単語・パラグラフの分散表現を用いたTwitterからの日本語評判情報抽出

芥子 育雄† 鈴木 優† 吉野 幸一郎† 大原 一人‡ 向井 理朗‡ 中村 哲†

†奈良先端科学技術大学院大学情報科学研究科 〒630-0192 奈良県生駒市高山町 8916-5 (けいはんな学研都市)
‡シャープ株式会社コンシューマーエレクトロニクスカンパニー 〒105-0023 東京都港区芝浦 1-2-3

E-mail: †{keshi.ikuo.ka9, ysuzuki, koichiro, s-nakamura}@is.naist.jp,
‡{ohhara.kazuto, mukai-toshiroh}@sharp.co.jp

あらまし Le と Mikolov は文書の分散表現を単語と同様にニューラルネットで学習できるパラグラフベクトルのモデルを提案し，英語の感情分析（ポジティブ，ネガティブ，ニュートラル等の極性判定）ベンチマークを用いて State-of-the-Art の精度を示した[4]．実用上の課題は，性能改善に必要な分散表現（ベクトル）の可読性である．著者らは過去に単語の意味ベクトルとブートストラップ学習を提案した．意味ベクトルの次元は 266 種類の特徴単語に対応し，基本単語約 2 万語については専門家が特徴単語を付与した．意味ベクトルとパラグラフベクトルを統合し，可読性の高いベクトルを自動で学習させることを研究の目的としている．本稿では，クラウドソーシングを利用して Twitter の日本語極性判定ベンチマークを作成し, ツイート中に出現する基本単語を特徴単語に展開することにより，評価実験の結果，パラグラフベクトルによる極性判定の精度を平均 2.3%上回ることを確認した．提案手法は，ポジティブ・ネガティブの極性判定では 84.5%，ニュートラルを含めた極性判定は 67.5%の精度を示した．

キーワード 感情分析, 評判分析, 極性判定, Twitter, クラウドソーシング, パラグラフベクトル, 意味ベクトル, 分散表現, Sentiment Analysis, paragraph2vec, doc2vec, word2vec

1. はじめに

近年，単語やパラグラフの意味表現である分散表現に関する研究が急速に進展している．2013 年に Mikolov 等が発表した単語の分散表現は，文章群を入力としてニューラルネットで各単語の重みベクトルを学習させると，重みベクトルに各単語の意味が学習されることが報告されている[1],[2],[3]．Word2vec プロジェクトとして，オープンソースで公開されていることもあり，単語の分散表現の応用研究が活発化している．

単語やパラグラフの分散表現の応用分野として，Twitter からの評判情報抽出への活用が期待される．著者らの実験では，Twitter を製品ブランド名等の質問文で検索し，質問文に対して収集されたツイートの極性判定精度は,日本語評価極性辞書[5],[6]をベースに機能表現の解釈行った場合で 50%台前半であり実用上は課題があった．極性ラベルが付与された教師データによる機械学習を用いた場合は，テストデータに対してポジティブ，ネガティブの極性判定であれば 75%程度に精度は向上するが，ニュートラルも含めた 3 クラスの極性判定は 60%程度に精度が落ちる．

映画レビューの感情分析タスクにおいて，単語の分散表現の学習をパラグラフに拡張し，パラグラフベクトルをニューラルネットで学習させることにより，エラー率が改善することが報告されている[4]．しかし，文献[4]によると，非常にポジティブ，ポジティブ，ニュートラル，ネガティブ，非常にネガティブの 5 段階の極性判定精度は 50%程度に留まる．

単語やパラグラフの分散表現を工学的に応用する上での課題は，分散表現の可読性が無いため，応用システムの性能改善に必要な分散表現の中身を解釈する手段がないことである.著者らは過去に単語の意味ベクトルとそのブーストラップ学習を用いた連想検索技術を提案し，写真検索，新聞記事の要約，操作ナビゲーション，インターネット検索，番組推薦，音声会話等の実用システムや商用システムに適用してきた[7],[8],[9],[10]．連想検索技術における意味ベクトルは，各次元が特定の特徴単語（概念分類）に対応しているため，ブートストラップ学習のためのテキスト量が少なくても，精度を確保できる特長がある．しかしブートストラップ学習では，テキストにおける単語の出現順を考慮せず，専門家が付与した基本単語の意味ベクトル辞書と対象テキストを元にした統計的学習により単語やパラグラフの意味ベクトルを付与した．また，単語の意味ベクトルはビットベクトルとしたため,単語の意味の分散表現としての解像度は高くなく，次章で述べる“同じ意味の言葉のグルーピング”の要件を十分に満たしているとは言えない．

本研究は，人手で付与した基本単語の意味ベクトル辞書を元にニューラルネットを用いて可読性の高いベクトルを自動で学習させる方法を確立し，単語やパラグラフの分散表現を工学的に取扱い易いものとすることを目的としている．その第一段階として，クラウドソーシングを利用して Twitter の日本語極性判定ベンチマークを作成し，ツイート中に出現する基本単語を特徴単語に展開することにより，短文テキストである

ツイートの概念（個人の意見）がより明確になり，極性判定の精度が改善できるか評価実験を行った．

本稿の提案手法，提案システムの特徴，貢献は以下の 2 点である．

1. 商品開発や品質サポートに役立つ評判情報抽出の視点から，Twitter の極性判定ベンチマークを作成し，Twitter からの日本語評判情報抽出システムを提案する．
2. 基本単語と 266 種類の特徴単語との関係を表した分散表現（意味ベクトル辞書）を用いることにより，Twitter からの日本語評判情報抽出の精度が向上する．

