Vol.2017-NL-231 No.8
Vol.2017-SLP-116 No.8
2017/5/15

分散的意味表現学習のための単語意味ベクトル辞書 Ver.2 と日本語 Twitter 極性分析ベンチマークについて

芥子 育雄[1,a)]　鈴木 優[1]　吉野 幸一郎[1]　大原 一人[2]　向井 理朗[2]　中村 哲[1]

概要：自然言語処理の応用システムでは，大規模文書を対象に単語やパラグラフの意味を数 100 次元のベクトルに埋め込む分散的意味表現学習により精度向上が図られている．これらの応用システムの課題は，品質保証・改善に必要な分散表現の中身を解釈する手段が無いことである．本研究では，分散表現の可読性を向上させることを目的に人手により構築された単語意味ベクトル辞書 Ver.2 を提案する．単語意味ベクトル辞書とは，約 2 万語の基本単語に対して，関係のある特徴単語を列挙した辞書である．Ver.2 では特徴単語を 264 種類とし，日本語の単語意味ベクトル辞書を元にクラウドソーシングを利用して英語版を作成した．ツイートベクトルの可読性を高めるために，基本単語に付与された特徴単語を再帰的に展開することにより生成した多値ベクトルを分散的意味表現学習のシードベクトルとして用いる．この評価のため，日本語 Twitter を対象に製品，サービス，組織の全 8 カテゴリ，38,576 ツイートから構成される極性分析ベンチマークをクラウドソーシングを利用して作成した．本稿では，モニター公開を予定している日英単語意味ベクトル辞書 ver.2 および評判分析のための大規模かつ多様性のある日本語 Twitter 極性分析ベンチマークについて，ケーススタディと共に報告する．

キーワード：意味ベクトル，分散表現，word2vec，パラグラフベクトル，極性分析，Twitter，ベンチマーク

1. はじめに

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], [2], [3]．また，Le と Mikolov は，単語の分散的意味表現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 [4].

Association for Computer Linguistics は，シェアードタスクとして 2013 年から継続して英語 Twitter を対象に極性分析タスクを SemEval で開催している [5], [6]．2016 年は 43 チームが参加し，SemEval では最も参加チームの多いタスクである．

日本語 Twitter においても製品，サービス，組織などに対する評判分析の必要性は高まっている．評判分析サービスも提供されているが，文長が短く，ノイズが多く，単語のスパース性が高いといった Twitter の課題から，評判分析の精度において利用者のニーズを満たしているとは言えない．著者らは，パラグラフベクトルの学習において，Twitter の課題を解消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した [7], [8], [9]．約 1 万 2 千ツイートから構成される特定スマートフォン製品ブランドの極性分析ベンチマークにおいて，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，パラグラフベクトルによる評価結果を 3.2 ポイント上回った [9].

Twitter の課題を解消するために文字（Unicode）単位の Twitter データをエンコーダ・デコーダモデルで学習させる 2 種類の Tweet2Vec が同時期に発表された [10], [11]. Vosoughi らの Tweet2Vec[11] は，WordNet[12] を用いた同義語拡張により，ランダムに選択した 300 万ツイートを複製して学習データを増やし，LSTM-CNN モデルを学習させた．SemEval2015[5] の極性分析ベンチマーク（訓練セット：9,520 ツイート，テストセット：2,380 ツイート）で評価を行い，パラグラフベクトルのマクロ平均 F 値を 1.9 ポイント上回り，他の SemEval 参加チームの F 値も上回り最高水準を示した．分散表現の分類器にはロジスティック回帰を用いている．Dhingra らの Tweet2Vec[10] は，ツイー

[1] 奈良先端科学技術大学院大学情報科学研究科
8916-5 Takayama-cho, Ikoma, Nara 630–0192, Japan
[2] シャープ株式会社 IoT 通信事業本部
1-9-2 Nakase, Mihama-ku, Chiba, Chiba 261-8520, Japan
[a)] keshi.ikuo.ka9@is.naist.jp

