単語意味ベクトル辞書を用いたTwitterからの評判情報抽出

芥子　育雄[†a)]　鈴木　優[†]　吉野幸一郎[†]　グラム　ニュービッグ[†*]
大原　一人[††]　向井　理朗[††]　中村　哲[†]

Reputation Information Extraction from Twitter Using a Word Semantic Vector Dictionary

Ikuo KESHI[†a)], Yu SUZUKI[†], Koichiro YOSHINO[†], Graham NEUBIG[†*],
Kazuto OHARA[††], Toshiro MUKAI[††], and Satoshi NAKAMURA[†]

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ワークで学習できるパラグラフベクトルのモデルを提案し，極性分析ベンチマークを用いて最高水準の分類精度を示した．パラグラフベクトルを用いたツイートの極性分析における実用上の課題は，単語のスパース性を解消するパラグラフベクトルの構築のために大規模文書が必要なことである．本研究では，Twitter の文に対して評判情報抽出を適用する際，その出現単語のスパース性に由来する性能低下を解決するため，人手により構築された単語意味ベクトルを導入する．意味ベクトルとして，各次元が 266 種類の特徴単語に対応し，約 2 万語に付与されている単語意味ベクトル辞書を使用する．この辞書を用いて単語拡張したツイートをパラグラフベクトルのモデルで学習するという，単語意味ベクトルとパラグラフベクトルの統合化手法を提案する．これにより，単語がスパースでも特定分野の文脈情報を学習できることが期待される．この評価のため，クラウドソーシングを利用してスマートフォン製品ブランドに関する極性分析ベンチマークを作成した．評価実験の結果，約 1 万 2 千ツイートから構成される特定のスマートフォン製品ブランドのベンチマークにおいて，提案手法は，ポジティブ，ニュートラル，ネガティブの 3 クラス分類におけるポジティブ予測とネガティブ予測のマクロ平均 F 値 71.9 を示した．提案手法は従来手法であるパラグラフベクトルによるマクロ平均 F 値を 3.2 ポイント上回った．

キーワード　極性分析，Twitter，word2vec，パラグラフベクトル，意味ベクトル

1. まえがき

企業において，新製品や品質に関する顧客の声を素早く掴むために，Twitter からの評判情報抽出の必要性が高まっている．文長が短い Twitter から個人の意見を明確にし，自社製品に対してのポジティブ，ネガティブな意見を高精度に抽出できる手法が求められる．

この目的を達成するために，単語や文書の類似度を計算する手法として，分散表現を活用することができると考えられる．Mikolov らが発表した word2vec は，文脈情報を素性としてニューラルネットワークにより学習を行うと，語義の似た単語や語句が似たような重みをもつベクトルを構築することができると報告されている [1]〜[3]．また，単語の分散表現の学習を文書に拡張し，パラグラフベクトルをニューラルネットワークで学習させることにより，複数の極性分析ベンチマークにおいて最高水準の分類精度を示したことが報告されている [4]．

1 ツイートに含まれる単語を元にパラグラフベクトルを学習する場合は，単語のスパース性が課題となる．例えば，**4.** の実験で用いる Twitter データにおいては，総語彙数約 9 万 3 千語（内 64%は出現頻度 2 回以下），1 ツイートあたり平均 7.6 単語であり，出現頻度 2 回以下の単語と製品ブランド名を除くと，平均 2 単語程度で文脈を推定する必要がある．しかし，パラグラフベクトルの学習ではウインドウ長（文脈情報の

† 奈良先端科学技術大学院大学情報科学研究科，生駒市
Graduate School of Information Science, Nara Institute of Science and Technology, 8916–5 Takayama-cho, Ikoma-shi, 630–0192 Japan
†† シャープ株式会社 IoT 通信事業本部，千葉市
IoT Communication Business Unit, Sharp Corporation, 1–9–2 Nakase, Mihama-ku, Chiba-shi, 261–8520 Japan
* 現在，カーネギーメロン大学言語技術研究所
a) E-mail: keshi.ikuo.ka9@is.naist.jp
DOI:10.14923/transinfj.2016DEP0013

 ⓒ一般社団法人電子情報通信学会 2017

単語数）は少なくとも 4 単語程度必要なため，単語がスパースだとツイートの類似度の精度は悪くなる．

学習によって得られる分散表現とは対照的に，著者らは過去に人手構築による単語意味ベクトルとそのブートストラップ学習を提案し，実用システムに適用した [5], [6]．単語意味ベクトルは，単語と特徴単語との関係を，関係あり，関係なし，の 2 値で表現する．特徴単語とは，単語意味ベクトルの各次元に対応するもので，266 種類の概念分類から成る．基本単語とは，新聞記事や百科事典の頻度解析により選択した約 2 万語の重要単語から成る．単語意味ベクトル辞書とは，基本単語に関係のある特徴単語を列挙した辞書である．この論文での考え方をパラグラフベクトルに応用することにより，単語のスパース性についての課題を解決することができると考えた．単語意味ベクトル辞書を用いて，各単語に関係のある特徴単語を単語拡張した文書をパラグラフベクトルのモデルで学習するという，単語意味ベクトルとパラグラフベクトルの統合化手法を提案する．単語意味ベクトル辞書を用いた単語拡張により，文長が短いツイートに文脈情報を追加することができる．

単語意味ベクトルとパラグラフベクトルを統合するとなぜ極性分析の精度が向上するかを具体的に説明する．ツイートの学習，及び単語拡張された文脈情報の学習にパラグラフベクトルを用いる．ツイート「新しい製品 B を推す」では，ツイートのパラグラフベクトルと「新しい，製品 B，推す」の 3 単語のベクトルを，語順の情報を利用するパラグラフベクトルの PV-DM モデルと，語順の情報を利用しないパラグラフベクトルの PV-DBOW モデルを用いて学習する．パラグラフベクトルの 2 種類のモデルに関しては，**3.2** で述べる．単語意味ベクトル辞書を用いた単語拡張では，基本単語「新しい，推す」に対して，共通の特徴単語「流行・人気，価値・質，優良，肯定的，強力」が文脈情報としてツイートに追加される．他に基本単語個別の特徴単語がツイートに追加される．このように論理的関連性と連想的関連性を考慮できることが，単語意味ベクトルとパラグラフベクトルを統合する利点である．文脈情報は PV-DBOW モデルを用いて単語ベクトルを更新し，ツイートのパラグラフベクトルに類似したベクトルを学習する．本ツイートがポジティブのラベルが付与されたツイートだとすると，後段の教師あり学習により，単語「新しい，推す」がツイートに出現すると，このツイートがポジティブに分類されるように分類器は学習される．これにより，もともとのツイートには存在しなかった特徴単語もポジティブに分類される単語ベクトルに更新されていく．例えば，ツイート「流石に製品 B は捗る」の教師データがない場合は，基本単語「流石，捗る」両方とも特徴単語「優良，肯定的，強力」に単語拡張されるため，提案手法ではポジティブに分類される．特徴単語を文脈情報として付加することにより，ツイートをパラグラフベクトルで学習する場合と比較して，大規模なツイートを収集しなくても単語のスパース性を改善できる．

