様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ４ 日現在

機関番号：１６１０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ１２０２８

研究課題名（和文）テキストから想起した印象抽出によるコンテンツ信憑性判定法の開発

研究課題名（英文）Development of a method for judging the authenticity of content by extracting impressions from texts

研究代表者
森田　和宏（MORITA, Kazuhiro）

徳島大学・大学院社会産業理工学研究部（理工学域）・准教授

研究者番号：２０３２５２５２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：SNSの普及により，ネットに溢れるテキストデータの取り扱いに対する重要性は極めて高い。一方，フィッシングメールなどは人間の不注意を誘うように工夫されている。また，自然災害などの大きな事件が起こるたびにデマやフェイクニュースの拡散が社会問題となっている。本研究では，テキスト情報のみから信憑性を判定するため，文書から受ける印象を抽出し，信憑性分類照合により信憑性を判定する技術の開発と評価をおこなった。

研究成果の学術的意義や社会的意義
人はある情報「地震で電車が脱線した」を見たときに，「大変だ」（内容を信用した肯定的印象），「本当なのか」（疑問を感じた否定的印象）といった何らかの感想（印象）を想起する。この印象に着目し，文章の信憑性を判定する学術的意義がある。また，テキストから受ける印象を起点とすることにより，信憑性辞書の変更のみで結果を制御できるため，SNS拡散への注意喚起，フィッシングメールへの対策のほか，誹謗中傷の検出などにも発展できる社会的意義がある。

研究成果の概要（英文）：With the spread of SNS, the importance of handling text data overflowing on the Internet is extremely high. On the other hand, phishing e-mails are designed to induce human carelessness. In addition, the spread of hoaxes and fake news has become a social problem every time a major incident such as a natural disaster occurs. In this study, in order to judge the authenticity of textual information alone, we developed and evaluated a technology to extract impressions from documents and judge their authenticity by collating classifications.

研究分野：感性情報処理，自然言語処理

キーワード：印象知識

１．研究開始当初の背景

SNS の普及により誰もが簡単に情報発信できるようになった現在において，ネットに溢れるデータはテキストだけでも膨大であり，このデータの取り扱いに対する重要性は極めて高い。情報セキュリティ分野ではアクセス先 URL などの検査，コンテンツ内容の検査など，データそのものの確からしさ（改竄されていないか）に主眼がおかれている。テキスト内容の検査においてもキーワードなどのパターン検出であり，文章の意味をとらえた検査をおこなうわけではない。

一方，フィッシング詐欺の手口では電子メールにもっともらしい文面を載せて受信者を誘導するが，もっともらしい文面を区別して検出することは難しい。従って，受信者が注意して判断しなければならないが，フィッシングメールなどは人間の不注意を誘うように工夫されているため，気づかずに騙されることも少なくない。SNS においては自然災害などの大きな事件が起こるたびにデマ，フェイクニュースの拡散が発生し，社会問題となっている。しかし，デマに関する研究は拡散の分析による推定研究が主であり，投稿文（テキスト）を主体としたものは少ない。詐欺メールやデマ，フェイクニュースはテキスト情報であるため，ウェブ情報における信頼性確保の観点からもテキスト情報を切り口にした研究は重要課題となる。

２．研究の目的

テキスト情報のみから，その文書の信憑性を判定するため，文書から受ける印象を印象語群として抽出し，印象語との信憑性分類照合により信憑性を判定する技術の開発をおこなう。研究期間中に印象抽出，信憑性判定技術の開発と，印象知識，信憑性辞書の構築・拡充，評価を実施する。

３．研究の方法

（１）文書から受ける印象の抽出と信憑性を判定する手法の提案

文書「動物園からライオンが逃げ出した」中のすべての語彙「動物園」「ライオン」「逃げ出す」に対する印象を取得し，それぞれの関連性を探索して文書から受ける印象 {怖い，ミス,...} を決定する手法を考案する。また，信憑性の分類照合により【低い】と判定する手法を開発する。

（２）印象知識と信憑性辞書の構築

小規模の文書データとして，一般的な文書，デマ投稿や詐欺メールなどと，対応する正規の文書を収集する。これらの文書の信憑性分類（【高い】【低い】【やや】【注意】など，分類名は検討する）を実施し，それぞれについて印象を取得することで，印象語に信憑性の分類を紐付けた信憑性辞書を構築する。また，単語を意味成分となる数百次元のベクトルによって表現する分散表現を用いて印象知識を構築する手法の考案もおこなう。

４．研究成果

（１）SNS に対する話題の信憑性判定

信憑性を分類照合により判定する手法の開発を進める中で，深層学習モデルである Self-Attention モデルが単語の重み，つまり単語ごとの注目度合いを学習することに着目し，このモデルを用いた信憑性判定手法を提案した。この手法は機械学習であるため【高い】【低い】といった分類判定を数値，スコアで示す。

具体的には，話題に関するキーワードを含むツイートを形態素解析し，Self-Attention モデルで分類する。分類の結果話題を信じているツイートには 0，疑っているツイートには 0.5 のスコアを付与する。疑っていると分類されたツイートについて，分類の根拠となった重みの最も大きい単語について，学習タスクで算出し重みの平均値をツイートごとにスコアに加算する。ツイートごとのスコアはサンプル時間ごとに平均値を算出し，各時間のスコアとする。スコアは高いほど信憑性が低いことを示す。

「スマホのカーナビ利用が違反になる」という誤った内容の話題に関して，時間ごとのスコアの変化を示したものが図１である。この話題では破線の分類器による結果をそのままスコアにするよりも，提案手法を用いたスコアは正解ラベルに近い値を推移しており，分類誤差を補正することができている。後半にスコアが大きく上がっているのはファクトチェック記事により，デマであることが周知され，Twitter 上で拡散されたためであり，提案手法は Twitter ユーザの反応を正しく反映している。