トからハッシュタグを予測するエンド・ツー・エンドのシステムである．

2 種類の Tweet2Vec では，ニューラルネットワークへの入力を単語単位ではなく文字単位で行うことにより，Twitter の課題を解消すると共に言語に依存せず高い性能を示すことが期待される．しかし，品質を確認する手段がタスクの精度であることは変わらず，精度が実用レベルに満たない場合は学習データを追加で準備する必要がある．Vosoughi らの Tweet2Vec[11] における WordNet を用いた同義語拡張は学習データを追加で準備することと等価と考えられる．これに対して，著者らの単語意味ベクトル辞書を用いた特徴単語の展開は，2.2 節で述べる通り，タスクの精度が向上するだけでなく，エラー解析においても効果がある．また，単語意味ベクトル辞書をニューラルネットワークのシードベクトルとして用いる場合は，ニューラルネットで学習した重みの大きな特徴単語を品質の確認に利用できる可能性がある．

本研究では，基本単語に付与された特徴単語を再帰的に展開することにより，単語意味ベクトル辞書をパラグラフベクトルのシードベクトルとする方法を示し，パラグラフベクトルを用いて学習を行なったツイートベクトルの重みの大きな特徴単語がツイートの内容を表す例を示す．そして，ニューラルネットワークのシードベクトルとしての利用を前提とした単語意味ベクトル辞書 Ver.2 を提案する．Ver.2 では特徴単語を従来の 266 種類から 4 の倍数となる 264 種類に削減し，新たに英語版の構築を行なった．この評価のため，日本語 Twitter を対象に製品，サービス，組織の全 8 カテゴリ，38,576 ツイートから構成される極性分析ベンチマークを構築した．本ベンチマークを用いたケーススタディについて報告する．なお，単語意味ベクトル辞書 Ver.2 と日本語 Twitter 極性分析ベンチマークはモニター公開を予定している．

2. 単語意味ベクトル辞書

本章では，最初に単語意味ベクトル辞書の設計方針を明確にする．次に Twitter の課題を解消するために単語意味ベクトル辞書を用いた特徴単語の展開方法とエラー解析例について説明する．最後にツイートベクトルの品質確認を目的とした単語意味ベクトル辞書に基づくシードベクトルの作成方法と実例を示す．

2.1 単語意味ベクトル辞書の設計方針

基本単語の意味ベクトル [7], [9] 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, …, 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．基本単語は，百科事典[*1]や新聞記事[*2]の説明に使われる用語，WWW ホームページの分類用語，取扱説明書などで使われる操作用語，および形容詞などの感性語から 2 万 336 語を選択した．

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	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，…

表 2 論理的関連による特徴単語の付与基準

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

これら選択した基本単語に対して，辞書編纂の専門家が，論理的関連性と連想的関連性から，特徴単語を付与した．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．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．

2.2 単語意味ベクトル辞書を用いた Twitter の単語拡張

パラグラフベクトルの学習において，Twitter の課題を解消するために著者らが提案した単語意味ベクトル辞書の

*1 ブリタニカ小項目事典 CD 版, TBS ブリタニカ, 1992.
*2 CD-毎日新聞’94, ’95 データ集, 毎日新聞社, 1994, 1995.

表 3 連想的関連による特徴単語の付与基準

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

導入方法とその効果を示す [8], [9]．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ことにより，文長が短い Twitter では適切に捉えることが難しい文脈情報の学習が改善する．ツイート中の基本単語を特徴単語に展開した例を図 1 に示す．このツイートでは，「真偽，製，端末，インチ，画面，非常，魅力，的」の 8 個の基本単語が抽出され，単語意味ベクトル辞書を用いて単語拡張を行った．パラグラフベクトルには，図 2 に示す通り，2 種類のモデルがある．PV-DM は語順の情報を保持し，次単語を予測するモデルである．PV-DBOW（単語ベクトル学習時は Skip-gram）はパラグラフ中の文脈情報を学習するためのモデルである．特徴単語を単語拡張したツイートの学習には PV-DBOW を用いて学習し，従来のパラグラフベクトル（PV-DM, PV-DBOW）と結合した分散表現と正解ラベル（ポジティブ，ニュートラル，ネガティブ）を用いて，SVM により分類器を構築した．これにより，1 章で述べた通り，極性分析の F 値がパラグラフベクトルに対して 3.2 ポイント改善した．