本論文では，実データで提案手法の有効性を検証するため，クラウドソーシングを利用して Twitter からの評判情報抽出ベンチマークを作成する．ツイートの学習と単語拡張した文脈情報の学習を統合した提案手法により，ツイートのパラグラフベクトルと比較して，極性分析精度を改善できるか確かめるための評価実験を行う．本論文の貢献は以下の 2 点である．

- パラグラフベクトルの学習において，単語のスパース性を解決するために著者らが構築した単語意味ベクトル辞書の導入手法を提案．
- 単語意味ベクトルとパラグラフベクトルの統合により，単語がスパースでも特定分野の文脈情報を学習し，極性分析の精度が向上．

2. 関連研究

本章では，Twitter を対象とした極性分析のシェアードタスクにおいて，シソーラスによる素性拡張を用いた手法及び最高水準を示した手法と提案手法との比較を行う．

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 を対象に極性分析のコンテスト SemEval を開催している [7]．SemEval での評価尺度は，ポジティブ予測 F 値の F_{pos} とネガティブ予測 F 値の F_{neg} のマクロ平均 F 値 $= (F_{pos} + F_{neg})/2$ が採用されている．したがって，ポジティブとネガティブのツイート数の偏りは考慮されず，ニュートラル予測に関しては間接的に評価される．本論文でも同じ評価尺度を採用した．以降，ポジティブ予測とネガティブ予測のマクロ平均 F 値を F 値と呼ぶ．

TwitterHawk [8] は，シソーラスの一種である WordNet(注1) による素性拡張の実験を行った．N-gram,

(注1)：http://wordnet.princeton.edu/

極性辞書に加えて，WordNet から素性を拡張し，リニアカーネル SVM を用いて極性分析を行った．しかし，WordNet を用いない場合と比較して F 値が 2 ポイント程度下がった．WordNet は 1 次元的に単語の概念を分類したものであり，単語に 11 万 7 千の同義語グループとの関係が付与されている．素性拡張に利用した場合にはツイートの文脈とは関係のない多くの同義語に拡張される可能性があり，これが極性分析の精度を落とした原因と考えられる．これに対して，提案手法の単語意味ベクトルは，266 種類の特徴単語との意味的な関係を 266 次元でベクトル表現したものである．TwitterHawk では拡張された素性ベクトルをリニア SVM に直接掛けているが，提案手法では単語拡張された特徴単語も文脈情報を学習するように単語ベクトルを更新している．これにより，提案手法では単語拡張を行わない場合と比較して F 値が 3 ポイント程度向上した．

SemEval-2013 で最高水準を示した NRC-Canada [9] は，N-gram，極性辞書（3 種類の人手構築による辞書，2 種類の Twitter からの自動生成辞書）など 100 万以上の素性に対して，リニアカーネル SVM を用いて分類器を構築した．結果，F 値は 69.02 となり，775,000 ツイートから自動生成した極性辞書により，5 ポイント程度 F 値が向上することを示した．

この種の Feature Engineering を行わない word2vec に類似した Structured Skip-gram モデルによる単語ベクトルの学習を用いた INESC-ID [10] では，5,200 万（21 万語彙）のラベル無しツイートを対象に 600 次元の単語ベクトルを学習させた．分類器も同じニューラルネットワークを用いて，教師データを元に 600 次元から 10 次元の部分空間への写像を誤差伝搬で学習させた．SemEval-2013 のテストセットは 72.09，SemEval-2015 は 65.21 と最高水準を示した．

NRC-Canada は，100 万次元以上の素性ベクトルを用いており，INESC-ID は 5,200 万ツイートで単語ベクトル辞書の初期学習を行った．これらは，評判情報抽出の実利用を考えると，経済合理性の観点から非現実的と言える．極性分析の精度を高めるために大規模な素性やツイートが必要な理由は，Twitter における単語のスパース性が原因である．提案手法では，NRC-CANADA の 0.1%程度の素性，INESC-ID の 1%程度のラベル無しツイートを用いて，同レベルの F 値を示した．これは単語意味ベクトルとパラグラフベクトルの統合により，単語がスパースでも特定分野の文脈情報を学習した効果と考えられる．

3. 提案手法

3.1 Twitter からの評判情報抽出手法（提案法）

本節では，商品開発や品質サポートに役立つ評判情報抽出手法を提案する．

本論文では，単語意味ベクトル辞書を用いて単語拡張した文書をパラグラフベクトルのモデルで学習する統合化手法を提案する．ツイート中から，単語意味ベクトル辞書に登録されている基本単語を特徴単語に展開する手法により，文長が短い Twitter では適切に捉えることが難しいパラグラフベクトルの学習を目指す．

図 1 に提案法の流れを示す．実線の矢印は訓練セットの流れを示す．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る．二重線の矢印は開発セット（ツイート・ラベル）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調整を示す．具体的には，ベンチマークの訓練セットと開発セットを準備した上で，以下の手順で提案法を構築する．

【Step 1】 訓練セットのラベル無しツイートを対象に **3.2** 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）を用いて単語ベクトルを学習する．同時に **3.3** で述べる単語意味ベクトル辞書を用いて，ラベル無しツイート中の基本単語を特徴単語に展開し，PV-DBOW モデルを用いて単語ベクトルを学習する．本手順（Step 1）は省略可能である．

【Step 2】 パラグラフベクトルの学習において，Step 1 で学習した単語ベクトルを単語辞書の初期値として読み込む．Step 1 を省略した場合は単語辞書の初期値はランダム値とする．次に訓練セットの “ラ

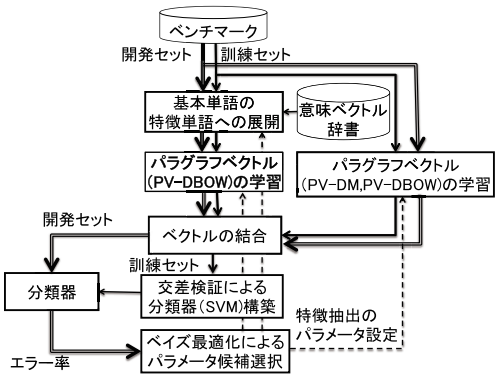


図 1 提案法の流れ
Fig. 1 Overall flow of the proposed method.