2. Twitter からの日本語評判情報抽出の要件

Twitter は人気のあるオープンなソーシャルネットワークサービスの一つであり，個人の意見を匿名あるいは実名で発信するツールとして活用されている．しかし 1 ツイート最大 140 文字の制限や匿名の情報発信が可能なこともあり，ノイズが非常に多い問題がある．企業にとっては，Twitter から自社の製品・システム・サービス等に関しての評判情報を高精度に抽出できれば，ブラディング，マーケティング，商品企画，品質サポートにおいて有益なツールになると期待されている．

Twitter からの評判情報抽出の企業における用途は，以下の通り分類できる．

- テレビでの紹介やメディア報道に対してのリアルタイムの反響調査
- 一定期間の広告等プロモーション効果の測定
- 新製品とライバル製品との比較や製品の品質に関する顧客の声（Voice of Customer）の活用

リアルタイムの反響調査やプロモーション効果の測定に関しては，単位時間あたりのツイート数やポジティブ，ニュートラル，ネガティブの意見の割合のトレンドグラフと代表ツイートの表示が要件となる．これに対し，自社製品の商品企画や品質サポートにとって有益な顧客の声を捉えるためには，自社製品に対してのポジティブ，ネガティブな意見を発信しているツイートのみを抽出する必要があり，ツイートの解析には以下の要件がある．

- 同じ意味の言葉のグルーピング
  - （ノイズ，ブチブチ音），（反応なし，反応し/ない/難い/辛い），（モッサリ，遅い）
- Twitter ならではの表現のポジネガ判定
  - 死ぬほど好き ⇒ ポジ
  - 頭が痛くなるほど欲しくなる ⇒ ポジ
- 機能表現，係り受けを踏まえたポジネガ判定
  - どうあっても事態が改善する見込がない ⇒ ネガ
  - 理系脳でモノづくり進めて豊かになるなら，A 社や B 社の苦境はないわ ⇒ ネガ

提案システムでは，同じ意味の言葉のグルーピングを目的に大量のラベル無しツイートによる単語ベクトルの学習を検証する．また，Twitter ならではの表現，及び機能表現や係り受けを踏まえた文の特徴を学習するように語順を保持するパラグラフベクトル（4.2 節で述べる PV-DM の結合）の効果を検証する．提案手法の特徴単語の展開は，ツイートの文脈や意味を補完し，文字数の制限やノイズに対して頑健になることを検証する．

3. 関連研究

本章では，Twitter を対象とした英語の極性判定ベンチマーク及びシェアードタスクについて述べ，著者らが構築した極性判定ベンチマーク，及び提案手法との比較を行う．

Twitter を対象とした共通ベンチマークに Stanford Twitter Sentiment corpus (STS) がある[11]．訓練セットはポジティブ，ネガティブ共に 8 万ツイートと多いが，感情を表現する顔文字のエモティコンを用いて，機械的に収集されている．テストセットの 359 ツイートは，製品名，会社名，人名等のトピックで収集し，ツイート中にエモティコンが存在するかどうかとは関係なく，ツイートに感情が含まれているかどうかの判断により，ポジティブ，ネガティブに分類している．ユニグラム等の特徴表現，SVM 等の機械学習のアルゴリズムを用いて 80%以上の精度で 2 クラスに分類できることを示した．しかし，本ベンチマークでは検索対象に対して，個人の意見を発信しているかの判断を行っていない．

Association for Computer Linguistics は，自然言語処理のシェアードタスクの一つとして，2013 年から Twitter を対象に極性判定のコンテストを開催しており，3 年目の SemEval-2015 Task10 は，以下の 5 つのサブタスクから構成される[12]．

- サブタスク A: フレーズ対象の極性判定
- サブタスク B: メッセージ対象の極性判定
- サブタスク C: メッセージに含まれるトピックを対象とした極性判定
- サブタスク D: 一定期間のメッセージを対象にトピックに対する極性トレンド
- サブタスク E: フレーズを対象にポジティブ極性のレベル判定

サブタスク A と B は 2013 年から継続して開催されており，サブタスク C-E は 2015 年に新設されたものである．本稿の要件を満たすサブタスクは C である．

最も人気のあるサブタスクは B で，参加 41 チーム

中, 40 チームがテストセットの回答を提出した．サブタスク B の訓練セットと開発セットは 2013 年に公開されたもので各々9,728 件と 1,654 件，テストセットは毎年公開され 2015 年は 2,392 件である．サブタスク C は 7 チームが回答を提出したが，訓練セット 530 件，テストセット 2,386 件と少ない．極性判定の精度は 50%程度であり，トピックを考慮せずメッセージレベルの極性判定を行うサブタスク B と比較して約 15%精度が落ちる．サブタスク B のトップのチームは，過去のコンテストで好成績だった 4 手法のアンサンブル学習によるものである[13]．各手法は素性や言語リソースに違いがある．この種の Feature Engineering を行わない word2vec を用いたチームは 4 位だった[14]．大規模なラベル無しツイートで学習した単語ベクトルを用いたことが好成績の要因と分析している．サブタスク C のトップのチームは，SVM を用いたシステムだが，言語学的な素性の選択を行っている[15]．このチームは，サブタスク B でも 10 位に入っており，サブタスク B の 1 万件規模の教師データで学習させた分類器を利用して，トピックを考慮した極性判定に適用したことが好成績の要因との分析である．