特徴単語の単語拡張は，少ない学習データにおいても文脈情報を明確にすることが目的である．表 4 にエラー解析の例を示す．特徴単語の単語拡張により，改善（不正解→正解），失敗（正解→不正解）したツイート群において，対群に対して出現比率の高い特徴単語を頻度順に並べたものである．表の括弧内の数値が対群に対する出現比率を示す．改善したツイート群では，ポジティブでは「肯定的」など，ネガティブでは「否定的」などの特徴単語の出現頻度が高く，特徴単語の単語拡張により文脈情報が明確になったことが分かる．一方で，失敗したツイート群では，ポジティブで「否定的」，ネガティブでは「肯定的」などの逆の意味を持った特徴単語の出現頻度が増えており，これは否定形などのツイートであり，PV-DBOW では対応出来ないことが原因と考えられる．

例を示した通り，特徴単語による単語拡張は，機械学習の可読性を高めることにも効果があると考えられる．

2.3 単語意味ベクトル辞書に基づくシードベクトルの作成

辞書の単語に対する定義文を再帰展開することにより単語ベクトルを生成する手法は提案されている [13]．単語意味ベクトル辞書は，266 種類の特徴単語で基本単語を定義しているとみなすことが出来る．特徴単語も基本単語となるため再帰展開が必要だが，定義文が 266 語に限定されるため，数回展開すれば収束する．また，単語ベクトルを辞書の関連語に合わせて修正する手法も提案されている [14]．ここでは，特徴単語の初期ベクトルを対応する次元を 1 とする One-hot ベクトル，他の基本単語はゼロベクトルとし，Faruqui らの Retrofitting ツール[*3][14] を用いて，単語意味ベクトル辞書を再帰的に展開することにり基本単語のシードベクトルを生成した．基本単語「旅行」のシードベクトルを設定した Skip-gram（PV-DBOW）モデルの例を図 3 に示す．中間層は特徴単語に対応する 266 種類のノー

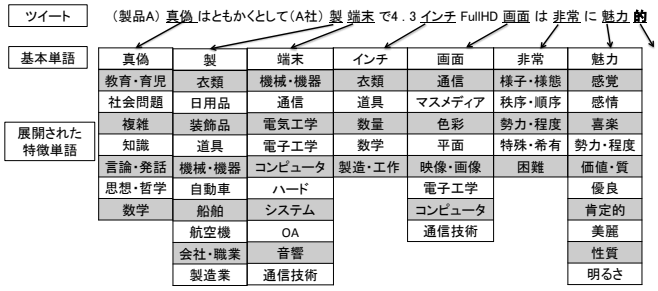

図 1 特徴単語の展開例

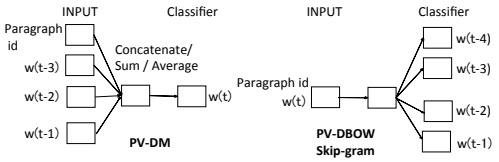

図 2 パラグラフベクトルの 2 種類のモデル

表 4 特徴単語の単語拡張により改善，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）

	開発セット		テストセット	
	不正解 → 正解	正解 → 不正解	不正解 → 正解	正解 → 不正解
Posi-tive	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)
Nega-tive	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)		

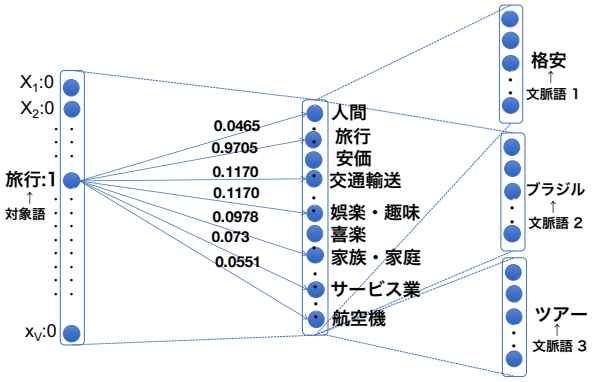

図 3 Skip-gram モデルへのシードベクトルの設定

[*3] https://github.com/mfaruqui/retrofitting

製品B すげ ぇ
勢力・程度, 強力, 人間, 価値・質, 様子・様態, 施設・設備, 教育・育児, 肯定的, 優良,

製品B は 神
様子・様態, **思想・哲学**, **人間**, 関係・関連, 生死, **実存**, **勢力・程度**, 水棲生物, 国家, **強力**,

あー やっぱり 製品B 音 いい w なんか 深み ？ が ある w
音響, 宣伝広告, 様子・様態, 映像・画像, **感情**, 価値・質, **音楽**, 勢力・程度, 感覚, 劣悪,

