P09

自然言語処理手法を用いたクックパッドデータの構造解析

川上瑛之　伊藤浩之　京都産業大学先端情報学研究科

概要

本研究では，国立情報学研究所からクックパッドデータの提供を受け，レシピ名単語とそれに付随する材料名単語を並べたコーパスを作成し，CBOWモデルにより前後10個の単語からの単語の予測の学習を行った。学習の過程で埋め込み層に自動的に構成される料理単語の高次元ベクトル表現をword2vecとして用いて，料理単語間のベクトル演算を行ったところ，ある程度妥当な結果が得られたことから高次元ベクトル空間に料理単語間の関係性が構造化されていることを確認した。この構造を更に解析するために，様々な次元圧縮法により高次元ベクトル空間を次元圧縮し，視覚化を行ったところUMAPにおいて類似した単語が近接して分布するクラスター構造が得られた。また，高次元ベクトル空間上のレシピ名単語をK-meansクラスタリングを用いた教師なし学習で分類した結果とクックパッドデータで各レシピ名単語に付与されている経験的なレシピ分類との対応の整合性の解析を行った。「お肉のおかず」，「野菜のおかず」，「サラダ」のカテゴリー分類に整合が見られ，さらに「お菓子」のカテゴリーに関しては教師なし学習は更に細分化した分類を提供することが示された。この結果は一般的な要素列データが与えられた時に，経験的分類とは異なる新たな分類法を機械学習が提供する可能性を示唆すると考える。

