ARG WI2 No.13, 2018
WI2-2018-20

時系列深層学習を用いた言い換え表現の獲得

近江 龍一† 西原 陽子†† 山西 良典††
† 立命館大学大学院情報理工学研究科 †† 立命館大学情報理工学部
{*is0157vv@ed, nisihara@fc, ryama@fc*}*.ritsumei.ac.jp*

概要 本研究では，ドメインや文脈により単語の意味が変化することに着目し，不適切な表現の言い換え表現を時系列深層学習器の LSTM を用いて獲得する手法を新しく提案する．不適切な表現には直接的に表現されるものと間接的に表現されるものがある．間接的に記述された不適切な表現をフィルタリングするためには，単語やその表現が使われているドメインや文脈に応じて意味が変化することを捉える必要がある．提案する手法では LSTM を用いて掲示板のレスポンスの系列を学習し，不適切な文脈のモデルを作成する．LSTM による評価が不適切であるレスポンスに対し，明らかな不適切表現がないものから単語を取り出し，頻度情報を用いて言い換え表現として獲得する．予備的な実験の結果，言い換え表現の一部を獲得できることを確認した．

キーワード 言い換え表現，時系列深層学習，電子掲示板，単語のベクトル表現

1 序論

インターネット上には未成年に対して不適切と指摘される情報が多くある．不適切な情報とは法に触れてしまうものや，差別に関するもの，暴力表現，ギャンブルに関するもの，グロテスクな表現と様々なものがある．不適切な情報は，特にインターネット上の電子掲示板やブログ，同人サイトなど，一般の人が自由に書込み，アップロードができる所で確認できる．これらのサイトには不適切な情報が掲載されることも少なくはない．

不適切な情報を未成年に見せないように，年齢制限や不適切な表現でフィルタリングする対策が行われている．フィルタリング技術の多くは，情報中に不適切な表現が含まれるかどうかを判定し，不適切な表現を含むならばその情報をフィルタリングする．菊池らは 2 つの単語の共起確率を用いることにより不適切さの確率を計算し， 不適切なサイトを高精度に検知する手法を提案している [1]．この手法では，迷惑メールのフィルタリングに多く用いられているベイジアンフィルタを応用している．ベイジアンフィルタでは，過去の不適切なサイトとそうでないサイトから得られた情報から任意の単語について不適切さの確率を計算する．そして，対象とする情報中の全ての単語の不適切さの確率を統合して対象が不適切であるかどうかの判断をしている．

多くの手法は不適切な意味を持つ単語を集めた辞書を用いる．一方で，不適切な表現を含むかどうかを判定する場合に，間接的な表現があると判定が困難になることも多い．フィルタリングされることを避けるために，あえて間接的な記述で不適切な事柄を表現することもある．たとえば，大麻を表す間接的な表現として「葉っぱ」があるが,「葉っぱ」自体は樹木の「葉」としても使われ，表現の意味はそれが使用されるドメインや文脈によって異なる．間接的に記述された不適切な表現をフィルタリングするためには，単語やその表現が使われているドメインや文脈に応じて意味が変化することを捉える必要がある．

本研究では，ドメインや文脈により単語の意味が変化することに着目し，不適切な表現の言い換え表現を時系列深層学習器の LSTM（Long short-term memory）を用いて獲得する手法を新しく提案する．

2 提案手法

提案手法では不適切な表現が含まれる掲示板のスレッドデータを入力として用いる．初めに，スレッドデータに含まれるそれぞれのレスポンスを時間の古いものから新しいものへと順番にインデックスをつける．続いて，レスポンスを形態素解析し，明らかな不適切表現が含まれるかどうかを調べ，含まれている場合には不適切ラベルを付与する．その後，レスポンスを fastText を用いて単語のベクトルにより表現する．そして，この単語のベクトルを LSTM の入力，不適切ラベルを出力として学習させ，モデルを作成する．モデルにより閾値以上の確率が付与され，かつ明らかな不適切表現が含まれないレスポンスから単語を抽出する．抽出された単語の頻度が閾値以上であれば，不適切な表現の言い換え表現として出力する．

2.1 入力：掲示板のスレッドデータの集合

本研究では，電子掲示板の年齢制限のある掲示板を使用する．それぞれのスレッドのレスポンスを形態素解析する．そしてレスポンスに明らかな不適切表現があるかどうかを調べる．明らかな不適切表現は，ニコニコ動画で使用されている NG ワードとした．明らかな不適切表現が含まれるならレスポンスにラベル 1 を付与し，含ま