製品B で 音楽 きい たら 音質 めちゃくちゃ よく て ビックリ し た w …
様子・様態, **音楽**, 施設・設備, 動作, 倫理・道徳,サービス業, **感情**, 数量, 価値・質, 文化,

製品B って 充電 終わっ て も led 点灯 し た まんま な ん だ …
明るさ, 機械・機器, **発熱・発光**, 様子・様態, コンピュータ, 活動, 実質・本質, 新しさ, 建造物,

図 4 ツイートと重みが大きな特徴単語（中間層ノード）の例

表 5 基本単語に付与された上位の特徴単語

特徴単語	基本単語数
様子・様態	5,188
関係・関連	4,460
勢力・程度	3,217
秩序・順序	2,689
強力	2,645
肯定的	2,455

ドから構成される．

図 4 にスマートフォンの製品ブランドに関する約 56 万ツイートでパラグラフベクトル（PV-DBOW）を学習させたときの 5 例について，各ツイートと重みの大きな特徴単語（中間層のノード）を上位から順に示す．「製品 B すげぇ」「製品 B は神」に対しては共通の特徴単語「勢力・程度」「強力」「人間」の重みが大きく，「神」のツイートに対しては「思想・哲学」や「実存」のように連想的関連のある特徴単語の重みが大きい．次の 2 例のスマートフォンの音質に関するツイートでは，共通の特徴単語「音楽」「感情」の重みが大きい．最後の充電と led 点灯に関するツイートでは，「明るさ」「発熱・発光」の内容に関連した特徴単語の重みが大きい．

3. 分散的意味表現学習のための単語意味ベクトル辞書 Ver.2

3.1 単語意味ベクトル辞書 Ver.2 の要件

単語意味ベクトル辞書 Ver.2 は，ニューラルネットワークの中間層のノードとして特徴単語を与えることにより，パラグラフベクトルの可読性を高めることを目的とする．本目的における Ver.2 の要件は以下の 3 点である．

- 特徴単語がパラグラフベクトルの各次元に対応するため，特徴単語数は計算およびメモリ効率の良い 4 の倍数であること．
- 特徴単語を単語拡張として利用する場合は高頻度語はダウンサンプリングされるが，中間層のノードとする場合は多くの基本単語に付与されている特徴単語の重みは大きくなるため削除する必要がある．
- 特徴単語は概念分類のため世界共通と考えられるが，基本単語の英語化は必要である．英語化においては，ソーシャルメディアで発信されるような平易な英単語・語句を選択すること．

表 6 単語意味ベクトル辞書 Ver.2 の仕様

	基本単語数	基本単語の平均特徴単語数
日本語版	20,330 語	8.77 語
英語版	21,912 語	11.73 語

以上の要件から，日本語の単語意味ベクトル辞書において，基本単語に付与された上位の特徴単語を表 5 に示す．表 1 の「抽象概念」に属する特徴単語「様子・様態」および「関係・関連」が極端に多くの基本単語に付与されていることが分かる．特に「様子・様態」は，図 4 に示した全ての例において，ツイートベクトルの重みが大きな特徴単語となっており，影響が強過ぎることが分かる．それ以下の特徴単語に関しては，表 4 に示した通り，単語拡張に利用した場合はポジネガの推定において重要な役割を果たしている．従って，「様子・様態」および「関係・関連」を削除し，特徴単語数を 264 種類と 4 の倍数とする．