研究背景

英語や日本語などの自然言語データを対象とした単語間の関係性の解析は数多く為されてきた。

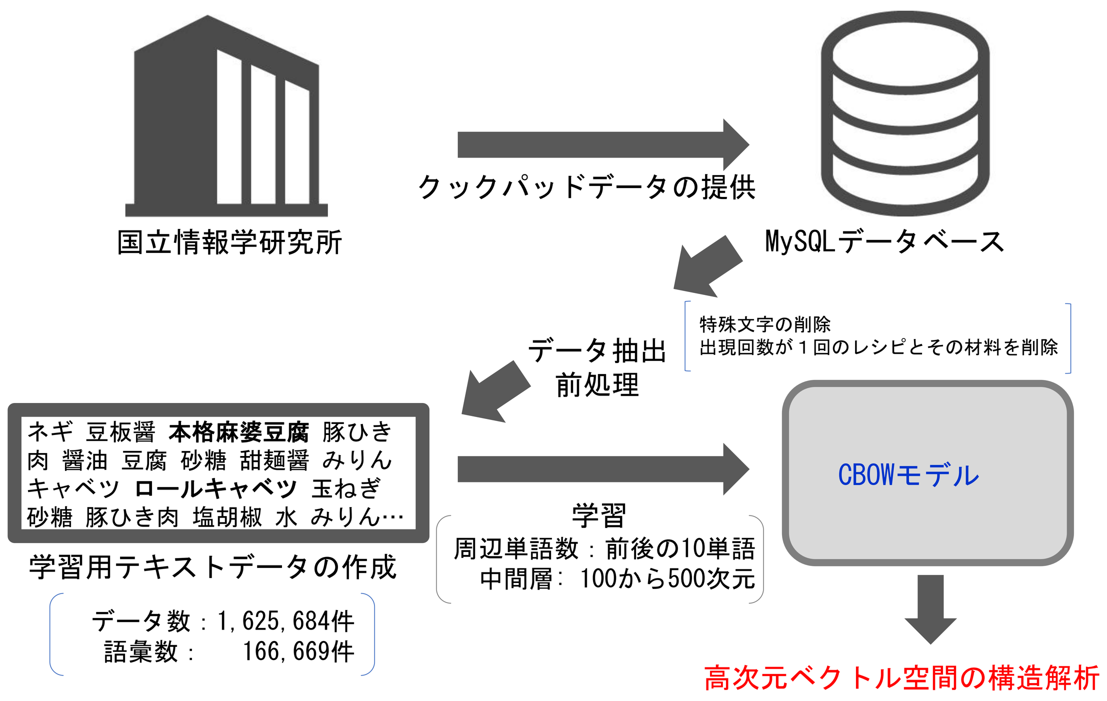
word2vecは単語の意味空間での位相構造を再現する

$\vec{king} - \vec{man} + \vec{woman} = \vec{queen}$

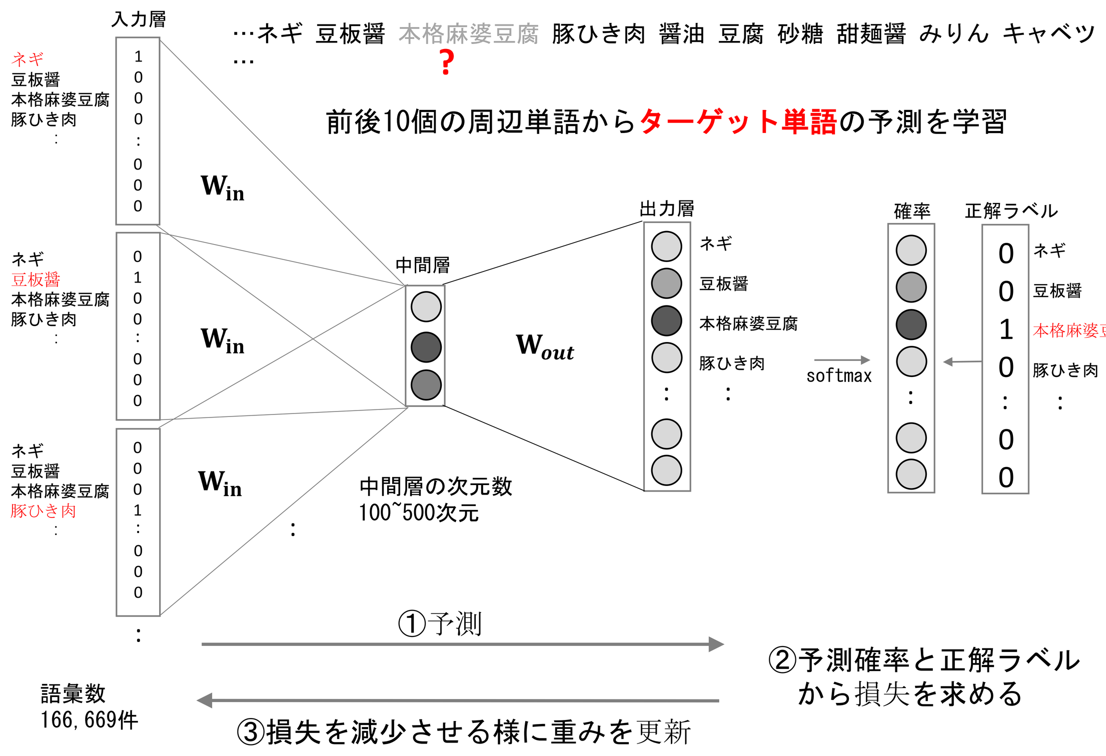
自然言語以外のデータでword2vecを行った際にも単語間の関係性を解明できるか？

