Semantic Relation Analysis among Domains Using Word Embeddings

単語の分散表現を用いた領域における出現単語の特徴分析

佐々木 稔（茨城大学工学部情報工学科）†
古宮 嘉那子（茨城大学工学部情報工学科）
新納 浩幸（茨城大学工学部情報工学科）

Semantic Relation Analysis among Domains Using Word Embeddings

Minoru Sasaki (Ibaraki University)
Kanako Komiya (Ibaraki University)
Hiroyuki Shinnou (Ibaraki University)

要旨

自然言語処理では，コーパス中の単語に対して意味的な特徴をベクトルで表現し，様々な自然言語処理タスクにおいて利用することが多い．単語をベクトルで表現することにより，単語間の類似度を計算し，単語や文の違いなどの比較が可能である．これまでの研究では，単語ベクトルを生成するためには，ひとつの大規模な文書データから生成する必要があった．そのため，書籍・新聞・雑誌など，文書の分野による分散表現の比較や特徴分析は行われていなかった．すなわち，分野による単語の類似性や違いは明らかになっていない．そこで，本研究では領域ごとの単語ベクトル生成手法を提案し，各領域における単語の特徴分析を行う．書籍・雑誌・新聞の３つの分野（領域）の日本語のコーパスに対し，指定した対象単語に対して単語ベクトルを作成する．対象単語のベクトルを用いて，対象単語が類似する単語を各領域において抽出し，単語の使用傾向などの分析を行う．実験の結果，同一単語であっても他領域で使われる意味と異なった語義で使用されている単語があることが分かった．加えて，動詞は領域からの依存度が低い，副詞は領域への依存度が高いなど品詞によって領域の依存度が異なるといった傾向があった．また，書籍領域では様々な種類の語句，新聞領域では政治関連語句，雑誌領域においてはカタカナ語が多く登場するなど，類似した単語には領域によって特徴がみられた．

１．はじめに

自然言語処理で単語の特徴を分析する際，文書中の単語をベクトルで表現することがよく行われている(Schnabel et. al 2015)．単語をベクトルで表現することにより，単語の意味的な特徴を表現することや，単語間の類似度を計算することが可能となる．しかし，これまでの研究では，単語ベクトルを生成するためには，ひとつの大規模な文書データから生成する必要があった．そのため，書籍・新聞・雑誌など，文書の分野による分散表現の比較や特徴分析は行われていなかった．すなわち，分野による単語の類似性や違いは明らかになっていない．

そこで，本研究では領域ごとの単語ベクトル生成手法を提案し，各領域における単語の特徴分析を行う．書籍・雑誌・新聞の3つの分野（領域）の日本語のコーパスに対し，領域ごとに単語ベクトルを作成する．単語ベクトルはコーパスから１単語に対し1つのベクトルを与えるword2vecを用いて作成する(Mikolov et. al. 2013)．単語ベクトルを使いて対象単語の分散表現や類似度の比

† minoru.sasaki.01@vc.ibaraki.ac.jp

較を行い，文書の分野（領域）ごとにどのような特徴の違いや，類似性があるかについて分析することを目的とする．

２．分析方法

本節では，領域の違いによる語義の特徴分析手法について述べる．

２.１　各領域の文書から単語集合の抽出

各領域の文書集合に対して，形態素解析を用いて単語に分割する．単語ベクトルを作成するためには，各領域の文書集合を，単語が半角スペースを区切り文字とした単語列へ変換する必要がある．形態素解析を用いて単語列に変換する際，単語の活用形は見出し語に変形する．本稿では，形態素解析ツールとして日本語形態素解析システム「mecab[1]」を使用する．

２.２　対象単語へのタグ付け

形態素解析を行った各領域の文書集合に対し，対象単語に領域タグを付けて区別する．対象単語 w に対し，領域 1 であれば単語の末尾に領域 1 を表すタグを付けた w_1 に変換する．同様に，領域 n 中の w に対しては w_n に変換する．この手順を繰り返し，すべての領域の対象単語 w にタグ付けを行う．その後，タグ付けが行われた各領域の文書集合を 1 つのコーパスとしてまとめ，これを併合コーパスとする．