ベルあり” ツイートを対象としたパラグラフベクトル（PV-DM，PV-DBOW）の学習，及び “ラベルあり” ツイート中の基本単語を特徴単語に展開したパラグラフベクトル（PV-DBOW）を学習する．ただし，特徴単語に展開した基本単語は学習対象として残す．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，各ツイートのラベルを教師データとして，サポートベクターマシン（SVM）により，ツイートの分類器を学習する．4 分割交差検証とグリッドサーチにより，SVM のパラメータを決定し，分類器を構築する．なお，訓練セットの学習時には PV-DBOW モデルで特徴単語のベクトルも更新する．

【Step 3】 開発セットのツイートを対象に Step 2 で訓練セットを学習した単語ベクトル辞書を元に 3 種類のパラグラフベクトルを学習する．開発セットのパラグラフベクトルとラベルを Step 2 で作成した分類器に与えて，エラー率 $[= 100-(F_{pos}+F_{neg})/2]$ を測定する．エラー率を目的関数として，目的関数の出力を最小化するようにガウス過程を用いたベイズ最適化によりパラグラフベクトル学習のパラメータを自動的に調整する [11]．

エラー率が収束するまで Step 2 と Step 3 を繰り返した後，パラグラフベクトル学習のパラメータの値を決定する．

3.2 パラグラフベクトル

パラグラフベクトルの二つのモデルを図 2 に示す [4]．PV-DM（paragraph vector with distributed memory）モデルは，ウインドウ内の周辺単語の文脈語ベクトル（入力層と中間層の間の重み）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) の対象語ベクトル（中間層と出力層の間の重み）をニューラルネットワークにより予測する．中間層のノード数がベクトル次元数に対応する．PV-DM は，文脈ベクトルと次単語 w(t) の対象語ベクトルとの内積が，周辺単語以外に出現する単語の対象語ベクトルとの内積より大きくなるように次単語 w(t) の対象語ベクトルを予測する．ウインドウをシフトしながら予測する単語を変えていくが，常にパラグラフベクトルを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，周辺単語をウインドウ長分選択し，ランダムに選択した周辺単語の文脈語ベクトル（中間層と出力層の間の重み）を予測するようにパラグラフベクトルを学習する．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できる．PV-DBOW モデルにおいて単語ベクトルを学習する場合は，word2vec の Skip-gram モデルが用いられる [2]．対象単語 w(t) の対象語ベクトル（入力層と中間層の間の重み）とランダムに選択した周辺単語の文脈語ベクトル（中間層と出力層の間の重み）との内積が，周辺単語以外の文脈語ベクトルとの内積より大きくなるように対象語ベクトルを学習する．

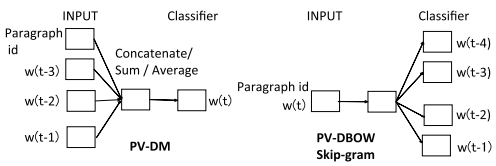


図 2 パラグラフベクトルの 2 種類のモデル
Fig. 2 Two types of models for learning paragraph vectors.

PV-DM や PV-DBOW は，単語ベクトルやパラグラフ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対象語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 は中間層で周辺単語の文脈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．

提案手法におけるツイート中の基本単語の特徴単語への展開は，単語拡張により語順の概念がなくなるため，PV-DM モデルとは合わない．一方で，単語拡張によりツイートの文脈情報が追加されるため，対象単語からランダムに周辺単語の文脈語ベクトルを予測する PV-DBOW モデルとは良く合う．このため，提案手法では単語拡張したツイートの学習に PV-DBOW モデルを用いる．

パラグラフベクトルは，後段の SVM 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，学習回数など）を調整する必要がある．この問題に対して本研究では，ベイズ最適化を用いてパラメータを探索することにより対応した．

3.3 基本単語の意味ベクトル

基本単語の意味ベクトル [5] 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，各次元が一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 $\mathrm{X} = (x_1, \cdots, x_n)$ 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」は特徴単語「人間，科学，興奮」と関係があるので，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単語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 に示すとおり，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．基本単語は，以下の 4 種類の選択基準から，2 万 336 語を選択した．

- 百科事典(注2)，新聞記事(注3)の説明に使われる現代用語
- WWW ホームページなどの分類用語
- 取扱説明書で使われる操作用語
- 形容詞などの感性語

辞書編纂の専門家が，以下の特徴単語の付与基準に基づいて，各基本単語に特徴単語を付与した．

- 論理的関連と連想的関連から，特徴単語を付与する．論理的関連は，基本単語に対して特徴単語が表 2 に示すような直接的関連性を有するものを指す．連想的関連は，基本単語に対して特徴単語が感覚的関連性，連想により想起される関連性を有するものを指す．例を表 3 に示す．

表 1　特徴単語の分類
Table 1　Classification of feature words.

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

(注2)：ブリタニカ小項目事典 CD 版，TBS ブリタニカ，1992.
(注3)：CD-毎日新聞'94，'95 データ集，毎日新聞社，1994，1995.

- 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」に付与する．例を表 4 に示す．

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．このベクトルに関して，下記の仮説を置く．

- 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される．
- 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，その単語が使われる適切な文脈情報を獲得できる．

3.4　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

表 2　論理的関連による特徴単語の付与基準
Table 2　Criteria for manual entries of feature words for core words by logical relationship.

論理的関係	基本単語	特徴単語
集合包含	秋	季節
同義関係	アイデア	思想
部分全体関係	足	人間の身体

表 3　連想的関連による特徴単語の付与基準
Table 3　Criteria for manual entries of feature words for core words by associative relationship.

基本単語	特徴単語
愛	優しさ，温かさ
アップ	経済，映像
足	自動車，交通輸送

表 4　基本単語に付与された特徴単語の例
Table 4　Examples of feature words given for core words.

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
爽やか	環境，感覚，様子・様態，優良，肯定的，新しさ
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

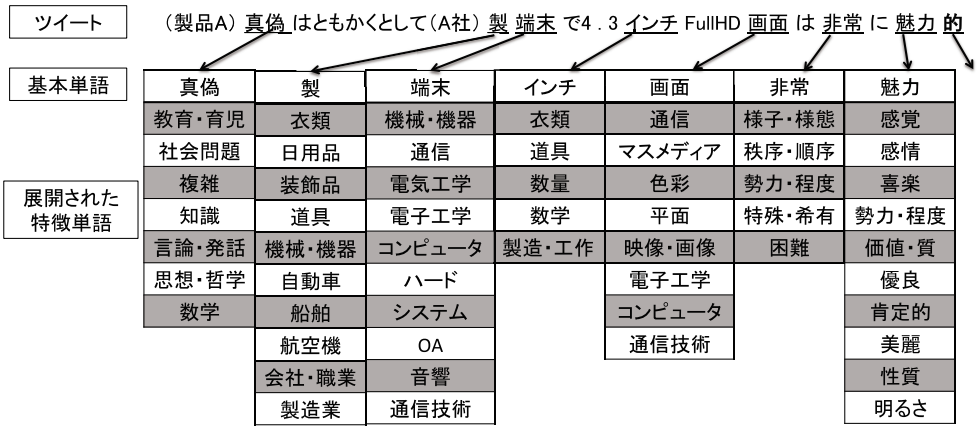


図 3　特徴単語の展開例
Fig. 3　An example of extracted core words from a tweet and extended feature words from the core words.