➡ 料理データ（クックパッドデータ）を対象として検証する。

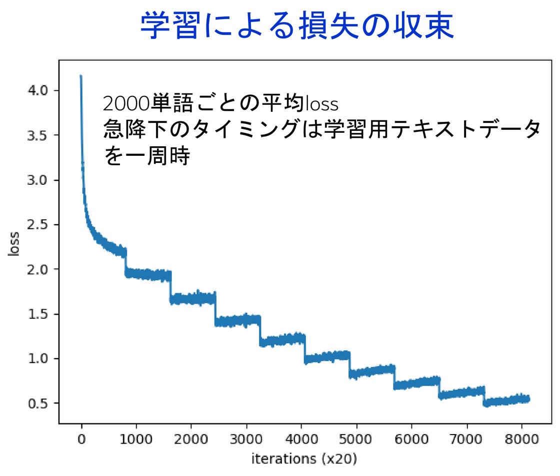
研究方法の全体像



CBOWモデルでの学習

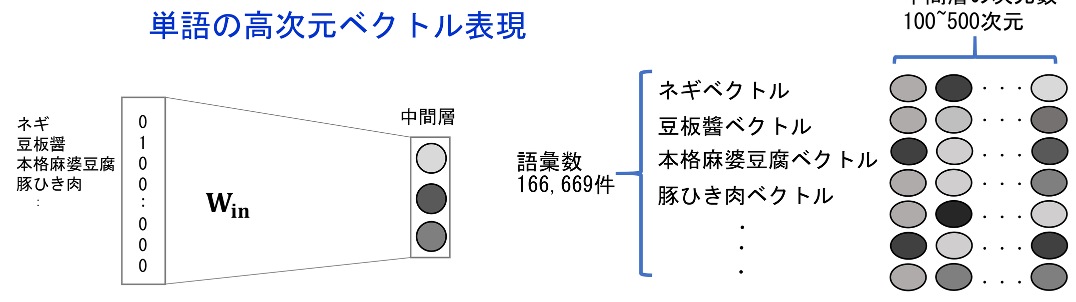


学習による損失の収束



データ数：1,625,684件
語彙数：　166,669件
中間層の次元数：300次元
周辺単語数：前後10単語（計20単語）

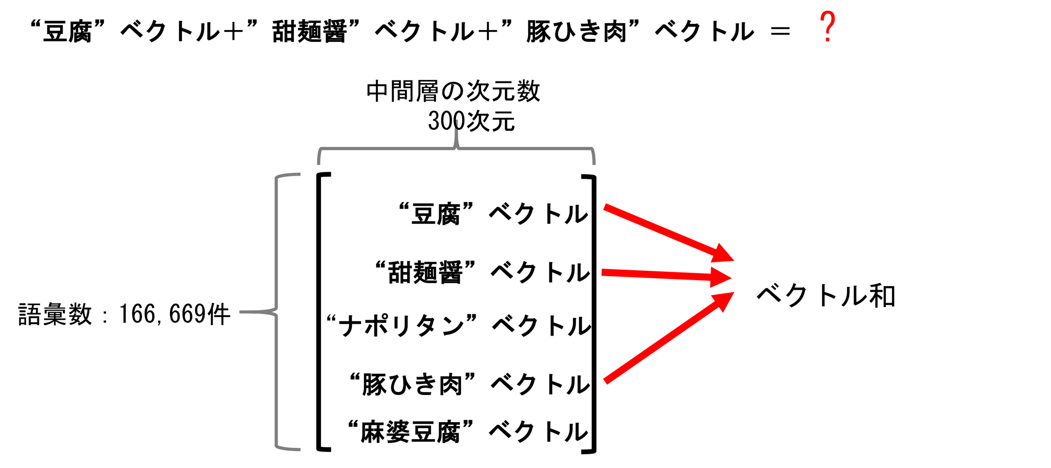
単語の高次元ベクトル表現



結果１：単語ベクトル間の演算

① 加算ベクトルの作成

“豆腐”ベクトル＋”甜麺醤”ベクトル＋”豚ひき肉”ベクトル ＝ ？



② 加算ベクトルに（内積orコサイン類似度）が近い上位10件のレシピ名単語のみを抽出

単語演算の結果

“豆腐”＋”甜麺醤”＋”豚ひき肉” = “？”

内積による類推（レシピ名のみ）	コサイン類似度による類推（レシピ名のみ）
ジャージャー麺: 3.94921875	担々麺: 0.352783203125
担々麺: 3.7578125	ジャージャー麺: 0.3359375
エビチリ: 3.654296875	ピリ辛肉味噌: 0.335693359375
肉味噌: 3.3828125	白菜のトロトロ煮: 0.328125
回鍋肉: 3.357421875	肉味噌: 0.32568359375
ピリ辛肉味噌: 3.251953125	汁なし担々麺: 0.3154296875
中華風コーンスープ: 3.052734375	チャーシュー麺: 0.31396484375
回鍋肉ホイコーロー: 3.0234375	万能ねぎダレ: 0.31201171875
麻婆豆腐: 3.005859375	担々麺: 0.30615234375
白菜のトロトロ煮: 2.98828125	担担麺: 0.294921875
ゴーヤの肉詰めフライ: 2.982421875	つくね: 0.291748046875
ラーメン: 2.947265625	味噌麻婆豆腐: 0.290771484375
中華風にゅうめん: 2.94140625	納豆のお揚げ包み焼き: 0.285400390625
肉みそ: 2.896484375	我が家の野菜たっぷり餃子: 0.283447265625
担々麺: 2.890625	エビチリ: 0.283203125

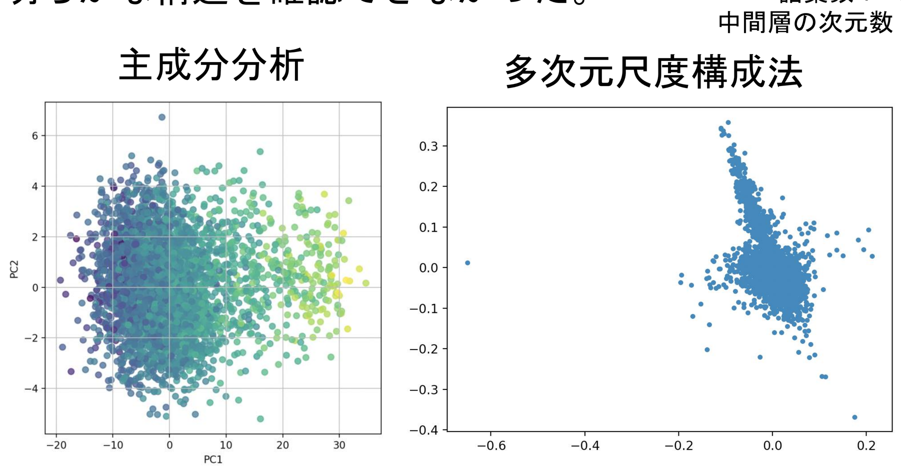
料理単語による演算結果は妥当に思える。
高次元ベクトル空間に何らかの関係性が反映されている。