２.３　単語ベクトルの生成

併合コーパスに対して word2vec[2]の CBOW モデルを用いて，各領域について出現単語の単語ベクトルを作成する．単語ベクトルとは，ひとつの単語をひとつのベクトルで表した知識表現のことである．CBOW モデルは入力層，中間層，出力層からなり，対象単語の周辺単語を入力とし，対象単語を予測するニューラルネットワークである．入力層と出力層は辞書中の単語と一対一に対応したノードから成るベクトルを表している．中間層は指定した数のノードを隠れ変数として与えることができる．入力層と中間層の間には，それらをつなぐ重みがあり，重み行列として表現する．この重み行列を更新することによって，周辺単語から対象単語を予測するように学習を行う．word2vec は文書集合全体で学習した重み行列を用いて各出現単語の単語ベクトルを出力する．

２.４　領域の違いによる単語ベクトルの比較

対象単語の単語ベクトルを求めることができれば，意味的に関連が強い単語は単語ベクトルが近くなることを用いて，それぞれの領域で単語ベクトルの比較を行う．2 つの領域における対象単語の単語ベクトル w_1 と w_2 の類似度が高く，w_1 の類似単語と w_2 の類似単語がほとんど同じであるとき，w_1 と w_2 は 2 つの領域においてほぼ同一の意味で用いられていると推測できる．その領域や単語によって，領域の異なる同一単語の単語ベクトルにどのような特徴や違いが存在するか比較・分析を行う．

３．実験

本節では，単語ベクトルを用いて単語の意味的な用法が領域ごとにどのような違いや相違があるかについて調査するために，日本語文書集合を対象とした実験を行う．

３.１　実験データ

本研究における実験用の文書データには，国立国語研究所が開発した現代日本語書き言

[1] http://taku910.github.io/mecab/
[2] https://code.google.com/p/word2vec/

葉均衡コーパス(BCCWJ)を利用する．BCCWJ は日本語の様々なジャンルの文書を収録した，書き言葉の全体像を把握するために構築されたコーパスである．今回の実験では「新聞」「雑誌」「書籍」の 3 つのジャンルを領域として，これらの領域に含まれる文書データを使用する．

３.２　対象単語

名詞単語である「日本」「市場」「円」「政策」「口紅」「表明」「凹凸」「協力」，動詞である「出す」「見せる」「成る」，形容詞である「早い」「少ない」「可愛い」「憂い」，副詞である「きらきら」「わくわく」の全 19 単語を対象単語として，それぞれの領域で単語ベクトルの比較を行う．それにより，対象単語の意味を調査する．また，同一単語であっても領域による違いが発生しているか，調査を行う．

３.３　実験データの準備

文書集合から word2vec に入力可能な単語列への変換を行う．文書集合に対して mecab を用いて形態素解析を行って単語の基本形に分割する．単語列への変換を行った各領域（書籍・新聞・雑誌）の文書中の対象単語に，書籍領域の「日本」であれば「日本_b」というように，領域ごとにタグ付けを行う．書籍領域の場合は「単語_b」，新聞領域の場合は「単語_n」，雑誌領域の場合は「単語_m」とする．その後，単語ベクトルの類似度算出の際の精度向上のため，各領域の文書データを 10 回繰り返し，これを併合コーパスとする．

３.４　単語ベクトルの生成

2.4 節で説明した word2vec を用いて文章データから単語ベクトルを生成する．word2vec のモデルには，Countinuous Bag-of-Words（CBOW）モデルを利用する．次元数は 300 とし，文脈長は 5 である．反復回数は 20 とする．表 1 に，word2vec の訓練オプションの一覧を示す．

表 1　word2vec で指定した訓練オプション

Skip-gram or CBOW	-cbow	1
次元数	-size	300
文脈長	-window	5
負例サンプリング	-negative	5
階層化ソフトマックス	-hs	0
最低頻度閾値	-sample	1e-3
単語最低出現回数	-min-count	2
反復回数	-iter	20

４．実験結果

各領域に分けた対象単語に対して，word2vec で得られた単語ベクトルが類似する上位 10 単語を表 2 に示す．表 2 の結果を見ると，多くの対象単語において，類似する単語の上位に異なる領域の同一単語が現れた．

まず，各対象単語について類似する単語を分析する．「日本」では，どの領域においても類似単語はおおむね他国名であっが，新聞領域においては聖書の登場人物である「ヤペテ」や島の名前である「トカラ」なども現れた．「市場」では，すべての領域において共通する類似単語は「企業」しか存在しなかった．しかし，各領域の単語ベクトルは互いに類似していた．また，新聞分野では価格に関わる単語が多く類似単語として出現した．「円」では，各領域とも価格の単位として多く使われていることがわかる．また，類似単語に距離や重さなど，数値と共起する他の単位も出現した．「政策」では，別領域の同一単語がどの領域においても最も類似する結果となった．しかし，書籍領域においては主に作戦という意味で用

いられている一方で，新聞・雑誌領域では国の政治情勢に関わる単語が類似単語として多く登場した.「口紅」では，3 領域ともタグの異なる同一単語が類似単語として表れなかった.