3.2 英語版単語意味ベクトル辞書の構築

基本単語の英語化のフローを図 5 に示す．最初に 264 種類の特徴単語を英語化し，特徴単語が全く付与されていない基本単語を削除した．基本単語 2 万 300 語について，クラウドサービスとして提供されているニューラル機械翻訳の Microsoft Translator API[*4]を利用して，基本単語を英単語あるいは英語句に機械翻訳を行なった．日英辞書ではなく，ニューラル機械翻訳を利用した理由は，難解な百科事典や新聞記事の説明用語を日常良く使われる英単語や語句に翻訳することが目的である．結果として，翻訳誤りを含めて約 14,000 語のユニークな英単語・語句に翻訳された．この機械翻訳された英単語・語句を翻訳元の日本語単語，および列挙された特徴単語との意味的，連想的関連性から基本単語の校正をクラウドソーシングを利用して依頼した．日本語新聞が読める英語ネイティブ，もしくは TOEIC900 点以上の日本人を募集し，応募者に基本単語 100 語の校正を無償のトライアルとして依頼した．100 語の校正結果，校正に掛った時間，見積もりから 3 名のクラウドワーカーを選択した．特にオンライン辞書を調べなくても校正が可能な日本語の知識があり，英語のセンスがある方を優先した．例えば，クラウドワーカー A は TOEIC 満点の米国在住経験 10 年以上（日本語・英語ネイティブ）の女性であり，トライアルを 10 分で行なった．ユニークな英単語・語句は 13,551 語であり，内 7,692 語はニューラル機械翻訳の結果をそのまま採用したことを示す．3 名の校正結果をマージし，21,912 語の英語版意味ベクトル辞書を作成した．単語意味ベクトル辞書 Ver2. の仕様を表 6 にサ

[*4] https://www.microsoft.com/en-us/translator/

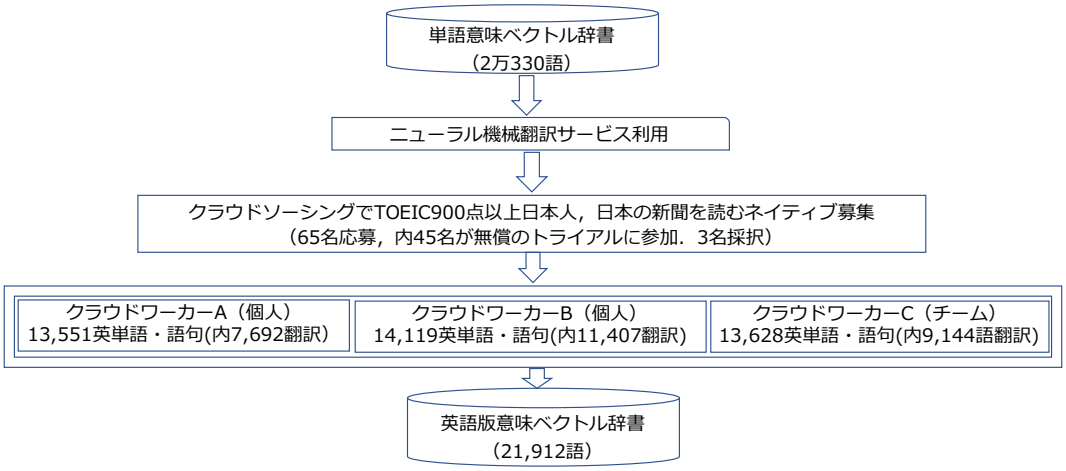

図 5 英語版構築のフロー

```
演奏会 娯楽・趣味 共同 多数・多量 文化 芸術 音楽 音響
生化学 微生物 植物 人間の身体 内臓器官 生物の身体 健康・美容 医学・薬学…
遠大 勢力・程度 大規模 広大 未来 思想・哲学 計画
短銃 殺生 機械・機器 軍事・防衛 戦争・紛争 短さ 軍事技術
建つ 住居 建造物 施設・設備 立体 形状 高さ 土木・建築 建設業
出来 誕生 活動 社会問題 災害 変化 価値・質 因果 肯定的 新しさ 容易 強力 動作 …
建て 住居 建造物 会社・職業 施設・設備 国家 経済 金融 秩序・順序 勢力・程度…
雪氷 気象・気候 凝固・凍結 寒冷 白
厳し 活動 財政 エネルギー問題 国際関係 極地 感覚 否定的 複雑 困難
沖積 陸地 環境 変化 地理 地学
平然 感覚 感情 勢力・程度 一般・平凡 単純 容易 強力 理性的 個人 静的 軽さ…

auctions economy behavior vast commerce customer
audicious emotion strong property behavior substantial thickness idea
audience hobby sport enjoyment majority general individual art image music customer
audience_seats sport structure facility space art image music customer
audio human language sound hobby machine music electronics
audio_source machine computer software sound
audio_training japan knowledge discussion book literature language
audiovisual human_body visceral_organ health sense medicine
audiovisual_senses human_body visceral_organ health sense medicine
```