➡ 高次元ベクトル空間（300次元）を2次元や3次元に次元圧縮して関係性の可視化を試みる。

結果２：次元圧縮による構造の視覚化

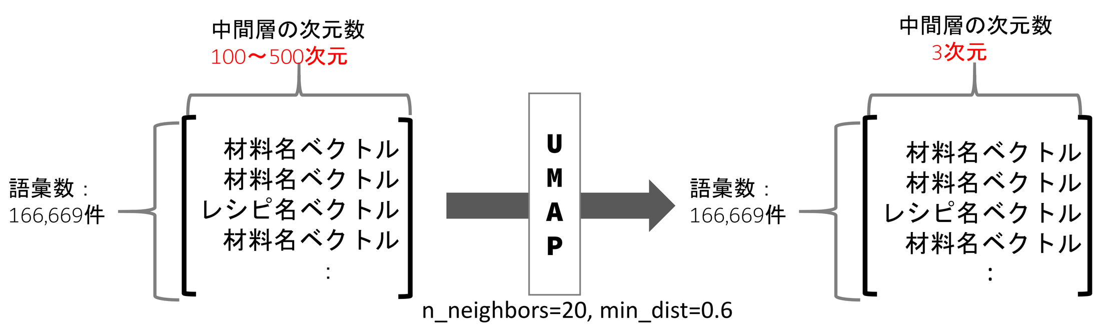
主成分分析と多次元尺度構成法では、クックパッドデータの明らかな構造を確認できなかった。