著者らが構築したベンチマークは，トピックとして 2 種類の製品ブランドを選択し，ツイート中のトピックに対して，個人の意見を発信しているかを判断し，極性を付与した．ベンチマークの規模は，1 種類の製品ブランドに関しては，訓練セット 8,949 件，テストセット 2,983 件と SemEval のサブタスク B に匹敵する規模である．ただし，開発セットは用意せず，訓練セットを用いて提案システムのパラメータ設定を行った．また，単語の分散表現や言語学的な素性を用いた研究は存在するが，本研究のように単語の分散表現として概念分類（特徴単語）との関係を用いた特徴抽出は行われていない．

4. 提案手法

4.1 Twitter からの日本語評判情報抽出システム（提案システム）

意味ベクトルとパラグラフベクトルを統合し，可読性の高いベクトルを自動で学習させることを研究の目的としている．その第一段階として，ツイート中の基本単語を特徴単語に展開することにより，日本語極性判定の精度を検証する．これは，ツイートに加えて，その中の基本単語を特徴単語に展開することは，Twitter の短いテキスト（最大 140 文字）では表現されていない発信者の感情や状況を補って，パラグラフベクトルを学習する効果があるとの仮説に基づく．

提案システムでは，ツイート中の基本単語を特徴単語に展開した後，パラグラフベクトルの PV-DBOW モデルを利用して，ツイートのパラグラフベクトルを学習する．図 1 に提案システムの流れを示す．実線の矢印は訓練セット（ツイート）の流れ，二重線の矢印はテストセット（ツイート）の流れ，破線の矢印はパラグラフベクトルの学習に必要なパラメータの調整を示す．このパラメータは事前学習により決定した．また，訓練セットの正解ラベルは教師あり学習の入力となる．

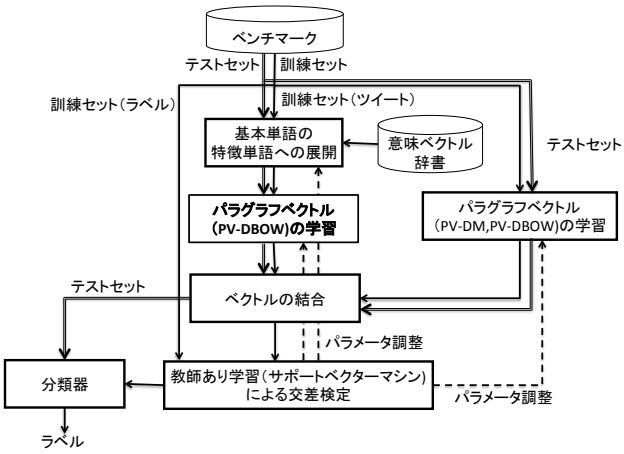

図 1. 提案システムの流れ

提案システムでは，まず，訓練セットのツイートを対象に 4.2 節で述べるパラグラフベクトルの 2 種類のモデル（PV−DM, PV-DBOW）を用いて学習させる．訓練セットは，“ラベルあり”ツイートと大量のラベル無しツイートから構成され，最初にラベル無しツイートの単語ベクトルを学習させる．次に“ラベルあり”ツイートの単語ベクトル，パラグラフベクトルを学習させる．同時に 4.3 節で述べる意味ベクトル辞書を用いて，ツイート中の基本単語を特徴単語に展開し，PV-DBOW モデルを用いてパラグラフベクトルを学習させる．ツイートごとにこれら 3 種類のパラグラフベクトルを結合し，各ツイートの正解ラベルを教師データとして，サポートベクターマシン（SVM）で学習する．交差検定により，SVM のパラメータを最適化し，テストセットの分類器を構築する．

4.2 パラグラフベクトル

パラグラフベクトルの 2 つのモデルを図 2 に示す[4]．PV-DM（paragraph vector with distributed memory）

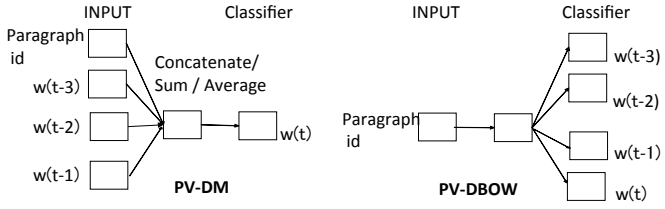

図 2. パラグラフベクトルの 2 種類のモデル

モデルは，ウインドウ内の周辺単語のベクトルにパラグラフのベクトルを追加した文脈ベクトルから，次単語 w(t)のベクトルをニューラルネットにより予測する．PV-DM は，文脈ベクトルと次単語のベクトルとの内積が，周辺単語以外の単語ベクトルとの内積より大きくなるように次単語のベクトルを予測する．パラグラフベクトルは，パラグラフ中の全ウインドウに追加することにより，文脈情報のメモリとしての役割を果たす．PV-DBOW (paragraph vector with distributed bag of words)モデルは，パラグラフベクトルを元にパラグラフ内の単語をランダムにウインドウ長分選択し，単語ベクトルを予測する．語彙数次元が必要な bag of words を数百次元に縮退させたものと捉えることが出来る．