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．図 3 で示したツイートでは，「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）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，及び仮説 2 の単語の意味ベクトルの更新には 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グラフベクトル学習のパラメータとして，後段の教師あり学習によるエラー率を考慮して決定する．

4.　Twitter からの評判情報抽出の実験

本章では，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．

4.1　手　　順

Twitter からの評判情抽出の実験手順を図 4 に示す．本論文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．

① 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．

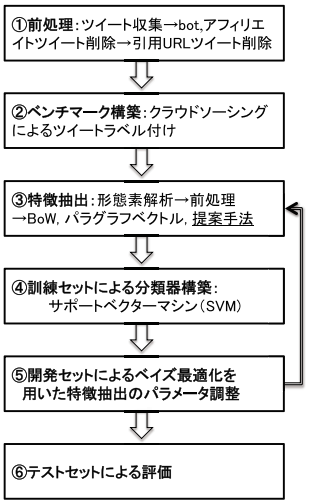


図 4　Twitter からの評判情報抽出の実験手順
Fig. 4　Experimental procedure of the extraction of the reputation information from Twitter.

② 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．

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．

③ 特徴抽出では，パラグラフベクトル，提案手法の単語意味ベクトル辞書による単語拡張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，あらかじめ以下の前処理を行った．

- ユーザ名（@user），改行の削除
- 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(注4)，及び Web 上の言語資源から数百万の新語や固有表現を拡充した MeCab 用辞書 mecab-ipadic-NEologd(注5) を用いた．

④ 訓練セットによる分類器構築では，SVM を用いて交差検証を行い，SVM の各種パラメータを決定した．実験方法については，**4.3** で述べる．

⑤ 開発セットによるベイズ最適化を用いた特徴抽出のパラメータ調整では，開発セットのエラー率を最小化するように特徴抽出のパラメータを調整し，③の特徴抽出にフィードバックする．調整するパラメータは，**3.2**，**3.4** で述べたウインドウ長，ベクトル長，学習回数，展開する特徴単語数の上限などである．開発セットのエラー率が収束した段階でフィードバックループを止める．なお，訓練セットの学習時は PV-DBOW の単語ベクトルを更新したが，開発セットではパラグラフベクトルのみ学習し，ツイート数が少ないため，PV-DBOW の単語ベクトルは更新しなかった．

⑥ テストセットによる評価では，⑤の開発セットを元に最適化した特徴抽出のパラメータを用いて，テストセットの 3 クラス分類により F 値を評価する．なお，テストセットの評価では，開発セットを訓練セットには含めず，PV-DBOW の単語ベクトルを更新しない場合と更新する場合の 2 種類の評価を行った．

(注4)：http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
(注5)：https://github.com/neologd/mecab-ipadic-neologd

4.2 ベンチマーク

製品 A と製品 B に関して，2014 年 10 月～2015 年 11 月の 13ヶ月分のツイート全数を検索式で収集した．製品 B に関しては製品名で収集を行ったが，製品 A に関しては同じ製品ブランド名を冠した製品ジャンルが他にあるためスマートフォンのトピックとの AND 検索を行った．10%程度にサンプリングした後，カタカナ列のみのツイートやボットなどを削除し，約 35,000 ツイートに対して，クラウドソーシングを利用して各ツイートに 5 人の作業者を割り当ててラベルを付与した．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．

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 5 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：3,210 件，開発セット：802 件，テストセット：802 件），製品 B のベンチマークは 11,774 ツイート（訓練セット：8,831 件，開発セット：1,471 件，テストセット：1,472 件）である．前処理を行った後のベンチマークの統計情報を表 6 に示す．製品 A が平均 14.5 単語/ツイート，製品 B が平均 10.0 単語/ツイート，ツイートから抽出された基本単語数は製品 A が平均 5.4 単語/ツイート，製品 B が平均 3.3 単語/

表 5 ベンチマークの構成
Table 5 Configuration of the benchmark.

Dataset	Positive	Negative	Neutral	合計
製品 A				
訓練	1,122 (35%)	1,023 (32%)	1,065 (33%)	3,210
開発	280 (35%)	256 (32%)	266 (33%)	802
テスト	280 (35%)	256 (32%)	266 (33%)	802
製品 B				
訓練	3,654 (41%)	2,375 (27%)	2,802 (32%)	8,831
開発	608 (41%)	396 (27%)	467 (32%)	1,471
テスト	609 (41%)	396 (27%)	467 (32%)	1,472
ラベル無し				
560,853				

表 6 各ベンチマークの統計情報
Table 6 Statistical information of each benchmark.

	製品 A	製品 B
総ツイート数	4,814	11,774
総語彙数	8,782	11,901
出現頻度 1 の割合	54.8%	54.3%
平均単語数/ツイート	14.5	10.0
基本単語含有ツイート数	4,606	10,352
平均基本単語数/ツイート	5.4	3.3
平均特徴単語種類/ツイート	26.3	19.1

ツイートである．これは，製品 A のベンチマークは製品 B に対してツイート数は 40%程度と少ないが，文長は長いことを示す．

クラウドソーシングにより，複数のラベルが 1 位で同数投票されたツイートや “無関係” のラベルが付与されたツイートに関しては，今回の実験対象から外した．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 56 万件ある．これらは製品 A，B に関して収集したツイートから前処理を行い，ベンチマークとの重複を除去したものであり，単語のスパース性を解消する目的で単語ベクトルの初期学習に用いた．

4.3 実験方法

（1） ベースライン（BoW）の評価実験

ツイートの特徴表現を製品ごとの訓練セットと開発セット，及び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(注6) を用いた．訓練セットの BoW に対しての 4 分割交差検証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，開発セット，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

最初に約 56 万件の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．最後に開発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(注7) を用いた．**3.2** 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適切な値の範囲を決めておき，ベイズ最適化ツール spearmint(注8) を用いてパラグラフベクトル学習のためのパラメータを決定した．spearmint は可能な限り少ない繰り返し回数により，目的関数の値が最小となるように複数のパラメータを調整して自動実験を行うように設計されている [11]．ここでの目的関数は開発セットのエラー率とした．

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 分割交差検証とグリッドサーチを元に構築した．ただし，訓練セットによる交差検証時は単語ベクトルも学習しているため，開発セットによる評価とは実験設定が異なる．構築した SVM の分類器と開発セットのパラグラフベクトルを用いて開発セットのエラー率を測定し，spearmint により次のパラグラフベクトル学習のパラメータを選択し，開発セットのエラー率が収束したと判断するまで自動実験を繰り返す．エラー率の収束条件は，100 回以上自動実験を繰り返しており，かつ連続 20 回以上，自動実験におけるエラー率の測定において，最小値が更新されなかった場合とした．