語彙数：　166,669件
中間層の次元数：300次元

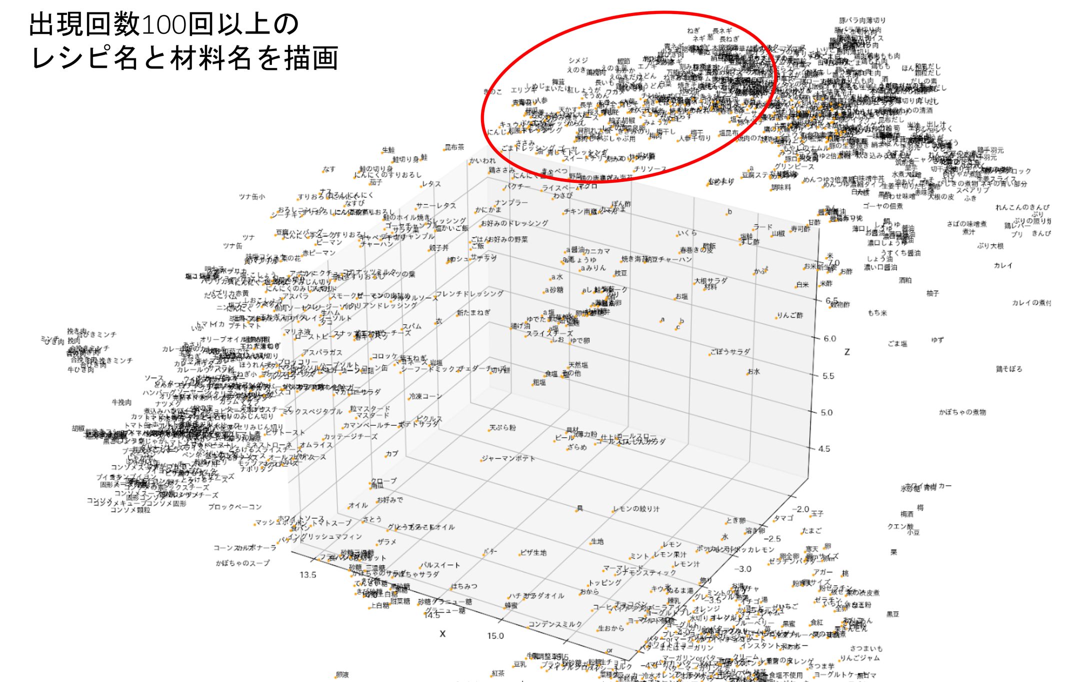


有効な次元圧縮による関係性の視覚化

クックパッドデータの学習で得られた高次元ベクトル空間をUMAPで低次元へ圧縮することで関係性の視覚化を試みる



出現回数100回以上のレシピ名と材料名を描画



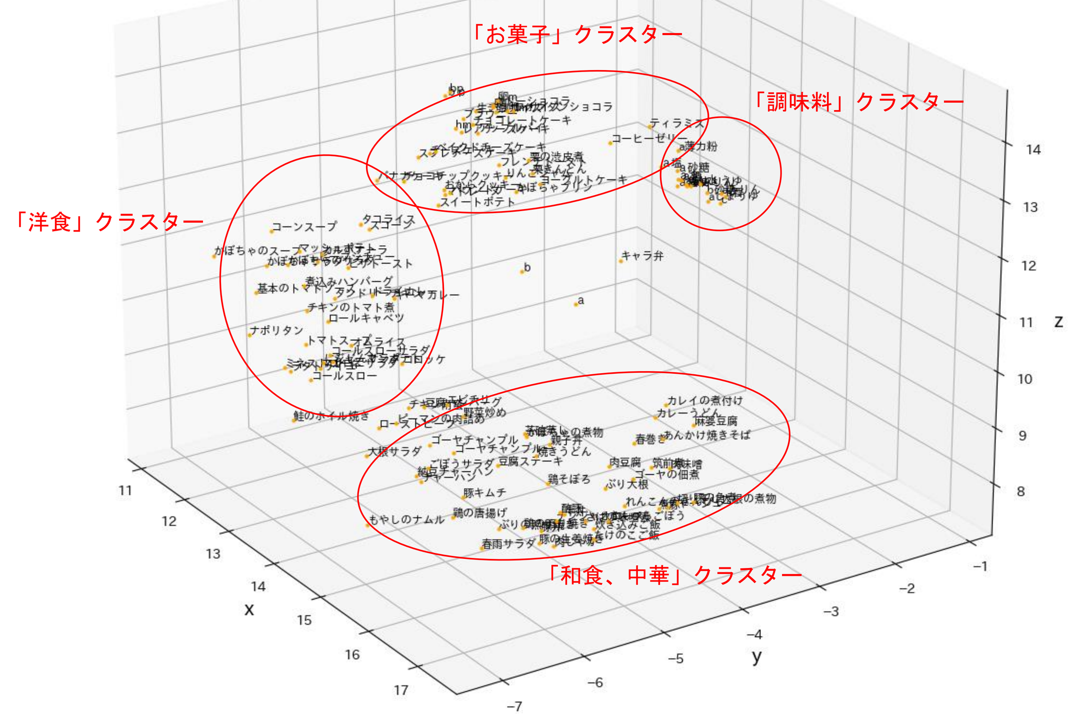
データ数：1,625,684件
語彙数：　166,669件
中間層の次元数：100次元
周辺単語数：前後10単語（計20単語）

類似した材料がクラスター化している

レシピ名のカテゴリー化

出現回数100回以上のレシピ名のみを描画

データ数：1,625,684件
語彙数：　166,669件
中間層の次元数：300次元
周辺単語数：前後10単語（計20単語）



ここまでの結論と次なる課題

料理データのCBOW学習によって得られた高次元ベクトル空間の構造をUMAPによって次元圧縮・視覚化したところ、何らかの関係性が抽出されていることが確認できた。

UMAPが視覚化したデータ点をK-meansクラスタリングにより教師なし分類した結果と我々が経験的に行う分類がどの程度合致しているのかを調査する。

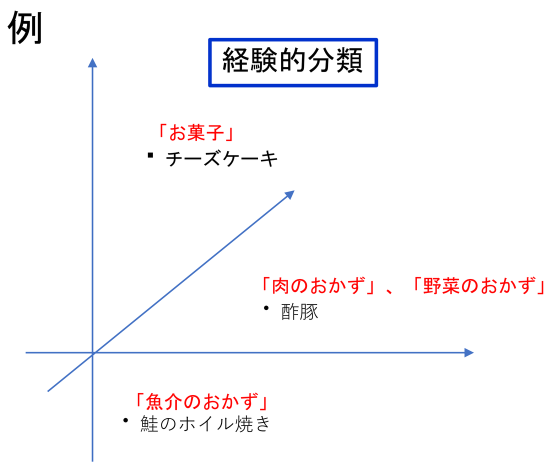
レシピ単語へのカテゴリーラベル付け

我々が経験的に行っている料理の分類カテゴリとして、クックパッドサイトの”レシピカテゴリ”を元に、各レシピに紐付けされている料理カテゴリを分類カテゴリとして付与した。