PV-DM や PV-DBOW は，単語ベクトルやパラグラベクトルの初期値をランダムに設定する．PV-DBOW は語順を考慮せずランダムにパラグラフ内の単語を選択するが，PV-DM はウインドウを順にシフトしながら次単語のベクトルを予測することで語順の情報を学習に利用する．また，PV-DM は中間層で周辺の単語ベクトルを結合することが可能であり，この場合は文脈ベクトルとして語順を保持する．

パラグラフベクトルは，後段の SVM，ロジスティック回帰，ニューラルネットなどの教師あり学習による分類器の特徴表現として利用される．パラグラフベクトルの学習が後段の教師あり学習のための特徴抽出として，最も精度が高くなるように各種パラメータ（ウインドウ長，ベクトル長，訓練セットやテストセットの学習回数),中間層でのベクトル生成方法(結合，和，平均）を調整する必要があり，高い精度を確保するためにはヒューリスティックが多く残されている．

4.3 基本単語の意味ベクトル

意味ベクトルは，単語の意味表現として，特徴単語との論理的，連想的関係をベクトル表現したものである．n 個の概念分類を特徴単語とし，各次元が 1 つの特徴単語に対応した n 次元ベクトル空間上の 1 点で，意味を表現するものである．単語の意味ベクトル $X=(x_1, \cdots, x_n)$の各要素を 2 値で表す場合は，単語が特徴単語と関係がある場合は 1，関係ない場合は 0 となる．例えば，特徴単語として{人間，悲しい，芸術，科学，興奮，政治}を採用した場合には，単語「パイロット」の意味ベクトルは(1, 0, 0, 1, 1, 0)となる．このように各特徴単語を関係あり，なしの 2 値で表現することで，分野に依存しない汎用的な意味ベクトル辞書を構築できると考えた.特徴単語として,表 1 に示す通り，6 種類の大分類，29 種類の上位概念に属する 266 種類の概念分類を選択した．2 万 336 語を基本単語として選択し，専門家が論理的関係と連想的関係から，各基本単語に特徴単語を付与した.例を表 2 に示す.また，特徴単語の上位概念，大分類は分類上の目安であり，付与判断の基準は特徴単語そのものである．例えば，特徴単語「温かさ」は上位概念「物理的特性」の下に分類されているが,“心の温かさ”からの連想によって基本単語「愛」に付与した．

表 1. 特徴単語の分類

大分類	上位概念	特徴単語例
人間・生命	人間 生物 ・	人間,人名,男性,女性,子供, 動物,鳥類,虫,微生物,植物, ・
人間環境	人造物 交通・通信 ・	道具,機械機器,建造物, 通信,交通輸送,自動車, ・
自然環境	地域 自然	地名,国名,日本,都会,地方, 陸地,山岳地,天空,海洋,環境,
抽象概念	精神・心理 抽象概念 ・	感覚,感情,喜楽,悲哀, 様子様態,変化,関係関連, ・
物理・物質	運動 物理現象 ・	運動,停止,動的,静的, 蒸発気化,凝固凍結,溶解液化, ・
文明・知識	人文 学術 ・	民族人種,知識.言論発話, 数学,物理学,天文学,地学, ・

表 2. 基本単語に付与された特徴単語の例

基本単語	特徴単語
愛	人間,家族・家庭,性,性問題,感情,喜楽,関係・間連, 肯定的,感情的,優しさ,温かさ,心理学,映像・画像
爽やか	環境,感覚,様子・様態,優良,肯定的,新しさ,明るさ
駅伝	スポーツ,日本,高速,困難,組織,行為,運動,大規模, 長さ,季節,地理

ブートストラップ学習では以下の 2 種類の仮説に基づいて，基本単語を基に全出現単語に意味ベクトル（266 種類の特徴単語）を付与する．

- 仮説 1：文書の意味ベクトル
  一定数以上の基本単語が含まれていれば，その基本単語の意味ベクトルの加重和によって，適切な文脈情報が表現可能
- 仮説 2: 単語の意味ベクトル
  単語が含まれている文書の意味ベクトルの加重和によって，適切な文脈情報を獲得

4.4 基本単語の意味ベクトル（特徴単語）からパラグラフベクトルの学習

ツイート中の基本単語を特徴単語に展開した例を図 3 に示す．(製品 A)，(A 社）はツイート中の固有名詞を置き換えたものである．学習には固有名詞や製品名の記号列そのものを用いる.例のツイートでは,“真偽”“製”“端末”“インチ”“画面”“非常”“魅力”“的”の 8 個の基本単語を含む．前節のブートストラップ学習の仮説 1 を満たしており，パラグラフの意味ベクトルは，基本単語の意味ベクトル（特徴単語の組合せ）