（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 **3.4** 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．各基本単語には平均 9 語の特徴単語が表 1 に示した分類順に付与されており，特徴単語の展開候補が 7 語より多い場合は，先頭から順に 7 語選択した．この理由は，製品の評判に関するツイートでは，大分類「人間環境」,「抽象概念」,「物理・物質」,「文明・知識」に分類される特徴単語の出現確率が高いが，このうち「人間環境」と特徴単語の並び順からは中央に位置する「抽象概念」に分類される特徴単語が，極性判定において特に重要と考えたためである．ラベル無しツイートは，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 10 万 9 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．ラベル無しツイートの学習に関しては，特徴単語の展開数などのパラメータは製品 A と製品 B の訓練セット，開発セットを用いた予備実験により決定した．各パラメータはデフォルト値を採用し，ラベル無しツイートの基本単語の名詞を特徴単語展開するかどうか，及び展開する特徴単語数の上限を 4 語，7 語，14 語と振り，開発セットの F 値が最も良くなる値を採用した．製品ごとの訓練セット，開発セットについては，基本単語を特徴単語に展開し，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．パラグラフベクトル学習に必要なパラメータ調整と同様に基本単語の特徴単語展開の上限もパラメータとして spearmint を用いて決定した．SVM の分類器の構

(注6)：http://scikit-learn.org/stable/
(注7)：http://radimrehurek.com/gensim/
(注8)：https://github.com/HIPS/Spearmint

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．

4.4 結　　果

3 クラス分類における F 値とエラー率を表 7 に示す．同じ条件設定でも，パラグラフベクトルは学習のたびにツイートをベンチマークからランダムに取得するため，異なる特徴ベクトルが生成される．したがって，SVM の分類精度が変動するため，開発セットに関しては 5 回試行の平均を，テストセットに関しては 10 回試行の平均を求めた．また，テストセットの学習において PV-DBOW の単語ベクトルを更新しない場合（表 7 のテスト列）と PV-DBOW の単語ベクトルを更新する場合（表 7 のテスト（適応）列）の 2 種類の評価を行った．

開発セットの評価では，提案手法は，製品 A：2.5 ポイント，製品 B：4.8 ポイント，パラグラフベクトル（以降，PVEC と呼ぶ）の F 値を上回った．エラー率の相対値では，提案手法により，製品 A：6.9% (= 2.5/36.1)，製品 B：14.7%改善した．この評価結果は，表 6 に示すとおり，製品 B のベンチマークは製品 A と比較して単語のスパース性が高いため，提案手法による単語拡張が製品 B のエラー率改善により効果があったと考えられる．

テストセットの評価結果は，PVEC，提案手法共に単語ベクトルを更新した場合（テスト（適応）列）の F 値が単語ベクトルを更新しなかった場合（テスト列）と比較して，同等あるいは向上している．テストセットの単語ベクトルを更新した場合に提案手法は製品 A：2.0 ポイント，製品 B：3.2 ポイント，PVEC の F 値を上回った．エラー率の相対値では，提案手法により，製品 A：5.8%，製品 B：10.2%改善した．提案手法によるテストセットの改善率は両製品共に開発セットに対して下がっているが，製品 B の下落率が大きい．この理由については，次節で考察する．

テストセットの単語ベクトル更新は，単語ベクトルを更新しない場合と比較して，製品 B の提案手法において F 値 1 ポイント（エラー率の相対値 3.4%）改善しており，最も効果があった．この理由は，製品 B は単語拡張の効果が大きいため，開発セットの F 値 72.1 と高くなり，過学習していることが考えられる．特徴単語は 266 種類と限定されているため，より少ないツイート数でもテストセットの文脈情報に適応させることにより，過学習の問題は解消されることを確認した．

一方で，語順を保持しないベースラインの BoW に対する従来手法 PVEC の改善率は，製品 A では開発セット：4.6 ポイント（エラー率の相対値 11.3%），テストセット：2.8 ポイント（同相対値 7.5%），製品 B では開発セット：0.6 ポイント（同相対値 1.8%），テストセット：0.5 ポイント（同相対値 1.6%）である．製品 A では PVEC が BoW に対して大きく改善しており語順を保持する効果は確認できるが，製品 B では PVEC は BoW に対して有意差を示したが，その差は小さい．この評価結果は，表 6 に示すとおり，製品 A のベンチマークは製品 B と比較して，1 ツイートあたりの平均単語数 45%増であり，ツイートの文長が長いためと考えられる．

表 8 に提案手法，PVEC における特徴抽出のパラメータ (Parameters)，振れ幅 (Min-Max)，及び製品 A，製品 B で採用したパラメータ設定を示す．「単語頻度しきい値」はツイートの低頻度語を語彙辞書から削除するしきい値，「高頻度語削減しきい値」はツイートの高頻度語を確率的に削減するしきい値を示す．これらのしきい値は，ラベル無しツイートの学習時に語彙辞書を作成しており，その設定が使われる．PV-DM の中間層を入力ベクトルの和とするか結合とするか

表 7　評価結果 1：ポジティブ・ネガティブ予測のマクロ平均 F 値（エラー率）
Table 7　Evaluation results1: macro-averaged F score for predicting positive and negative tweets in the dev. set and the test set (the error rates).

	製品 A			製品 B		
	開発	テスト	テスト（適応）	開発	テスト	テスト（適応）
BoW	59.3 (40.7)	62.9 (37.1)		66.7 (33.3)	68.2 (31.8)	
PVEC	63.9 (36.1)	65.4 (34.6)	65.7* (34.3)	67.3 (32.7)	68.3 (31.7)	68.7** (31.3)
提案手法	66.4 (33.6)	67.7 (32.3)	67.7*** (32.3)	72.1 (27.9)	70.9 (29.1)	71.9**** (28.1)

vs. BoW　　*p = 3.9e-12 < 0.05　　**p = 0.004 < 0.05
vs. PVEC　　***p = 4.1e-07 < 0.05　　****p = 5.6e-13 < 0.05

表 8　特徴抽出のパラメータ・振れ幅と製品 A，B の値
Table 8　Parameters of feature extraction and their values for the product A and the product B.

Parameters	Min-Max	製品 A		製品 B	
		提案手法	PVEC	提案手法	PVEC
訓練セット学習回数 (学習係数)	5〜15 (0.007〜0.025)	12 (0.007)	5 (0.024)	15 (0.007)	15 (0.007)
テストセット学習回数 (学習係数)	5〜15 (0.007〜0.025)	5 (0.018)	15 (0.021)	15 (0.010)	5 (0.014)
ベクトル次元数	100〜400	142	400	400	400
特徴単語展開の上限	7〜17	17		7	
名詞の特徴単語展開	0: No, 1: Yes	1		0	
PV-DBOW (提案手法) ウインドウ	6〜18	6		6	
PV-DM 中間層	和	和	和	和	和
PV-DM ウインドウ	4〜12	4	4	12	5
PV-DBOW ウインドウ	5〜15	5	5	5	5
単語頻度しきい値	2	2	2	2	2
高頻度語削減しきい値	1e-03	1e-03	1e-03	1e-03	1e-03
Negative Sampling	0〜10	10	10	10	10

表 9　提案手法による改善，失敗したツイート数（製品 A）
Table 9　The number of succeeded tweets and failed tweets by the proposed method in the product A.