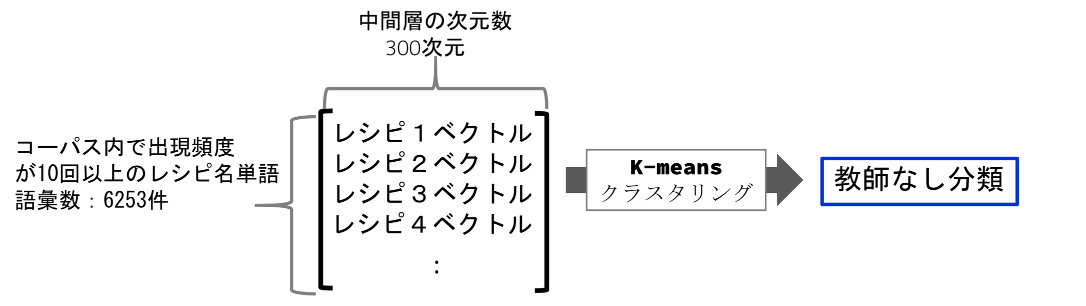
レシピカテゴリ

などの20種類

例

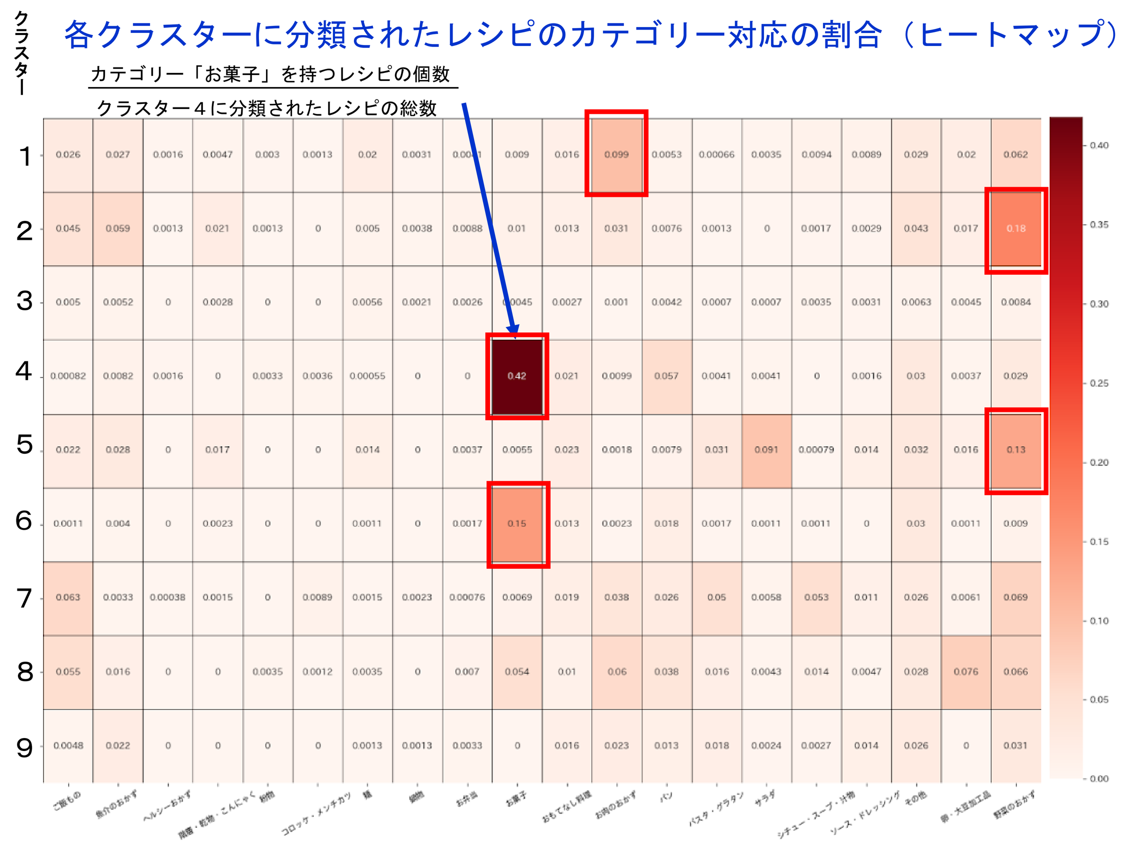


結果３「教師なし分類」と「経験的分類」の整合性



クラスターの分離度合の指標(CH基準値）はクラスター数９で最大となった

各クラスターに分類されたレシピのカテゴリー対応の割合（ヒートマップ）



クラスター1に分類されていたレシピ名単語の一部

['うま味調味料', '豚丼', 'aオイスターソース', '豚肉のケチャップ炒め', '豚切り落とし', '鶏ガラスープの素顆粒', '調味料a', '調味料b', 'もずくスープ', '豚キムチ', '豆腐ハンバーグ', 'キムチ鍋', '麻婆豆腐', '鮭のちゃんちゃん焼き', 'ビビンバ', 'トンカツ', 'しょ売の皮', 'おろし生姜チューブ', 'タンドリーチキン', 'そぼろ', '鶏スープ', 'チンジャオロースー', 'ニンニク生姜', '肉の下味', '土しょうが', '卵mサイズ', 'かぼちゃのそぼろあんかけ', '中華そば', 'チンゲンサイ', '豆乳鍋', '餃子鍋', 'チキンスープの素', 'さつまい南蛮漬け', 'にゅうめん', 'タコライス', 'もやし炒め', '野菜の肉巻き', 'ふき味噌', '肉豆腐', '米茄子', '豚肉の味噌焼き', '納豆キムチチャーハン', 'ショウガ汁', '豚バラかたまり', '焼きうどん', '肉味噌', 'ニンニクすりおろす', '生姜すりおろす', '豚の生姜焼き', '我が家の春巻き', '竹の子水煮', 'アミの塩辛', '豚キムチうどん', 'チャーシュー丼', 'チャーシュー', '鶏の手羽先', '肉そぼろ', 'ワンタンスープ', '鶏がら', 'ワンタン', 'エビチリ', '＜合わせ調味料＞', 'アップルパイ', 'カツオのたたき', '顆粒鶏ガラスープ', '豆鼓ムチ炒め', '中華風スープ', 'ゆで玉子', '春雨炒め', 'うどん玉', '我が家の餃子', '豚かたまり肉', '豚肉のしょうが焼き', '豚肉こまぎれ', '芝麻醤', '餃子のたれ', 'あんかけ焼き

肉のおかず

クラスター２に分類されていたレシピ名単語の一部

['ひじきの炊き込みご飯', 'カレイの煮付け', 'きのこご飯', '切り昆布の煮物', '粉末ダシ', 'タケノコご飯', '松茸ご飯', '松茸', 'ひじき煮', '山菜の水煮', '豚の角煮', '葱の青い部分', 'ぶり大根', 'もずく', 'じゃがいものきんぴら', '大学いも', '筍の土佐煮', '煮干ん草のお浸し', 'のりの佃煮', 'きんぴらごぼう', '手羽先と大根の煮物', '鶏そぼろ', 'ほうれん草の白和え', '肉うどん', '牛こま切れ', 