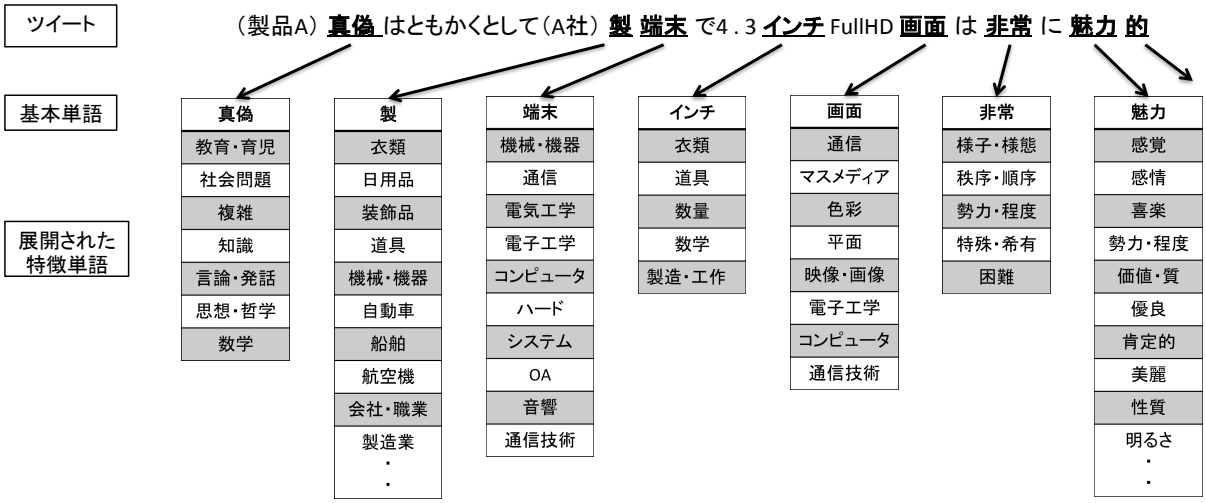

図 3. 特徴単語への展開例

から適切な文脈情報を学習できると考えられる．ここでは，パラグラフ（ツイート）の意味ベクトル構築には PV-DBOW をそのまま用いた．

ツイートの形態素解析を行った後，ツイート中から基本単語を抽出するが，特徴単語に展開する基本単語の品詞，及び展開する特徴単語数の上限はパラメータとして，後段の教師あり機械学習による分類精度を考慮して決定する．

5. Twitter からの日本語評判情報抽出の実験

本章では，提案システムのプロトタイプを作成し，提案手法の効果を確認することを目的に実施した評価実験について述べる．

5.1 手順

Twitter からの日本語評判情抽出の手順を図 4 に示す．本研究では商品企画や品質サポートにとって有益な個人の意見を Twitter から抽出することを目的とし，2 種類のスマートフォンの製品ブランド（以降，A 社製の製品ブランドを製品 A,B 社製の製品ブランドを製品 B と呼ぶ）を対象とした．

前処理では，各製品ブランドに関連したキーワードを元にツイートを収集し，ボットやアフィリエイトと推測される単語や引用 URL を含むツイート，リツイートを除外した．

ベンチマーク構築では，クラウドソーシングを利用して，製品ブランドごとに各ツイートに対してラベル付けを行った．ラベルは以下の 4 種類である．

- **ポジティブ**：対象の製品ブランドに対して，ポジティブな意見を発信しているツイート．
- **ネガティブ**：対象の製品ブランドに対して，ネガティブな意見を発信しているツイート．
- **ニュートラル**：対象の製品ブランドに対して，個人の意見を発信しているが，ポジティブでもネガティブでもないツイート．
- **無関係**：対象の製品ブランドに対しての個人の意見を発信していないツイート．

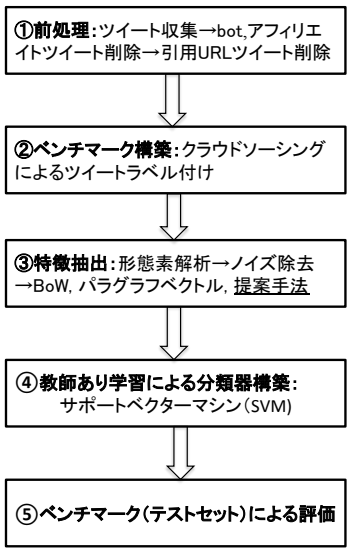

図 4. Twitter からの日本語評判情報抽出の手順

各ツイートに少なくとも 5 人の作業者を割り当て，投票結果を元に評価用ベンチマークを構築した．ベンチマークの詳細については次節で述べる．

特徴抽出では，パラグラフベクトル，提案手法である基本単語の特徴単語展開を利用したパラグラフベクトル，及びベースラインとしての全ツイートから抽出された語彙を次元とする Bag of Words(BoW)を特徴表現として作成した．ツイートからは，あらかじめ以下のノイズ除去を行った．

- ユーザ名（@user），改行の削除

表 3．ベンチマークの構成

	製品 A		製品 B	
	訓練セット	テストセット	訓練セット	テストセット
ポジティブ	**1,346**	**336**	**3,675**	**1,224**
ネガティブ	**1,228**	**307**	**2,385**	**795**
ニュートラル	**1,278**	**319**	**2,889**	**964**
ベンチマーク	**3,852**	**962**	**8,949**	**2,983**
合計	**4,814**		**11,932**	
無関係	8,216		5,998	
その他	1,473		2,891	
ラベル無し	**351,342**			

・ 句読点や記号[’:’，’;’，’(’,’)’，’{’，’}’，’[’，’]’，’．’，’，’，’ “’，’ #’ ］の削除

・ 英文字列は小文字に統一

ツイートから単語を抽出するための日本語形態素解析には MeCab[1]を用いた．

教師あり学習による分類器構築では，SVM を用いて，ベンチマークの訓練セットを対象に交差検定を行い，各種パラメータを決定した．実験方法については，5.3 節で述べる．