れないならラベル 0 を付与する．

2.2 レスポンスの単語ベクトル化

スレッドの集合をコーパスとして skip-gram を用いた fastText[2] を使い，単語ベクトルモデルを作る．作成したモデルを用いて，レスポンスの各単語を単語ベクトルにより表現する．

レスポンス内に含まれる各単語の単語ベクトルの平均を取り，レスポンスのベクトルとして扱う．

2.3 LSTM を用いた時系列モデルの作成

時系列深層学習器の一つである LSTM[3] を使用し，時系列モデルを作成する．1 レスポンスの平均ベクトルを LSTM への入力とする．

LSTM では時系列的には不適切表現が含まれる可能性が高いレスポンスに対して，高い確率が付与される．ここで，付与される確率が閾値以上であるが，明らかな不適切表現が含まれないレスポンスについては，不適切な表現の言い換え表現が含まれる可能性が高いとし，レスポンスの単語の中から言い換え表現を獲得する．

確率が閾値 T1 以上かつ，明らかな不適切表現が含まれない，つまりラベルが 0 であるレスポンスから単語を抽出し，頻度をカウントする．頻度が閾値 T2 以上であれば，不適切な表現の言い換え表現とみなす．

3 予備実験

確率が閾値以上のレスポンスから，言い換え表現が獲得可能かを調べる予備的な実験を行った．実験手順は以下の通りである．

まず形態素解析されたコーパスを学習させることで単語ベクトルモデルを得る．分かち書きの際の形態素解析器と辞書にはそれぞれ MeCab と 2018 年 8 月 18 日に更新された NEologd を用いた．コーパスとして 5 ちゃんねるの年齢制限のあるスレッド 10,300 件を使用した．このコーパスで単語ベクトルモデルを学習した．スレッドには明らかな不適切な表現が含まれるレスポンスが約 100 万件，明らかな不適切な表現が含まれないレスポンスが約 560 万件あった，

LSTM での学習においては，5 個の連続するレスポンスが与えられたときに 6 個目のレスポンスが不適切表現を含むかどうかを学習した．

本予備実験では，1 つのスレッドに対して，上記の処理を行い，言い換え表現が含まれる可能性が高いと判定されたレスポンス (閾値 T1=0.40 とした) を人手でチェックし，不適切な表現の言い換え表現を手動で取り出し，予備的な評価を行った．スレッドには 1,000 件のレスポンスがあった．

表 1　レスポンス内における不適切表現の有無それぞれに対する LSTM の出力の集計結果

		レスポンス内の不適切表現	
		ある	ない
出力	不適切	64	72
	不適切ではない	123	741
合計		187	813

3.1 結果に対する考察

表 1 に，実験対象のスレッドのレスポンス内における不適切表現の有無と，それぞれに対する LSTM の出力の集計結果を示す．実験の結果，閾値以上のレスポンスのうち，明らかな不適切表現が含まれるレスポンスは 187 件中 64 件あった．187 件中 64 件で正解率が 34 % と低い値になった．これは LSTM の学習のときに明らかに不適切な表現が含まれるレスポンスが約 100 万件に対し，明らかに不適切な表現が含まれないレスポンスが約 560 万件というデータの偏りが原因と考えられる．

そして閾値以上のレスポンスのうち，明らかな不適切表現が含まれないレスポンスは 72 件あった．そのうち，言い換え表現が含まれたレスポンスは 22 件あった．22 件のレスポンスにはそれぞれ異なる言い換え表現が含まれていた．本予備実験により，言い換え表現が獲得可能と見通しがたった．

4 結論と今後の課題

本論文では，時系列深層学習を用いた言い換え表現の獲得手法を提案した．本論文では予備的な実験を行った．5 ちゃんねるの年齢制限のあるスレッドを 10,300 件用意し，レスポンスの時系列を LSTM により学習させ，モデルを作成し，LSTM により確率が閾値以上であるが，明らかな不適切表現が含まれないレスポンスについて，言い換え表現が含まれるかどうかを調べた．学習した結果，1,000 レスポンス中，確率が閾値以上であるが明らかな不適切表現が含まれないレスポンスは 72 件あり，その中で 22 件のレスポンスには言い換え表現が含まれていた．今後は，全てのスレッドデータに対し評価実験を行い，提案手法の有用性の評価を行っていく．

参考文献

[1] 菊池 琢弥，内海 彰 : 語の共起情報に基づく有害サイトフィルタリング手法，情報科学技術フォーラム講演論文集，Vol.9，No.6，pp.1-6，(2013).

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Trans. of the Association for Computational Linguistics, Vol.5, pp.135–146, (2017).

[3] Sepp Hochreiter, and Jurgen Schmidhuber. Long short-term memory, Neural Computation, Vol.9, No.8, pp.1735–1780, (1997).