'シナモンあれば', '七味とうがらし', 'つくしの佃煮', 'つくし', '大豆粉', 'ボイルホタテ', 'かぼす', 'かぼちゃの煮物', 'ひじきご飯げんの胡麻和え', 'ブリの照り焼き', 'かつおパック', '練り胡麻', 'わらび', '白菜漬け', '揚げ', '豆腐そぼろ', 'あぶらあげ', '大根の皮のきんぴら', 'いりこ', 'アジの南蛮漬け', '栗おこわ', '鶏のさっぱり煮', 'ゴーヤの佃煮', '豚バラ肉スライス', 'イカと里芋の煮物', '筑前煮', '大学芋', '炒り豆腐', '豆ご飯', 'さばの味噌煮', 'さばの切り身', 'いか大根', '韓国風肉じゃが', 'プレーンスコーン', '鯛めし', '大豆とひじきの煮物', '筍ご飯', '田作り', 'いかと里芋の煮物', '生姜ご飯', 'サバの味噌煮', '里芋とイカの煮物', 'ぶりあらの煮物', '顆粒出汁', '茹で筍', 'ひじきと大豆の煮物', 'ひじきドライパック', '大豆ドラ

野菜のおかず

クラスター５に分類されていたレシピ名単語の一部

['イタリアンサラダ', 'はちみつレモン', '大根サラダ', 'ゆずこしょう', '甘酢しょうが', 'バジリコ', 'ブロッコリーの胡麻和え', '和風パスタ', 'チキンサラダ', 'サラダ', 'パンプキンサラダ', '飲むヨーグルト', 'グリーンレタス', 'さつまいもご飯', 'かつおのたたき', 'マグロのカルパッチョ', 'サーモンマリネ', '生春巻き', 'さっぱりサラダうどん', 'ライスサラダ', 'マヨネーズ塩コショウ', 'たらこスパ', '新玉ねぎサラダ', 'もやしのナムル', '生わかめ', 'さつまいものサラダ', 'マカロニサラダ', '切り干し大根のサラダ', 'レアチーズ子漬け', 'いくらの醤油漬け', 'ほうれん草のツナ和え', '練りごま白', 'ヘルシーポテトサラダ', '釜揚げしらす', 'ツナと大根おろしのパスタ', 'ビネガー', 'さつまいもチップス', 'かにカマ', 'マグロ刺身', '梅ぼし', 'アボカドとトマトのサラダ', 'きざみにんにく', '鮭のび', 'マグロの漬け丼', 'リーフレタス', 'グリーンサラダ', 'ホタテのカルパッチョ', '薄焼き玉子', 'かに風味かまぼこ', 'カルピス', '出しの素', 'しゃぶしゃぶ用豚肉', 'わかめときゅうりの酢の物', '玉ねぎドレッシング', '味の素無くても可', 'レッドオニオン', 'ひじきとアボカドのサラダ', 'アボカドグラタン', '春雨の中華風サラダ', 'スライス玉ねぎ', '酢砂糖', 'タラコ', '桜でんぶ', '春雨サラダ', '塩パスタ茹でる用', 'ほうれん草のサラダ', 'イカのマリネ', 'ゴーヤサラダ', 'ゴーヤのナムル', 'アボガドサラダ', '