図 6 単語意味ベクトル辞書 Ver.2 のサンプル

ンプルを図 6 に示す．辞書の各行先頭に基本単語を置き，スペースを空けて特徴単語が列挙されている．英語句は，アンダーバーで英単語を結合している．

4. 日本語 Twitter 極性分析ベンチマーク

これまで著者らは，日本語 Twitter を対象に 2 種類のスマートフォン製品ブランドのベンチマーク（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，およ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，提案手法の評価に用いてきた [8], [9]．しかし，製品数が 2 種類では評価は十分とは言えないため，多様なカテゴリー（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，組織）を含む小規模なベンチマークを作成し，提案手法の評価を行なった [9]．多様性のあるベンチマークで評価を行うことの有効性を確認できたため，本研究では，クラウドソーシングを利用して，多様性があり，かつ大規模なベンチマークを構築した．

各カテゴリーごとに収集したツイート数，クラウドソーシングによりラベル付与を行なったツイート数を表 7 に示す．スマートフォン A，B に関しては 2014 年 10 月〜2015 年 11 月の 13 ヶ月分，その他のカテゴリーに関しては 2015 年 1 月〜2016 年 2 月の 13 ヶ月分のツイートを製品名などのキーワードで収集を行なったものである。表 7 の総ツイート数は収集したツイートの全数を示す．各カテゴリーごとにクラウドソーシングを利用してラベル付けを行なった．各ツイートには 5 名の作業者を割り当て，多数決により，ラベルを付与した．多数決の結果，同点の場合は，1 ツイートに複数のラベルを付与した．ラベルは以下の 5 種類である．

- **ポジティブ**：対象カテゴリーの具体的な特徴に対して，ポジティブな意見を発信しているツイート．
- **ネガティブ**：対象カテゴリーの具体的な特徴に対して，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**：対象カテゴリーに対して，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．
- **ポジ&ネガ**：対象カテゴリーの具体的な特徴に対して，ポジティブな意見とネガティブな意見の両方を発信しているツイート．
- **無関係**：対象のカテゴリーに対しての個人の意見を発信していないツイート．

ここで注意が必要な点は，大規模ツイートに対してのクラウドソーシングによる作業を効率的に行うため，ポジティブおよびネガティブの判断基準を「具体的な特徴に対して」と明確にしたことである．従って，従来はポジティブと判定していた図 6 の最初の 2 例は今回のラベル付与作業ではニュートラルとなり，具体的な特徴「音」や「充電機能」について発信している後半 3 例がポジティブ，ネガティブの対象となる．

クラウドソーシングによるラベル付与結果を元に作成した日本語 Twitter 極性ベンチマークを表 8 に示す．クラウドソーシングによるラベル付与作業を行なったツイートの

表 7 クラウドソーシングによるツイートへのラベル付与数

カテゴリー	総ツイート数	ポジティブ	ネガティブ	ニュートラル	ポジ&ネガ	無関係	総ラベル付与数
スマートフォン A	130,650	2,906	5,188	16,054	594	68,158	92,900
スマートフォン B	482,036	5,655	9,531	51,900	603	18,884	86,573
スマートフォン C	1,155,034	3,543	6,176	45,568	408	28,844	84,539
ロボット掃除機 A	11,664	741	311	6,894	41	4,371	12,358
ロボット掃除機 B	307,156	954	1,089	20,654	55	48,092	70,844
コンビニプリントサービス	275,097	3,887	3,484	30,176	241	35,514	73,302
メーカー A	187,584	744	4,421	40,950	75	26,358	72,548
メーカー B	169,532	1,503	937	13,624	80	54,891	71,035
総数	2,718,753	19,933	31,137	225,820	2,097	285,112	564,099

表 8 日本語 Twitter 極性分析ベンチマーク

データセット	ポジティブ	ネガティブ	2 クラス計	ニュートラル	無関係	合計
訓練セット	10,100	15,618	25,718	137,089	180,186	342,993
開発セット	2,525	3,904	6,429	34,272	45,046	85,747
テストセット	2,525	3,904	6,429	34,272	45,046	85,747
合計	15,150	23,426	38,576	205,633	270,278	514,487
ラベル無し			2,204,266			

表 9 コーパスの統計情報

項目	語数
語彙（出現頻度 5 以上）	126,213 語
コーパス中総出現単語	79,640,916 語
ダウンサンプリング対象高頻度語	2,910 語
基本単語	12,937 語

