく人以外のモノや概念の場合を考慮することでその精度は向上すると考えられる．

6.2. 再構成された対話コーパスの精度結果の考察

正解率，適合率，再現率，F 値ともに高い評価を得る結果となった．しかし真陰性率は約 71%の精度であり，精度が低くなった．原因としては，悪口コーパスを作成する際，その精度が 70%と低かったためといえる．さらに，悪口対話コーパス作成と同様に，係り受けや否定の意味を持つ単語が出現した場合の対応を行うことで対話の善悪を誤って判定する割合も減るのではないかと考えられる．

7. 今後の課題

悪口対話の抽出精度を上げるため，SO-PMI 判定での善悪判定のみではなく，単語の類似度をベクトルとして表現する Word2Vec[5]の活用も検討する．Word2Vec により悪口単語同士のベクトルの近さを利用することで悪口判別をより適切に行えるのではないかと考えられる．また，BERT[6]による 2 値分類を用いて悪口判定に使用できないかを検討する．BERT は機械学習による文章理解や分類などに使用されており，BERT を悪口判定に応用することで SO-PMI 判定との比較研究を行い精度を調査する．

8. おわりに

本研究では，非悪口対話と悪口対話からなる対話コーパスを作成し，対話コーパスから悪口対話を抽出・除去するシステムの開発を試みた．結果として，悪口対話の収集に関しては，収集された悪口対話としての精度は約 70%となり，SO-PMI 判定による対話の善悪判定に関しては，非悪口対話として再構成された対話コーパスの適合率は約 99.7%で高い評価を得ることができた．しかし，悪口対話の正答率である真陰性率が約 71%であったことから，悪口対話コーパスが多量になると，非悪口対話と誤判定されたものも増えていくため適合率は下がると予測される．そのため，真陰性率を向上させ，誤判定されるものを少なくすることが今後の検討課題である．

参考文献

[1] Microsoft, "Learning from Tay’s introduction", <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>,2022 年 2 月 3 日閲覧

[2] Guangwei Wang and Kenji Araki , “Modifying SO-PMI for Japanese Weblog Opinion Mining by Using a Balancing Factor and Detecting Neutral Expressions” , Proceedings of NAACL-HLT 2007, pp.189-192, Rochester, New York, USA, April, 2007.

[3] 石坂達也,山本和英, “Web 上の誹謗中傷を表す文の自動検出”, 言語処理学会第 17 回年次大会,発表論文集, pp.131-134,2011.

[4] 高村大也, 乾孝司, 奥村学”, スピンモデルによる単語の感情極性抽出", 情報処理学会論文誌, Vol.47 No.02 pp. 627-637, 2006.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, “Distributed Representations of Words and Phrases and their Compositionality”, Advances in Neural Information Processing Systems Vol.26, p.9, 2013.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv preprint arXiv:1810.04805, 2018.