	開発セット		テストセット（適応）	
	不正解 → 正解	正解 → 不正解	不正解 → 正解	正解 → 不正解
Positive	28	20	31	13
Negative	10	12	13	24
Neutral	41	23	31	17

表 10　提案手法による改善，失敗したツイート数（製品 B）
Table 10　The number of succeeded tweets and failed tweets by the proposed method in the product B.

	開発セット		テストセット（適応）	
	不正解 → 正解	正解 → 不正解	不正解 → 正解	正解 → 不正解
Positive	126	99	82	120
Negative	109	82	112	72
Neutral	125	112	134	90

に関してもラベル無しツイート学習時の設定（入力ベクトルの和）を変えることができない．「Negative Sampling」は，ツイート内の周辺単語に対して負例をサンプリングする語数である．固定値以外のパラメータについては，ベイズ最適化における局所最適の問題を避けるため，あらかじめデフォルト値（ツールや論文の推奨値）に固定し，各パラメータを順に振って，開発セットのエラー率を元にパラメータの許容範囲を調査した．この予備実験を元に各パラメータの振れ幅 (Min-Max) を決めた．

4.5 考　　察

提案手法による単語拡張の効果について検証する．PVEC に対して，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を表 9（製品 A），表 10（製品 B）に示す．提案手法は，製品 A では 16%程度のツイートの極性分析に影響を与えたが，製品 B では 45%程度のツイートの極性分析に影響を与えた．これは表 6 に示したとおり，製品 A は製品 B と比較して，1 ツイートの平均単語数が 4.5 単語多いため，単語拡張の影響を受け難いためと考えられる．また，製品 A では，開発セットとテストセットの単語拡張がポジティブとニュートラルの分類に効果があることを示す．製品 A において，ポジティブ，ネガティブを想起する出現頻度の高い特徴単語と付与されたツイート数を以下に示す．

- ポジティブ：肯定的：1794，新しさ：1607，優良：1489，流行・人気：929，明るさ：624，美麗：356
- ネガティブ：否定的：1088，困難：590，劣悪：411

ネガティブを想起する特徴単語が付与されたツイート数がポジティブと比較して少ないことが，ネガティブの分類において，単語拡張の効果が小さい要因と考えられる．一方，製品 B においては，開発セットでは，ポジティブ，ネガティブ，ニュートラルいずれも改善した．しかし，テストセットでは，ネガティブとニュートラルは更に改善したが，ポジティブは単語拡張による悪影響が出た．この理由を調べるため，製品 B のポジティブとネガティブにおいて，提案手法により改善したツイート群（不正解→正解），失敗したツイート群（正解→不正解）における典型的な特徴単語を開発セット，テストセットごとに表 11 に示す．各ツイート群において，各特徴単語の出現頻度を各群の

表 11　製品 B の提案手法による改善，失敗したツイート群における典型的な特徴単語（対群に対する出現比率）
Table 11　The typical feature words in the succeeded tweets and the failed tweets by the proposed method in the product B (their appearance ratio to their pair group).

	開発セット		テストセット（適応）	
	不正解 → 正解	正解 → 不正解	不正解 → 正解	正解 → 不正解
Positive	感情 (2.0)	勢力・程度 (2.7)	肯定的 (2.2)	機械・機器 (4.4)
	肯定的 (2.2)	否定的 (2.4)	数量 (2.1)	日用品 (3.3)
	感情的 (2.6)		経済 (2.3)	道具 (3.1)
	道徳・倫理 (3.6)		安価 (2.4)	施設・設備 (5.0)
	強力 (2.5)		税制 (2.4)	否定的 (3.2)
	流行・人気 (2.1)		流行・人気 (3.9)	複雑 (4.4)
Negative	否定的 (1.6)	数量 (2.9)	否定的 (6.3)	変化 (2.2)
	機械・機器 (2.0)	肯定的 (3.7)	性質 (3.1)	新しさ (14)
	施設・設備 (2.0)	経済 (3.0)	秩序・順序 (2.6)	肯定的 (1.6)
	劣悪 (2.4)	安価 (3.0)	劣悪 (4.8)	
	運動 (2.3)	税制 (3.0)		
	病気 (2.3)	道徳・倫理 (4.0)		

ツイート数で正規化し，対群に対しての出現比率が高い特徴単語を抽出し，各群における頻度順に並べた．

以下に表 11 の群ごとに対群に対する典型的な特徴単語を分析し，従来手法の不正解が提案手法により正解に変わった理由，及び従来手法の正解が提案手法により不正解に変わった理由について分析する．

- **ポジティブ：開発セット（不正解→正解）**：典型的な特徴単語は，「良い，綺麗，好き，新しい，推す，流石，捗る」などの単語拡張であり，これらの特徴単語がポジティブの極性判定に寄与した．
- **ポジティブ：開発セット（正解→不正解）**：典型的な特徴単語は，「無駄に，間違い，無理，迷う，やばい，傾く」などの単語拡張であり，否定形，他製品の話題など，前後の文脈を考慮しないとポジティブとの極性判定が困難なツイートが多い．
- **ポジティブ：テストセット（不正解→正解）**：「買う」（典型的な特徴単語の「数量，経済，安価，税制」）に関連したツイートが多く，他の典型的な特徴単語は，「綺麗，安心，お洒落，便利，羨ましい」などの単語拡張であり，ツイートの傾向は開発セット（不正解→正解）と類似している．
- **ポジティブ：テストセット（正解→不正解）**：「使う」（典型的な特徴単語の「機械・機器，日用品，道具，施設・設備」）に関連したツイートが多く，他の典型的な特徴単語は，「難い，嫌い，不便，残念」などの単語拡張であり，ツイートの傾向は開発セット（正解→不正解）と類似している．
- **ネガティブ：開発セット（不正解→正解）**：「使う」に関連したツイートが多く，他の典型的な特徴単語の「病気」は「危うい，酷い，不調，異常，弱い」などの単語拡張，「運動」は「割れる，繰り返し，急速」などの単語拡張であり．これらの特徴単語がネガティブの極性判定に寄与した．
- **ネガティブ：開発セット（正解→不正解）**：「買う」に関連したツイートが多く，「肯定的，道徳・倫理」は「良い」の単語拡張であり，否定形あるいは他製品の話題など，前後の文脈を考慮しないとネガティブとの極性判定が困難なツイートが多い．
- **ネガティブ：テストセット（不正解→正解）**：不具合に関連したツイートが多く，典型的な特徴単語の「否定的，秩序・順序，性質」は「勝手」の単語拡張であり，これらの特徴単語がネガティブとの極性判定に寄与した．
- **ネガティブ：テストセット（正解→不正解）**：機種変更に関連したツイートが多く，典型的な特徴単語は「変える，新しい」などの単語拡張であり，ツイートの傾向は開発セット（正解→不正解）と類似している．