表 10 極性分析による 2 クラス分類精度（標準偏差）

シードベクトル	開発セット	テストセット
単語意味ベクトル辞書	89.2%(0.3%)	88.2% (0.1%)
ランダム設定	88.7%(0.4%)	88.6% (0.2%)

うち，ポジ&ネガのラベルや多数決により複数ラベルが付与されたツイートについては，今回はラベル無しツイートに含めた．ポジティブ，ネガティブの 2 クラス分類では，38,576 ツイートと SemEval のベンチマークと比較しても大規模なベンチマークである．また，ポジティブ，ネガティブの分類基準が明確なため信頼性が高く，製品，サービスや組織の具体的な特徴が明記されているため，商品企画や品質サーポトにとって役に立つツイートと考えれる．

本ベンチマークはニュートラルを含めた 3 クラス分類に使うことも可能である．しかし，単に製品やサービスが好き，欲しい，嫌い，必要ないのようなツイートは今回はニュートラルに含めたため，ニュートラルが 1 桁大きな不均衡データとなった．このため，3 クラウス分類において，ポジティブ予測とネガティブ予測のマクロ平均 F 値を高めることは非常に難しいタスクである．実応用としては，本ベンチマークから無関係をフィルタリングし，ニュートラルをフィルタリングした後，ポジティブやネガティブを高精度で分類できなければ，商品企画や品質サポートの要望に応えることはできない．

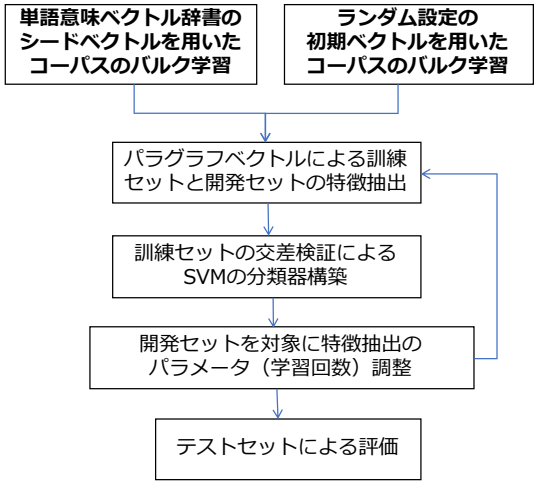

図 7 極性分析の手順

5. ケーススタディ

単語意味ベクトル辞書 Ver.2 を利用して日本語基本単語のシードベクトルを作成し，パラグラフベクトル（PV-DBOW）を用いて，日本語 Twitter 極性ベンチマークの 2 クラス分類を対象にケーススタディを行なった．表 7 の約 272 万ツイート，2 種類のスマートフォン製品ブランドに関する約 56 万ツイート，日本語単語意味ベクトル辞書 2 万 330 行を対象に出現頻度 5 回以上の単語をボキャブラリとした．その統計情報を表 9 に示す．

日本語 Twitter 極性ベンチマークを用いた極性分析の手順を図 7 に示す．シードベクトルとして単語意味ベクトル辞書を用いた場合とパラグラフベクトルのデフォルト設定であるランダムの初期ベクトルを用いた場合の 2 クラス分類の精度を比較した．表 10 に示す通り，シードベクトルとして単語意味ベクトル辞書を用いた場合，開発セット

89.2%，テストセット 88.2%と高い精度を示し，5 回試行の標準偏差も低く，パラグラフベクトルに匹敵する．また，以下に例を示す通り，ロボット掃除機，コンビニプリントサービス，組織の多様なツイートにおいてもツイートに関連した特徴単語が大きな重みを持つことを確認した．

ロボット掃除機 B すげ ー わ 部屋 の 空気 が すごい きれい に なっ た

[住居, 家族・家庭, 施設・設備, 環境, 建造物, 新しさ, 機械・機器]

A 社 さん の ロボット掃除機 A と 政宗 様 へ の 愛 を ひしひし と...