表 2:対象単語の単語ベクトルと類似した単語ベクトル

対象単語と領域	単語ベクトルが類似する上位 10 単語
日本_b	日本_m，アメリカ，日本_n，中国，ドイツ， 託い讒ず，エトルリア，アラビア，外国，戦前
日本_n	日本_m，日本_b，中国，韓国，ヤペテ， イラク，トカラ，タイワン，ロシア，外国
日本_m	日本_n，日本_b，フランス，ヨーロッパ，アメリカ， 中国，韓国，ドイツ，写声，ロシア
市場_b	市場_n，企業，システム，競争，取り引き， 化，市場_m，需要，構造，銀行
市場_n	市場_b，市場_m，利下げ，企業，大手， 貿易，下落，じり高，軟調，買戻
市場_m	市場_n，市場_b，経済，デフレ，企業， 大手，株安，資本，グローバル，集権
円_b	円_m，円_n，ドル，トン，￥，キロリットル， グラム，ルピア，キロメートル，カトゥン
円_n	円_m，円_b，ドル，￥，トン， ルピア，カトゥン，株，キロリットル，件
円_m	円_n，円_b，￥，メートル，カトゥン， ドル，バーツ，CC，ルピア，グラム
政策_b	政策_n，政策_m，財政，施策，措置， 緊縮，路線，策，経済，戦略
政策_n	政策_b，政策_m，対中，大綱，戦略， 党内，制裁，対日，機構，構想
政策_m	政策_n，政策_b，対中，枠組み，対米， 税制，対日，国策，大綱，外交
口紅_b	リボン，ピンク，チーク，パウダー，ウール， 同系，裏地，ラメ，ブラウス，コート
口紅_n	ドリトス，衣料，子馬，特売，押麦， 茶粥，プラスチック，後払い，ステンレス，コーンビーフ
口紅_m	前著，パープル，茶系，パール，マフラー， イミテーション，メロー，ラメ，風姿，重修
ライン_b	ライン_m，スクエア，カホウ，ライン_n，シャフト， ハンドル，オン，ライン_b，ルート，エンド
ライン_n	カホウ，勝敗，プル，ライン_b，ディフェンス， スクエア，エイボン，終 盤，CK，パリーグ
ライン_m	シルエット，フレア，ウエスト，ピン，フェース， トーン，ベージュ，タイ ト，ベルト，毛先

戦争_b	戦争_n，戦争_m，侵略，内戦，日露， 戦役，終結，冷戦，占領，敗戦
戦争_n	戦争_b，戦争_m，テロ，イラク，終結， 報復，侵略，内戦，冷戦，蜂起
戦争_m	戦争_n，内戦，戦争_b，チュウトウ，湾岸， 大戦，敗戦，日露，侵略，事変
凹凸_b	紋様，飾り，タペスト リー，取々，白磁， 袋帯，兜，花弁，陶器，水差し
凹凸_n	木目，円錐，中空，新粉，金具， 突起，背面，　色調，生え揃う，太め
凹凸_m	最深，ビジター，聞き 違い，流し台， モーテル，軟式，ポトス，堅木，湾奥，インサイド
表明_b	表明_n，非難，堅持，看過，表明_m， 吐露，容認，　主張，否定，払拭
表明_n	表明_b，言明，堅持，表明_m，了承， 先送り，明言，辞任，首班，会談
表明_m	首班，打倒，宥和，アラファト，施政， 表明_n，辞任，表明_b，蔵相，陸相
協力_b	協力_n，協力_m，連携，支援，賛同， 援助，啓蒙，尽力，提言，友好
協力_n	協力_b，協力_m，連携，支援，交流， 協定，承認，協調，提言，合意
協力_m	協力_b，協力_n，無償，援助，JICA， ソーシャル，有償，アナリスト，依頼，ワーカー
出す_b	出す_m，出す_n，出る，書く，聞く， 入れる，渡す，持つ，張り上げる，流す
出す_n	出す_m，出す_b，出る，引き出す_m，減らす， 入れる，注ぐ，見せる_n，上げる，合わせる
出す_m	出す_n，出す_b，出る，渡す，見せる_m， 入れる，使う，付ける，揃える，残す
成る_b	成る_m，成る_n，有る，言う，分かる， 思う，見える，為る，出来る，考える
成る_n	成る_m，成る_b，入る，繋がる，思う， 言う，為る，出る，陥る，付く
成る_m	成る_n，成る_b，思う，見える，入る， 付く，分かる，出る，変わる，言う
見せる_b	見せる_n，見る，見せる_m，教える，感ずる， 与える，伝える，せる，広げる，引き立てる
見せる_n	見せる_m，見せる_b，見る，面変わり，語る， 決める，見せ付ける，知る，熟す，広げる
見せる_m	見せる_n，見せる_b，見る，感ずる，教える， 付ける，与える，出す_m，覚える，変える