野菜のおかず
サラダ系

クラスター４に分類されていたレシピ名単語の一部

['低脂肪乳', '基本のマフィン', 'ショートブレッド', 'ガトーショコラ', 'ｂｐ', '抹茶パウンドケーキ', 'りんごケーキ', 'バナナスコーン', 'チョコケーキ', 'チョコバナナパウンドケーキ', 'マーブルケーキ', 'bp', 'チーズケーキ', 'バター好きのためのマーガリン', 'ケーッケ', 'チョコスコーン', 'アーモンドクッキー', 'ホットケーキミックスでバナナケーキ', '基本のスコーン', 'さんまの蒲焼', 'かき揚げ', 'ココアクッキー', 'りんごのケーキ', 'スイートポテト', '紅茶のパウンドケーキ', 'チキンナゲット', 'バナナケーキ', 'スコーン', 腐ドーナツ', 'チョコレートシフォンケーキ', 'パンプキンタルト', 'かぼちゃクリーム', 'パートシュクレ', 'りんごのマフィン', 'オートミールクッキー', 'えびマヨ', '牛乳か豆乳', 'チョコチップマフィン', 'プリンケーキ', 'ロールケーキ', '抹茶シフォンケーキ', 'シューパン', 'ココナッツサブレ', 'さくさくクッキー', 'フォンダンショコラ', '豆乳か牛乳', 'ブラックココア', 'ブロッコリーとゆで卵のサラダ', '基本のパン生地', 'バナナマフィン', 'スイートポテトパイ', '生イースト', 'アイスボックスクッキー', 'プレーン生地', 'ココア生'牛乳または水', 'チョコチップスコーン', 'リンゴケーキ', '太刀魚', 'キャロットケーキ', '米粉パン', 'ピーナッツバター無糖', 'ねぎ焼き', '揚

お菓子
パン・ケーキ
小麦粉系

クラスター６に分類されていたレシピ名単語の一部

['あさりのワイン蒸し', '苺ジャム', 'マンゴー缶', 'マンゴーピューレ', 'ミルクプリン', 'スイートチョコ', '仕上げ用粉糖', '型に塗るバター', '桃のコンポート', 'ココナッツミルクパウダー', 'オレンジマフィン', '焼きりんご', 'グレープフルーツゼリー', 'グラハムパン 'サイダーゼリー', '赤紫蘇ジュース', 'ﾖｰｸﾞﾙﾄ', '栗の渋皮煮', '牛スジ', '梅味噌', 'パイナップルジュース', 'キャラメルポップコーン', 'いかなごのくぎ煮', '〜ソース〜', '餡子', '海塩', 'スダチ', 'チョコレートチーズケーキ', '新生姜の甘酢漬け', 'ゴルゴンゾーラ'イチジク', 'ヨーグルト加糖', 'オレンジマーマレード', '溶かしバター無塩', '完熟梅', 'チョコレートシロップ', 'ポッキー', 'フルーツヨーグルト', '黄桃', 'ロール生地', 'ヨーグルトムース', '水きりヨーグルト', '100%オレンジジュース', 'ヨーグルトパフェ', 'ダノンビ', '抹茶粉末', '紅茶豚', '加糖ヨーグルト', 'ハイジの白パン', 'アレンジ', '塩水', 'ぜんざい', 'ほうじ茶', '好きなフルーツ', 'ｸﾘｰﾑﾁｰｽﾞ', 'ﾚﾓﾝ汁', '台', 'オレオクッキー', 'マンゴープリン', '脱脂粉乳', '生チョコ', 'レモンシャーベット', '卵m', 'きんかん', 'お好ホール', '梅ジャム', 'リッツ', '鶏卵', 'りんごジャム', '中身', 'ローマジパ

お菓子
その他

まとめ

自然言語処理手法であるword2vecを用いて，クックパッドデータの料理単語を高次元ベクトル空間へ埋め込み，データを構成する料理単語間の関係性をUMAPにより次元圧縮して視覚化することができた。

高次元ベクトル空間を K-meansクラスタリングにより教師なし学習で分類したクラスターと我々が経験的に行っているレシピの分類では一部で対応が見られることや新たな分類の可能性を確認できた。

料理データに限らず，要素間の関係性が未知のデータ列に対して本研究と同様なアプローチを行うことにより，我々の経験とは異なる新たな分類を発見する可能性があると考える。

自然言語処理手法を用いたクックパッドデータの構造解析
情報処理学会研究報告　Vol.2023-IFAT-149 No.1 2023/2/27