分析の結果，開発セット，テストセット，ポジティブ，ネガティブにかかわらず，PVEC の極性判定誤りを提案手法により正しく極性判定できたツイートは，製品 B に「満足」，製品 B を「買う」，製品 B を「使う」，製品 B の「不具合」の 4 種類のケースである．いずれのケースにおいても単語拡張により，ツイートの文脈情報を補い，極性判定に成功したことが分かった．

一方，PVEC で正しく極性判定できたものが提案手法により極性判定を誤ったツイートは，製品 B に「不満足」，製品 B を「使う」，製品 B を「買う」，製品 B から「変える」の 4 種類のケースである．いずれのケースも極性に影響を与える単語の否定形や他製品についての話題であり，語順を考慮しない PV-DBOW では対処が困難なケースが多いことが分かった．これは，PVEC と比較すると，提案手法では PV-DM の効果が相対的に弱まったことが原因と考えられる．

4.6　提案手法の有効性について

本論文では，約 1 万 2 千ツイートから構成される製品 B のベンチマーク，及び総ツイート数は製品 B と比べて半分以下だが，1 ツイートの平均単語数が約 45%多い製品 A のベンチマークを構築し，提案手法の有効性を示した．しかし，製品数が 2 種類では評価は十分とは言えない．そこで，多様な製品について，

表 12　実験的ベンチマークの構成
Table 12　Configuration of the experimental benchmark.

製品	訓練セット			テストセット		
	Positive	Negative	Neutral	Positive	Negative	Neutral
スマートフォン						
製品 A	7	9	8	2	2	2
製品 B	5	5	6	2	2	2
製品 C	1	5	2	0	1	0
その他	0	0	9	0	0	3
ロボット掃除機						
ロボット A	6	2	1	1	1	1
ロボット B	4	3	3	2	0	0
コンビニプリントサービス						
サービス A	2	2	1	1	1	0
サービス B	1	5	2	0	1	0
メーカー						
メーカー A	8	5	1	2	2	0
メーカー B	3	3	2	1	1	0
合計	37	39	35	11	11	8

提案手法の有効性評価を目的に小規模なベンチマークを作成した．ベンチマークの構成を表 12 に示す．スマートフォン，ロボット掃除機，コンビニプリントサービス，メーカーのカテゴリー分類の元，各カテゴリー 2〜4 種類の製品から構成される．スマートフォンの「その他」は，製品 A〜C を含め，複数のスマートフォン製品ブランドについて言及したツイートであり，全てニュートラルのラベルが付与された．メーカーは，メーカー A，B に関しての評判や両社の製品に関するツイートである．ベンチマーク全体で 141 件（訓練セット 111 件，テストセット 30 件）である．

評価は，ラベル無しツイートを用いず，単語ベクトルの初期値をランダム値として，ツイート中の基本単語を特徴単語展開を行って PV-DBOW モデルを用いて学習した提案手法のパラグラフベクトルとツイートのみを PV-DBOW モデル用いて学習した従来手法のパラグラフベクトルを対象に行う．すなわち，提案手法による単語拡張した文脈情報の学習効果を評価することを目的に PV-DBOW（提案手法）と PV-DBOW (従来手法) のベクトルのみを特徴表現として用いて，訓練セットを対象に leave-one-out 交差検証により，SVM 分類器を構築する．訓練セットを基に決定した特徴抽出のパラメータ設定を表 13 に示す．「高頻度語削減しきい値」の設定により，従来手法では高頻度語の削減はなく，提案手法はほぼ全ての単語が削減対象となる．このため，提案手法の学習回数が非常に多いが，毎回異なる単語の組み合わせを学習していることになる．

評価結果を表 14 に示す．これまでの評価結果における 3 クラス分類の F 値に加えて，2 クラス分類の F 値も評価した．これはテストセットのニュートラルがスマートフォンに偏っているため，ポジティブとネガティブの 2 クラスのみの訓練セット，テストセットにより，評価を行った．3 クラス分類 F 値，2 クラス分類 F 値ともに 9〜10 ポイント提案手法が従来手法を上回った．標準偏差が大きいため，30 回試行の平均を採用した．以下に従来手法では極性判定を誤り，提案

表 13　特徴抽出のパラメータと提案手法，従来手法の値
Table 13　Parameters of feature extraction and their values for the proposed method and the conventional method.

Parameters	提案手法	従来手法
訓練セット学習回数	6000	600
テストセット学習回数	1500	300
ベクトル次元数	100	100
特徴単語展開の上限	7	
名詞の特徴単語展開	1	
ウインドウ長	30	15
単語頻度しきい値	2	2
高頻度語削減しきい値	1e-06	0.1
Negative Sampling	10	10

表 14　評価結果 2:PV-DBOW（従来手法）と PV-DBOW（提案手法）における 3 クラス分類と 2 クラス分類の F 値（標準偏差）
Table 14　Evaluation results2: F scores(SD) in the three class and the two class classification by PV-DBOW (the conventional method) and PV-DBOW (the proposed method).

	3class 分類 F 値 (SD)	2class 分類 F 値 (SD)
従来手法	50.3(±3.0)	64.6(±3.7)
提案手法	60.1*(±3.3)	73.8**(±4.7)

vs. 従来手法*p = 3.4e-17 < 0.05　**p = 2.2e-11 < 0.05

手法では正解した例を示す．

- ネットワークプリントサービス めっちゃ 便利$_1$ や [様子・様態，肯定的，容易，優しさ，顧客・ユーザ]$_1$

前節で示した製品 B の分析と同様に単語拡張がポジティブの極性判定に寄与している．また，提案手法で極性判定に失敗した例を以下に示す．

- コンビニプリントサービス A で 試し$_1$ 刷り した んだ けど 納得$_2$ の いく 緑$_3$ が 出 ない … [勢力・程度，価値・質，正確，動作，行為，反応，素材・材料]$_1$ [様子・様態，肯定的，理性的，思想・哲学]$_2$ [植物，交通・輸送，環境，肯定的，色彩，緑，農業]$_3$

このツイートは否定形が極性を反転しており，前節の分析のとおり，PV-DBOW（提案手法）のみでは対処できない．本ツイートは，PV-DM と結合することにより，ネガティブと判定されることを確認した．

5. むすび

本論文では，Twitter における単語のスパース性の問題を解決することを目的に単語意味ベクトル辞書を用いて単語拡張した文書をパラグラフベクトルのモデルで学習する統合化手法を提案した．

本論文ではまず，商品開発や品質サポートに役立つ評判情報抽出手法を提案した．次にベースラインとしての BoW，パラグラフベクトル，提案手法を用いた評判情報抽出の実験システムを構築し，評価実験を行った．結果として，提案手法の単語拡張は，パラグラフベクトルに対しては単語のスパース性が高いほど，優位性を示した．単語がスパースなベンチマークにおける開発セットへの過学習の問題は，テストセットに単語ベクトルを適応させたときに解消されることを確認した．語順を考慮しない BoW に対しては，ツイートの文長が長いほど，パラグラフベクトルに優位性があることを示した．また，提案手法のエラー分析の結果，単語のスパース性が高いベンチマークでは，語順を考慮する PV-DM モデルの効果が相対的に弱まり，誤判定となるケースがあることが分かった．更に多様な製品からなる小規模なベンチマークを構築し，提案手法の有効性を示した．