ベンチマークによる評価では，テストセットを用いて，ポジティブ，ネガティブ，ニュートラルのツイートを対象に極性判定を行う“3 クラスの分類”とポジティブ，ネガティブのツイートのみを対象に極性判定を行う“2 クラスの分類”により，分類精度を評価する．

5.2 ベンチマーク

クラウドソーシングを利用して構築したベンチマークを表 3 に示す．製品 A のベンチマークは 4,814 ツイート（訓練セット：3,852 件，テストセット：962 件），製品 B のベンチマークは 11,932 ツイート（訓練セット：8,949 件，テストセット：2,983 件）である．複数のラベルが 1 位で同数投票された“その他”のツイート（製品 A が 1,473 件，製品 B が 2,891 件）や“無関係”のラベルが付与されたツイートに関しては，今回の実験の学習対象，及び評価対象から外した．トータル 35,324 ツイートに対し，クラウドソーシングを利用して各ツイート 5 人の作業者を割り当てた．ラベル付与に要した費用は約 2 万円であり，ベンチマークは現実的な予算で構築可能である．これ以外にクラウドソーシングに掛ける前段階のラベル無しツイートが約 35 万件ある．これらは製品 A，B に関して収集したツイートでノイズ除去まで行ったものであり，2 章で述べた同じ意味の言葉のグルーピングを目的に単語ベクトルの学習に用いた．

5.3 実験方法

(1)ベースライン（BoW）の評価実験

製品 A，B の特徴表現（ツイートベクトル）を製品ごとの訓練セットとテストセット両方の語彙を次元とするベクトル（BoW）とし，各ツイートから抽出された語彙の頻度をベクトルの値として用いた．

BoW と SVM による分類器の構築には，Python の機械学習ライブラリである scikit-learn[2]を用いた．訓練セットの BoW に対しての 4 分割交差検定とグリッドサーチにより，SVM のカーネル関数とハイパーパラメータの値を決めた．構築した SVM の分類器を用いて，テストセットの評価を行った．

(2)パラグラフベクトルの評価実験

2 章で述べたツイート解析の要件を満たすために最初にラベル無しツイートを元に単語ベクトルの学習を行った．その上に製品ごとの訓練セットのパラグラフベクトルを学習させた．最後にテストセットのパラグラフベクトルを学習させた．パラグラフベクトルの学習には，Python 用トピックモデルのライブラリである gensim[3]を用いた．

SVM 分類器は，scikit-learn を用いて，訓練セットのパラグラフベクトルに対して，4 分割交差検定とグリッドサーチを元に構築した．また，4.2 節で述べたパラグラフベクトルの各種パラメータや中間層でのベクトル生成方法も予備実験により決定した．ただし，訓練セットによる交差検定時は単語ベクトルも学習しているため，テストセットによる評価とは実験設定が異なる．構築した SVM の分類器とテストセットのパラグラフベクトルを用いて，テストセットの評価を行った．

(3)提案手法の評価実験

(2)で述べたパラグラフベクトルによる評価実験に対して，ラベル無しツイートの学習に 4.4 節で述べたツイート中に含まれる基本単語の特徴単語展開を行った．ここでは基本単語の名詞，形容詞，形容動詞，動詞を対象に特徴単語を最大で 7 語展開した．クラウドソーシングに掛ける前のツイートのため，意味不明なツイートも数多く含まれており，上記品詞の基本単語が全く含まれていないツイート数が約 1 万 8 千件存在した．展開された特徴単語も含めて，単語ベクトルを学習させた．製品ごとの訓練セット，テストセットも同様に基本単語の特徴単語の展開を行い，上記単語ベクトルを読み込んで，パラグラフベクトルを学習させた．SVM の分類器の構築では，(2)の PV-DM と PV-DBOW に特徴単語展開を行って作成した PV-DBOW を結合して，特徴表現として用いた．

[1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[2] http://scikit-learn.org/stable/

[3] http://radimrehurek.com/gensim/

表 4. 製品 A の評価結果

	3 クラスの分類		2 クラスの分類	
	交差検定	テスト	交差検定	テスト
BoW	57.5%	58.6%	73.9%	78.1%
PVEC	62.4%	62.3%	78.9%	80.1%
提案手法	63.4%	**64.7%**	79.6%	**81.0%**

表 5. 製品 B の評価結果

	3 クラスの分類		2 クラスの分類	
	交差検定	テスト	交差検定	テスト
BoW	61.8%	65.1%	79.9%	80.1%
PVEC	64.7%	65.2%	81.7%	82.7%
提案手法	66.3%	**67.5%**	82.9%	**84.5%**

5.4 結果

製品 A の分類精度を表 4 に製品 B の分類精度を表 5 に示す. 表の PVEC はパラグラフベクトルによる結果を示す．同じ条件設定でもパラグラフベクトルは学習の度に異なる特徴ベクトルが生成され，SVM の分類精度が変動するため，5 回試行の平均を採用した．テストセットの評価では，提案手法は，3 クラスの分類では製品 A で 2.4%，製品 B で 2.3%，2 クラスの分類では製品 A で 0.9%，製品 B で 1.8%パラグラフベクトル（以降，PVEC と呼ぶ）の精度を上回った．ラベル無しツイート学習，及び訓練セット，テストセット学習のパラメータ設定を表 6，表 7 に示す．提案手法の製品 A においてのみ文脈ベクトルとして語順を保持する PV-DM の結合を採用したが, 3 クラスの分類では語順を保持しない BoW を 6.1%（相対値 10.4%）上回っており，語順を保持する効果を確認できる．同条件でのウインドウ長 7, 中間層が入力ベクトルの和の場合は, 3 クラスの分類 64.4%, 2 クラスの分類 80.6%である．