憂い_b	捨身，慈悲，菩提，情，候， ぬ，一揆，少弐，至り，ショウスケ
憂い_n	大意，憂い_m，穢，禾，小猿， エキケン，今様，霊言，舌舐り，スサノオ
憂い_m	ビノバープリー，黄肌，シバコウエン，タド，ヤスミ， コウシカイ，福泉，順大，カサマ，蓚酸
早い_b	早い_n，早い_m，遅い，カヅキ，良い， 起床，ない，寒い，ウタコ，近付く
早い_n	早い_b，早い_m，遅い，カヅキ，暑い， 間に合う，難しい，長い，無理，正確
早い_m	早い_n，早い_b，面映ゆい，長い，遅い， 安い，カヅキ，若い，良い，強い
少ない_b	少ない_n，少ない_m，多い，低い，思い為す， 増える，殆ど，減る，多く，有る
少ない_n	少ない_b，多い，少ない_m，増える，思い為す， せめて，多く，減る，大きい，遅い
少ない_m	少ない_n，少ない_b，多い，遅い，増える， 低い，大きい，思い為す，殆ど，難しい
可愛い_b	可愛い_m，退部，怖い，優しい，優雅， 寂しい，可愛らしい，可愛い_n，嫌，素敵
可愛い_n	可愛い_m，嬉しい，退部，嫌，可愛い_b， 良い，面白い，羨ましい，河馬，幸せ
可愛い_m	キュート，可愛い_b，嫌，優しい，ラブリー， フェミニン，可愛い_n，新 鮮，良い，甘酸っぱい
きらきら_b	文才，藤色，きら，光点，ぽっちり， ヨシウラ，土埃，銀色，ピスタチオ，垂線
きらきら_n	モンテベルディ，巡洋，セルロース，ミマン，ルンメイ， クレゾール，バーデン，尋ね求める，プログレッシブ，暴雨
きらきら_m	ゴージャス，ぱちくり，ぱっちり，ぎょろぎょろ，ほっくり， 潤む，ストーン，羽織物，愛敬，毳立つ
わくわく_b	うんざり，どきどき_m，感激，どきどき_n，狼狽， ぐずぐず，浮き浮き，ときめく，嬉々，苛立つ
わくわく_n	イレン，擽る，町歩き，しゃぎり，RU， フェスタ，触れ合い，むずかる，どきどき_n，宝恵
わくわく_m	どきどき_m，ほろり，　台無し，閉口，居た堪れる， 息継ぎ，おたおた，どぎまぎ，はらはら，まごまご

書籍領域ではファッションに関わる道具として，新聞領域では商品の一つとして，雑誌領域では化粧品として用いられている．「ライン」では，書籍領域では他の領域の同一対象単語と類似度が高かったが，書籍以外の 2 領域では対象単語以外の単語と類似度が高いという結果であった．「凹凸」では，3 領域とも他の領域の同一対象単語が類似単語として全く現れなかった．また，3 領域とも類似単語が共通していなかった．「表明」では，書籍領域と新聞領域が特に類似する結果となった．また，新聞領域と雑誌領域では政治に関わる単語が

類似単語として多く表れた．「協力」は，3 領域ともほとんど同じ意味で用いられ，「援助」「連携」「支援」といった単語が複数の領域で類似単語として出現した．

「出す」は，3 領域ともほとんど同じ意味で用いられる傾向があった．「出る」「入れる」のように 3 領域共に類似単語として表れた語句も存在した．「成る」は 3 領域ともほとんど同じ意味で用いられていた．「見せる」では，3 領域ともほとんど同じ意味で用いられていただけではなく「見る」との類似度も高い結果となった．