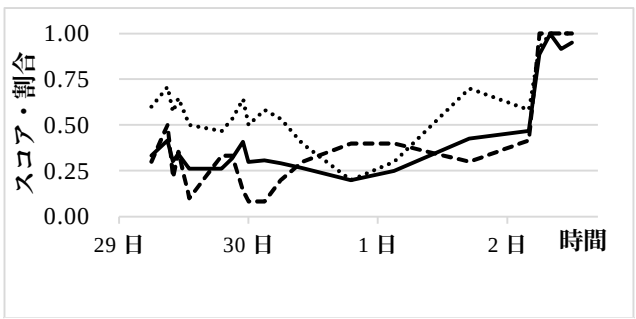

図１ スコアの変化

（2）文書情報のみからの信憑性判定

SNS に対する話題の信憑性判定では時間ごとのスコア算出のために複数のツイートを用いている。本研究の目標として詐欺メールなどへの注意喚起があり，これに対応するため文または文書単体から信憑性を判定する手法を提案した。【注意】の度合いを警戒度という数値で判定する。

具体的には，正規メールとスパムメールを対象に，収集したメールの文章を形態素解析し，単語単位に分割する。分割した文章を一方で，BERT を用いて正規メールかスパムメールか予測する。もう一方で，単語単体でポジティブ・ネガティブの印象を持つ単語を「感情語」と定義し，あらかじめコーパスから感情語辞書を作成しておき，照合しながら文章の感情語の割合を計測する。最後に，BERT での予測結果と感情語の割合を用いて警戒度を算出する。

既存手法である bsfilter，BERT のみの判定と，提案手法で判定したときの正解率を表 1 に示す。bsfilter と比較したとき，正規メールの分類では正解率が落ちるものの，スパムメールの分類では同等かそれ以上の正解率が得られている。また，警戒度を 3 段階に分けメールを分類すると表 2 のようになった。この結果から，スパムメールは正規メールよりも警戒度が顕著に高くなっていることが分かる。よって，危険な文章，つまり信憑性が低い文章を判断できると考える。

表 1　正解率の比較

	bsfilter	BERT	提案手法
正規メール	1.00	0.94	0.96
スパムメール	0.77	0.98	0.99

表 2　3 段階での分類

警戒度	0.0-0.5	0.5-1.5	1.5-2.0
正規メール	43	54	3
スパムメール	0	6	94

（3）分散表現を用いた印象獲得

印象知識と信憑性辞書の構築，拡充を進めていく中で，人手による作業効率を改善する観点から，深層学習モデルを用いた知識構築手法の考案もおこなった。分散表現を用いた手法では，Word2Vec と呼ばれるモデルを用いて意味的類似性の算出により印象語候補を決定する。

具体的には，入力単語に対して分散表現による類似度が高い単語を，形容詞と形容動詞のみ抽出する。分散表現は Word2Vec モデルにより事前学習している。また，文章のある単語と同時に出現することが多い単語を共起語と呼び，ある単語と共起語になっている単語は関連性が強いとされる。この共起語を文単位で抽出する。名詞と形容詞，形容動詞を抽出し，抽出された単語の中で印象語に適切でない単語を除外する。1 文の中で名詞と共起語になっている形容詞，形容動詞をペアにする。分散表現からの抽出語と共起語に共通する高頻度単語 10 語を印象語の候補とする。

印象語候補の抽出例は表 3 のようになる。抽出された候補が印象としてふさわしいかの評価では 60%程度の結果で人間がイメージする印象と異なる候補も存在するが，作業効率の観点からは有用だと考える。

表 3　印象語候補の例

単語	印象語候補
海	青い 自然 美しい 深い 近い 白い 広い 高い 暗い 静か
冬	寒い 暖かい 厳しい 長い 暑い 冷たい 早い 暗い 深い 涼しい

（4）BERT による連想知識抽出

分散表現では意味的類似性を用いるため，「冬」に似た「夏」でも同様の印象語候補が現れていた。このため深層学習モデルの別のアプローチを用いた手法の考案もおこなった。ここでは印象といった感性的な情報抽出の前段階として連想知識抽出を試みている。

具体的には，深層学習モデルの BERT MLM を用いる。MLM は文章中の単語を予測するモデルで，ある単語をマスクした文からその単語を予測する学習をする。この際の入力文を「<元の文章><マスクした文章>」の形で与えるように変更し，マスクの予測単語に元の単語の連想語を与えて学習させた。連想知識の抽出は，対象の単語とその単語を含む文を「<元の文章> <マスクした文章>」の形で与える。

同じ対象単語「皮」について異なるシンプルな文を入力として与えたときの抽出例は表 4 のようになる。抽出された候補が連想知識として相応しいかの評価は 61%程度の結果で分散表現を用いる手法と変わりなかったが，与える入力文によって出力が変化することから，信憑性判定技術へ応用できる可能性が見いだせた。

表 4　「皮」についての連想語候補の例

入力文	連想語候補
バナナの皮	危ない かわいい 工作 春 まち針 細い
魚の皮	固い 美味しい 白い 犬 ラーメン カルシウム

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 菊川 智揮，森田 和宏，泓田 正雄
２．発表標題 文章における信憑性判定のための警戒度算出
３．学会等名 令和4年度 電気・電子・情報関係学会四国支部連合大会
４．発表年 2022年

１．発表者名 西村 聡一郎，森田 和宏，泓田 正雄
２．発表標題 Twitter ユーザの反応に基づく話題の信憑性評価
３．学会等名 令和3年度電気・電子・情報関係学会四国支部連合大会
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関