表 6. ラベル無しツイート学習のパラメータ設定

	PVEC	提案手法
学習対象	ラベル無しツイート	同左＋ツイート中の基本単語（名詞，動詞，形容詞，形容動詞）を特徴単語展開（上限 7 語）
語彙辞書	学習対象＋訓練セットとテストセットの語彙	学習対象＋特徴単語展開を行った訓練セットとテストセットの語彙
学習方法	PV−DM：400 次元，ウインドウ長 7，中間層は入力ベクトルの和 PV−DBOW：400 次元，ウインドウ長 8	PV−DM，PV-DBOW：同左 PV-DBOW（特徴単語展開）：400 次元，ウインドウ長 10
学習回数	20 回	20 回

表 7. 訓練セット，テストセット学習のパラメータ設定

	PVEC	提案手法
学習対象	訓練セット テストセット	同左＋ツイート中の基本単語（製品 A：動詞，形容詞，形容動詞，製品 B：名詞，動詞，形容詞，形容動詞）を特徴単語展開（製品 A：上限 4 語，製品 B：上限 7 語）
語彙辞書	ラベル無しツイートで学習した単語ベクトル辞書の読み込み	ラベル無しツイートで学習した単語ベクトル辞書の読み込み
学習方法	PV−DM：400 次元，ウインドウ長 7，中間層は入力ベクトルの和 PV−DBOW：400 次元，ウインドウ長 10	PV−DM：400 次元， 製品 A：ウインドウ長 4，中間層は入力ベクトルの結合， 製品 B：ウインドウ長 7，中間層は入力ベクトルの和 PV-DBOW：同左 PV-DBOW（特徴単語展開）：400 次元，ウインドウ長 12
学習回数	訓練セット：5 回 テストセット：10 回	訓練セット：5 回 テストセット：6 回

表 8 製品 B のベンチマークの割合に応じた評価結果

		3 クラスの分類		2 クラスの分類	
	割合	精度（STDEV）	差分	精度（STDEV）	差分
BoW	50%	61.2%		76.0%	
	75%	61.2%		77.2%	
	100%	65.1%		80.1%	
PVEC	50%	62.6%(±0.5%)	1.4%	80.7%(±0.8%)	4.7%
	75%	62.7%(±0.7%)	1.5%	80.7%(±0.9%)	3.5%
	100%	65.2%(±0.3%)	0.1%	82.7%(±0.3%)	2.6%
提案手法	**50%**	**66.4%**(±0.7%)	3.8%	**83.5%**(±0.6%)	2.8%
	75%	**65.1%**(±0.5%)	2.4%	**82.4%**(±0.4%)	1.7%
	100%	**67.5%**(±0.5%)	2.3%	**84.5%**(±0.4%)	1.8%

5.5 考察

2 クラスの分類において，製品 B と比較して製品 A の提案手法による改善割合が小さい．これがベンチマークサイズに依存するものか，データに依存するものかを検証するため，製品 B のベンチマークを用いて，訓練セット，テストセットを一定の割合（75%, 50%）に削減した時のテストセットの評価結果を表 8 に示す．表 8 の STDEV は精度の標準偏差を示す．差分列は，提案手法では PVEC に対しての精度の差分，PVEC では BoW に対しての精度の差分を示す．表 8 から，提案手法による精度の改善はベンチマークサイズに依存しないこと（PVEC では 50%，75%の精度が同程度だが，提案手法では 75%の精度が最も低い), 及び全ての組み合わせにおいて提案手法の精度は PVEC を上回っていることが分かる. PVEC の分類精度は，製品 A の 100%を含めて，BoW を平均 2.3%上回り，提案手法は PVEC をさらに平均 2.3%上回った．

PV-DM の設定は, PVEC の 50%, 75%のみウインドウ長 4, 中間層は入力ベクトルの結合を採用したが，2 クラスの分類において BoW との差分が最も大きい．提案手法の PV-DM は入力ベクトルの和を用いた場合に精度が良かったが，差は平均 0.1%と小さい．他のパラメータは 100%と同じ設定とした．

次にラベル無しツイートによる単語ベクトルの学習効果を検証する．製品 B において，ラベル無しツイートの単語ベクトルを読み込まなかった場合の結果を表 9 に示す．差分列はラベル無しツイートの単語ベクトルを用いた場合との精度の

表 9. 製品 B でラベル無しツイートを利用しない場合

	3 クラスの分類		2 クラスの分類	
	精度（STDEV）	差分	精度（STDEV）	差分
PVEC	61.2%(±0.6%)	-4.0%	77.6%(±2.4%)	-5.1%
提案手法	**63.6%**(±0.2%)	-3.9%	**80.5%**(±0.2%)	-4.0%