「憂い」は，書籍領域では感情表現として，新聞領域ではネガティブなイメージを表す語として用いられていた．雑誌領域においては，「黄肌」や「順大」など，様々なバリエーションの語句が類似単語となった．「早い」は，3 領域ともほとんど同じ意味で用いられていた．また，「遅い」との類似度がどの領域においても高く，時間の単位として多く用いられている．「少ない」は 3 領域ともほとんど同じ意味で用いられていた．「多い」や「増える」など数量を表す単語が，3 領域の類似単語となった．「可愛い」では，書籍領域においては人柄を表す語句が類似単語として多く見られた．新聞領域では，感情表現を表す単語が，雑誌領域では愛らしいという意味の単語が多く見られた．その他，3 領域とも別領域の同一対象単語が類似単語として表れなかった．「きらきら」では，3 領域とも別領域の同一対象単語が類似単語として表れなかった．また，各領域ともカタカナ語の多く類似単語として出現した．「わくわく」では，3 領域とも別領域の同一対象単語が類似単語として表れなかった．また，書籍領域，雑誌領域においては「どきどき」といった畳語が類似していた．

次に，各領域の特徴を述べる．書籍領域に出現した類似単語は，様々な種類の語句が登場していた．新聞領域に出現する類似単語は，政治に関わる単語が他領域に比べ多く登場した．「憂い」「凹凸」「可愛い」など，雑誌領域の類似ベクトルにカタカナ語が他領域と比べ多く出現した．

５．考察

実験結果より，word2vec を用いた実験ではほとんどの単語は，領域の異なる同一対象単語との類似度が最も高くなるか，それに近い結果となった．しかし，全対象単語 19 単語のうち，「口紅」「凹凸」「きらきら」「わくわく」の 4 単語においては，類似する 10 件の単語ベクトルに異なる領域の同一単語は表れず，別領域の同一対象単語との類似度が高い結果となった．例えば，「口紅」では，書籍においては他の化粧品との類似度が高い一方，新聞では広告としての関連語句の類似度が，雑誌では種類や色味などの関連が強い傾向にあった．これは，書籍では数ある化粧品の種類の一つとして，新聞では商品として，雑誌では唇に塗布する化粧品の総称としてというように，各領域において異なる「もの」として扱われていることを示す．動詞である「出す」「見せる」「成る」は，どの領域においてもほぼ同じ意味で用いられていた．動詞の語義は領域への依存度が低いと考えられる．「きらきら」「わくわく」はどちらも副詞であり，この結果から副詞の語義は特に領域に依存すると推測できる．また，「きらきら」と「わくわく」はともに畳語であるが，「わくわく」の類似単語には畳語が多くあらわれたが，「きらきら」にはほとんど現れることはなかった．書籍領域の対象単語に対する類似単語は，様々な種類の語句が出現した．これは，書籍領域では他領域と比べ多様な語句が用いられているからだと考える．また，新聞領域では他領域と比べ政治的意味を持つ単語が，対象単語の類似単語として比較的多く登場した．新聞領域では社会情勢を示す記事が含まれていることが理由であると思われる．「憂い」「凹凸」「可愛い」の 3 単語においては，雑誌領域の類似ベクトルにカタカナ語が他領域と比べ多く出現した．この結果は，雑誌領域では他領域と比べカタカナ語が多く用いられていることを表していると推測できる．

６．結論

本稿では，word2vec を用いて日本語を対象とした単語ベクトルを使用し，領域の違いによる単語の特徴や類似度の分析を行った．実験の結果，同一単語であっても他領域における

意味と異なる意味で使用した単語が存在することがわかった．加えて，動詞は領域の依存度が低い，副詞は領域への依存度が高いなど，品詞によって領域の依存度が異なる傾向があった．また，書籍領域では様々な種類の語句，新聞領域では政治関連語句，雑誌領域においてはカタカナ語が多く登場するなど，類似した単語には領域によって特徴がみられた．この結果は，語義は領域に依存することを示している．

今後は，さらに多くの単語に対し，領域を増やして単語ベクトルの比較や分析を行うことが課題である．今回の実験では，3 領域中の 19 単語を用いることで，各領域における語義の分散表現の特徴を分析した．BCCWJ にある白書など領域を追加し，対象単語を増やすことで各単語の意味の分かれ方や領域による差異がよりわかりやすくなると考えられる．

文　献

Tobias Schnabel, Igor Labutov, David Mimno, Thorsten Joachims (2015). “Evaluation methods for unsupervised word embeddings”, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.298-307.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. (2013) "Distributed Representations of Words and Phrases and their Compositionality", Proceedings of the 26th International Conference on Neural Information Processing Systems, pp.3111-3119.

関連 URL

『現代日本語書き言葉均衡コーパス』　http://pj.ninjal.ac.jp/corpus_center/bccwj/