[顧客・ユーザ, 優しさ, 住居, 芸術, 平和, 音響, 会社・職業, 人間]

コンビニ プリント で 小冊子 作れる の ほんと 便利

[書物・書籍, 勢力・程度, 価値・質, 新しさ, 製造・工作, 教育・育児]

え ！ B 社 社内 同性婚 ok ！ ？ 企業 が すすめる って すごい ！ ！

[会社・職業, 商業・貿易, 価値・質, 公共制度, 経済, 国際関係, 社会活動]

6. おわりに

本稿では，分散的意味表現のための単語意味ベクトル辞書 Ver.2 および評判分析のための大規模かつ多様性のある日本語 Twitter 極性ベンチマークを提案した．また，単語意味ベクトル辞書 Ver.2 を元にパラグラフベクトルのシードベクトルを作成し，日本語 Twitter 極性ベンチマークを用いたケーススタディを示した．ポジティブとネガティブの 2 クラス分類では，88%以上の精度を示しデフォルト（初期ベクトルがランダム）のパラグラフベクトルに匹敵すること，および重みが大きな特徴単語はツイートとの関連性が高いことを確認した．

単語意味ベクトル辞書 Ver.1 は約 20 年前に作成されたため一部の単語は難解であり，Twitter のコーパスに 1 度も出現しない基本単語が 10%程度ある．ニューラル機械翻訳とクラウドソーシングを利用して作成した Ver.2 の英語版に関しては，より現代の感覚に合った基本単語が選択された．その結果として，アンダーバーで複数の単語を接続した語句が多くなった．英語版の利用においては，英文テキストからの語句の適切な抽出が課題となる．

日本語 Twitter 極性ベンチマークは，日本のみならず海外においても共通のベンチマークとして活用されることを期待している．1 章で述べた通り，言語に依存しない文字単位のエンコーダ・デコーダモデルによる極性分析システムとの性能比較が可能となる．

今後の課題としては，システム性能比較のためのベンチマークではない，実応用を視野に入れた本ベンチマークを用いたシェアードタスクの実現可能性を検討する．また，単語意味ベクトル辞書 Ver.2 は現代の感覚に合った辞書となるように基本単語・特徴単語の追加・削除や英語句は抽出し易いように改良するなど，継続的な辞書のアップデートを実現する仕組みを確立することが重要である．さらには，パラグラフベクトルの可読性による品質確認を超えた新規のアプリケーション提案を目指す．

謝辞 本研究の一部は，NAIST ビッグデータプロジェクトの助成を受けたものである．

参考文献

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Proc. of 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. of NIPS, pp.3111-3119, 2013.
[3] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations, " Proc. of NAACL HLT, pp.746-751, 2013.
[4] Q. Le, T. Mikolov, "Distributed representations of sentences and documents, " Proc. of ICML, pp.1188-1196, 2014.
[5] S. Rosenthal, P. Nakov, S. Kiritchenko, S. M Mohammad, A. Ritter, and V. Stoyanov, "SemEval-2015 Task 10: Sentiment analysis in Twitter," Proc. of SemEval-2015, pp.451-463, 2015.
[6] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 task 4: Sentiment analysis in Twitter," Proc. of SemEval-2016, pp.1-18, 2016.
[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，” 信学論 （D），vol.J79-D-II，no.4，pp.484-491，1996.
[8] 芥子育雄，鈴木優，吉野幸一郎，大原一人，向井理朗，中村哲，”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，” 第 8 回 データ工学と情報マネジメントに関するフォーラム論文集 (A1-3)，2016.
[9] 芥子育雄，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，”単語意味ベクトル辞書を用いた Twitter からの日本語評判情報抽出，” 信学論 （D），vol.J100-D，no.4，pp.530-543，2017.
[10] B. Dhingra, Z. Zhou, D. Fitzpatrick, M. Muehl, and W. Cohen, "Tweet2Vec: Character-Based Distributed Representations for Social Media," Proc. of ACL, vol.2, pp.269-274, 2016.
[11] V. Soroush, V. Prashanth, and R. Deb, "Tweet2Vec: Learning Tweet Embeddings Using Character-level CNN-LSTM Encoder-Decode," Proc. of SIGIR, pp.1041-1044, 2016.
[12] C. Fellbaum, "WordNet," Wiley Online Library, 1998.
[13] 鈴木敏，”辞書に基づく単語の再帰的語義展開，” 情報処理学会論文誌，vol.46, no.2, pp.624-630, 2005.
[14] M. Faruqui, J. Dodge, S. Jauhar, C. Dyer, E. Hovy, and N. Smith, "Retrofitting Word Vectors to Semantic Lexicons," Proc. of NAACL, pp.1606-1615, 2015.