提案手法の今後の展開として，単語ベクトルの初期値として人手構築による単語意味ベクトルを与えることにより，可読性の高いパラグラフベクトルの実現を目指す．これにより，少ない学習データから信頼性のある単語やパラグラフの意味表現学習を実現し，タスクの評価に依存しない意味表現学習の品質評価法を確立することが今後の課題である．また，**4.5** の製品 B のエラー分析の結果として，PV-DM を二つ結合して提案手法における PV-DM のウエイトを増やすことにより，F 値 72.1 まで向上することを確認している．今後は，ベンチマークに応じて，PV-DM，PV-DBOW，PV-DBOW（提案手法）のウエイトをパラメータとして調整できるようにし，また人手構築の単語極性辞書を単語意味ベクトル辞書に組み込むことにより，極性分析の更なる精度向上を目指す．更に単語意味ベクトル辞書の基本単語と特徴単語を英語化することにより，評判情報抽出の英語対応が実現できると考えられる．

本論文では，製品の評判情報抽出に限定して評価を行ったが，提案手法はこれに限定されるものではなく，特定分野の評判情報抽出に適用可能であると考えられる．また，提案手法は，評判情報抽出に限定されず，対話システムにおける意図推定 [12] のように発話のスパース性が課題となる自然言語処理のタスクの精度改善においても有効性を示すことが今後の課題である．

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．

文　献

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” Proc. Workshop at ICLR, 2013.

[2] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Proc. NIPS, pp.3111–3119, 2013.

[3] T. Mikolov, W. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” Proc. NAACL HLT, pp.746–751, 2013.

[4] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” Proc. ICML, pp.1188–1196, 2014.

[5] 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，” 信学論（D-II），vol.J79-D-II, no.4, pp.484–491, April 1996.

[6] 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, pp.15–20, 2000.

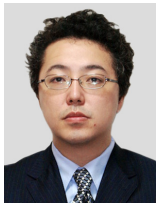
[7] S. Rosenthal, P. Nakov, S. Kiritchenko, S.M. Mohammad, A. Ritter, and V. Stoyanov, “SemEval-2015 Task 10: Sentiment analysis in Twitter,” Proc.

SemEval-2015, pp.451–463, 2015.
[8] W. Boag, P. Potash, and A. Rumshisky, “Twitter-Hawk: A feature bucket approach to sentiment analysis,” Proc. SemEval-2015, pp.640–646, 2015.
[9] S. Mohammand, S. Kiritchenko, and X. Zhu, “NRC-Canada: Building the sate-of-the-art in sentiment analysis of tweets,” Proc. SemEval-2015, pp.321–327, 2013.
[10] R.F. Astudillo, S. Amir, W. Lin, M. Silva, and I. Trancoso, “Learning word representations from scarce and noisy data with embedding subspaces,” Proc. ACL-IJCNLP, pp.1074–1084, 2015.
[11] J. Snoek, H. Larochelle, and R.P. Adams, “Practical Bayesian optimization of machine learning algorithms,” Advances in Neural Information Processing Systems, pp.2951–2959, 2012.
[12] 石川葉子，平岡拓也，水上雅博，吉野幸一郎，G. Neubig，中村　哲，“対話状態推定のための外部知識ベースを利用した意味的素性の提案，” 情処学音声言語情報処理研報，vol.2015-SLP-109, no.13, 2015.

（平成 28 年 6 月 29 日受付，11 月 4 日再受付，
29 年 1 月 6 日早期公開）

芥子　育雄　（学生員）

1981 年阪大･工･電子卒．1983 年同大大学院修士課程了．同年シャープ（株）入社．以来，知的文書処理，連想検索，ホームヘルスケアの研究開発，操作ナビゲーション・エージェント，レコメンデーションサービス，ビッグデータ解析の商品・事業開発に従事．1989 年 MIT AI Lab. 客員研究員，2003 年システム開発センター室長，2008 年先端映像技術研究所技師長，2015 年 9 月退職，同年 10 月から奈良先端大情報科学研究科博士後期課程に在籍．情報処理学会，人工知能学会各会員．



鈴木　　優　（正員）

1999 年神戸大学工学部情報知能工学科卒業．2001 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．2004 年同博士後期課程修了．博士（工学）．2004 年立命館大学情報理工学部講師，2009 年京都大学大学院情報学研究科特定研究員，2010 年名古屋大学大学院情報科学研究科特任助教を経て 2014 年奈良先端科学技術大学院大学情報科学研究科特任准教授．データ工学，情報検索に関する研究に従事．IEEE Computer，ACM，情報処理学会，日本データベース学会各会員．

吉野幸一郎

2009 年慶應義塾大学環境情報学部卒業．2011 年京都大学大学院情報学研究科修士課程修了．2014 年同博士後期課程修了．同年日本学術振興会特別研究員（PD）．2015 年より奈良先端科学技術大学院大学情報科学研究科特任助教．京都大学博士（情報学）．音声言語処理及び自然言語処理，特に音声対話システムに関する研究に従事．2013 年度人工知能学会研究会優秀賞受賞．IEEE，ACL，情報処理学会，言語処理学会各会員．

グラム　ニュービッグ

2005 年米国イリノイ大学アーバナ・シャンペーン校工学部コンピュータ・サイエンス専攻卒業．2010 年京都大学大学院情報学研究科修士課程修了．2012 年同大学院博士後期課程修了．同年奈良先端科学技術大学院大学助教．機械翻訳，自然言語処理に関する研究に従事．

大原　一人

1995 年東北大・工・情報工卒．1997 年同大学大学院修士課程修了．同年シャープ（株）入社．動画像符号化技術の応用に関する研究・開発に従事．大阪大学博士（情報科学）．

向井　理朗

1992 年東京理科大学理工学部卒業．1994 年同大学大学院理工学研究科修士課程修了．同年シャープ（株）入社．1995 年より 3 年間，技術研究組合新情報処理開発機構へ出向．以来，マルチモーダルインタフェース，動画像認識，対話システムの研究開発，レコメンデーションサービス，ビッグデータ解析の商品・事業開発に従事．

中村　　哲　（正員）

1981 年京都工芸繊維大学電子卒．京都大学博士（工学）．シャープ株式会社．奈良先端大助教授，2000 年 ATR 音声言語コミュニケーション研究所室長，所長，2006 年 (独) 情報通信研究機構研究センター長，けいはんな研究所長などを経て，現在，奈良先端大教授．ATR フェロー．カールスルーエ大学客員教授．音声翻訳，音声対話，音声・自然言語処理の研究に従事．情報処理学会喜安記念業績賞，総務大臣表彰，文部科学大臣表彰，Antonio Zampoli 賞受賞．IEEE SLTC 委員，ISCA 理事，IEEE フェロー．