表 10. 提案手法により改善，失敗したツイート数

	不正解→正解	正解→不正解
ポジティブ	103	60
ネガティブ	54	54
ニュートラル	106	63

差を示す．提案手法で平均 4.0%, PVEC では平均 4.6%単語ベクトルの効果があった．単語ベクトルを読み込まない場合でも平均 2.7%提案手法の精度は PVEC を上回った. 提案手法は，PVEC と比較して標準偏差も小さく，特徴単語を利用することにより安定性向上が期待できる．本実験では学習に用いるテキスト量が少ないため，頻度 5 以上の単語を学習対象とした．また，PV-DBOW のウインドウ長を PVEC では 7，提案手法では 8，PV−DM は単語ベクトル辞書を用いる場合と同じ設定とした．学習回数は，提案手法では訓練セット 15 回，テストセット 18 回，PVEC では両方 30 回とした．

最後に PVEC の極性判定から，提案手法により分類が正解に変わったツイート数，不正解に変わったツイート数の例を表 10 に示す．製品 B の 3 クラス分類を対象とした．この例では，提案手法により，ポジティブとニュートラルの極性判定が改善し，PVEC と比較して各 43 ツイート正解が増えた．

6. おわりに

本稿では，単語・パラグラフの分散表現を用いた Twitter からの日本語評判情報抽出システムを提案した. 基本単語と 266 種類の特徴単語との関係を表した分散表現（意味ベクトル辞書）を用いることにより，Twitter からの日本語評判情報抽出の性能が向上することを確認した．また，商品開発や品質サポートに役立つ評判情報抽出の視点から，Twitter を対象に日本語の極性判定ベンチマークを作成した．英語では関連研究で述べた通り，Twitter の極性判定に関してシェアードタスクが開催されている．映画レビューに関しては，極性判定の共通ベンチマークが公開されており，研究に利用されている[16]．パラグラフベクトルが 2014 年に State-of-the-Art を出した後，2 度記録が塗り替えられ，感情分析の研究が活発化している[17]．提案手法の有効性検証や日本語の感情分析技術の進展のためにも共通ベンチマークが必要である．著者らは本稿の評価結果を元により規模の大きい，信頼性のある日本語評判情報抽出のためのベンチマーク構築を進める予定である．

提案手法の今後の展開として，特徴抽出のパラメータ設定を教師あり学習からのフィードバックを入れて最適化することにより，さらなる精度向上を目指すことが挙げられる．また，パラグラフベクトルや単語ベクトルの初期値として，意味ベクトルを与えることにより，可読性の高いベクトルの学習を実現し, 提案システムの性能改善, 安定性向上を目指す．

謝辞: 本研究の一部は，NAIST ビッグデータプロジェクトによるものである．

参 考 文 献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. Proc. of Workshop at ICLR, 2013.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. 3111-3119, Proc. of NIPS, 2013.

[3] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. 746-751, Proc. of NAACL HLT, 2013.

[4] Quoc Le, Tomas Mikolov. Distributed Representations of Sentences and Documents. 1188-1196, Proc. of the 31st International Conference on Machine Learning, 2014.

[5] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一. 意見抽出のための評価表現の収集．自然言語処理，Vol.12, No.3, 203-222, 2005.

[6] 東山昌彦, 乾健太郎, 松本裕治, 述語の選択選好性に着目した名詞評価極性の獲得, 言語処理学会第 14 回年次大会論文集, 584-587, 2008.

[7] 芥子育雄, 池内洋, 黒武者健一. 百科事典の知識に基づく画像の連想検索．電子情報通信学会論文誌，Vol.J79-D-II, No.4, 484-491, 1996.

[8] 黒武者健一, 芥子育雄. 連想検索技術を利用した文書の要約. Proc.. Advanced Database System Symposium ’97, 135-142, 1997.

[9] 芥子育雄, 佐藤亮一, 宮川晴光, 黒武者健一, 清水仁. 機能意味検索と操作自動実行に基づくナビゲーションソフト．情処研報．HI,ヒューマンインタフェース Vol.83, 43-48, 1999.

[10] 芥子育雄, 黒武者健一, 佐藤亮一, 河村晃好, 清水仁, 宮川晴光, 伊藤愛, 松岡篤郎, 竹澤創, 紺矢峰弘. ディジタル情報家電のインタフェースエージェント技術の開発．シャープ技報, Vol.77, 15-20, 2000.

[11] Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision. Technical report, Stanforf University, 2009.

[12] Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, Veselin Stoyanov. SemEval-2015 Task 10: Sentiment Analysis in Twitter. 451-463, Proc. of the 9th International Workshop on Semantic Evaluation, 2015.

[13] Matthias Hagen, Martin Potthast, Michel Büchner, Benno Stein. Webis: An Ensemble for Twitter Sentiment Detection. 582-589, Proc. of the 9th International Workshop on Semantic Evaluation, 2015.

[14] Ramon F. Astudillo, Silvio Amir, Wang Ling, Bruno Martins , Ma´rio Silva, Isabel Trancoso. INESC-ID: Sentiment Analysis without hand-coded Features or Liguistic Resources using Embedding Subspaces. 652-656, Proc. of the 9th International Workshop on Semantic Evaluation, 2015.

[15] William Boag, Peter Potash, Anna Rumshisky. TwitterHawk: A Feature Bucket Approach to Sentiment Analysis. 640-646, Proc. of the 9th International Workshop on Semantic Evaluation, 2015.

[16] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts, Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013.

[17] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Modeling CNNs for text: non-linear, non-consecutive convoluteion. 1565-1575, Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015.