食品単語のベクトル空間の構築とその評価(第 2 報)

Construction of a dense vector space for food category words and its evaluation (2nd report)

○矢野 達也　林 豊洋　大橋 健
Tatsuya Yano　Toyohiro Hayashi Takeshi Ohashi
九州工業大学
Kyushu Institute of Technology

n238070t@mail.kyutech.jp

Abstract

自然言語処理分野の 1 つのアプローチとして word2vec が注目されている。word2vec は、ニューラルネットワークの学習に基づき、与えられた文章から単語のベクトル表現を生成する。本研究では、類似する食品単語に対して密なベクトル空間の構築とその評価を目的とした。word2vec の入力コーパスは、各食品のカテゴリや原材料情報を web 上で取得し、ジェネレータにより作成した。食品のカテゴリ、原材料を学習させたコーパス 1 、原材料のみを学習させたコーパス 2 を作成し、それぞれのベクトル空間を構築した。構築したベクトル空間は k-means 法により 30 のクラスタに分類し、評価を行った。具体的には、各クラスタ間のデータの分散値、含有率の比較、各クラスタ内でのユークリッド距離、コサイン類似度による類似語検索、PCA 法による低次元圧縮上での可視化により評価を行う。結果、どちらのコーパスも単語の類似性に対して優れたベクトル空間の構築が確認できた。コーパス 1 で構築したベクトル空間では、カテゴリを学習させているため、菓子、飲料の明確な区別の中で、類似度分類が可能であった。コーパス 2 では、原材料のみ学習させたため、カテゴリの区別なく類似語の分類ができた。自作の文章をベクトル空間構築のコーパスとして使用することの有用性が検証できた。

1. はじめに

近年、家庭用サービスロボットの開発が活発である。国際的ロボット競技大会である RoboCup@Home では、家庭環境において、人間とロボットのインタラクションを評価するテストを行っている。ロボットには、柔軟な言語処理の能力が求められる。例えば、ロボットに「コーヒーを持って来て」と指示し、そこにコーヒーが無かった場合、解決策の 1 つとしてコーヒーの代替品を持ってくることが挙げられる。こういった際、類似語を機械的に取得することができれば、より柔軟な対応が期待できる。本研究では。word2vec を用いて、自作した食品単語の文章データからベクトル空間を構築し、この課題への対応を試みる。

2. 関連研究

word2vec は Tomas Mikolov らによって提唱された、単語をベクトル表現にする手法である[1]。入力として文章を与え、文章中の単語の共起関係にもとづき、予め設定していた次元数のベクトル(分散表現)を学習する。通常、学習コーパスには数十万から数百万の語彙数を持つテキストデータを使用する。文章データに含まれる語彙数が多いほど幅広い表現が可能なベクトル空間を構築できる反面、特定の分野に関しては、疎なベクトル空間になる場合が考えられる。根本らは、雑談ができる知識表現の獲得を目指して、青空文庫の文学作品を著者別に学習し、雑談対話システムへの検討を行っている[2]。そこでは、「人間」の類似単語をみると、夏目漱石の作品では「価値」、「学問」、太宰治の作品データでは、「事態」、「思想」といった単語であった。著者の思想や性格にもとづき広義の類似語結果が取得できていることが確認できる。このように、word2vec では構築したいベクトル空間に合わせた入力コーパスを用意する必要がある。

本研究では、自作の文章ジェネレータを用いて文章データを作成することで、類似食品単語の検索に適したベクトル空間の構築を行う。また、構築したベクトル空間にクラスタリング、主成分分析を施しその評価を行う。本研究の第 1 報では、菓子、飲料それぞれベクトル空間の構築とその評価を行った[3]。第 2 報では、菓子、飲料を 1 つのベクトル空間に集約し、その評価を行う。

3. 食品単語のベクトル空間の構築

3.1 提案手法の概要

本研究では、家庭用サービスロボットでの運用に焦点を絞り、構築するベクトル空間は一般家庭を想定した菓子、飲料に限定する。文章は Web から取得した食品情報をもとに、自作の文章ジェネレータにより生成する。生成された文章を word2vec のコーパスとして用いて、食品単語のベクトル空間を構築する。

3.2 データ収集

まず、文章生成に必要なデータを Web から取得

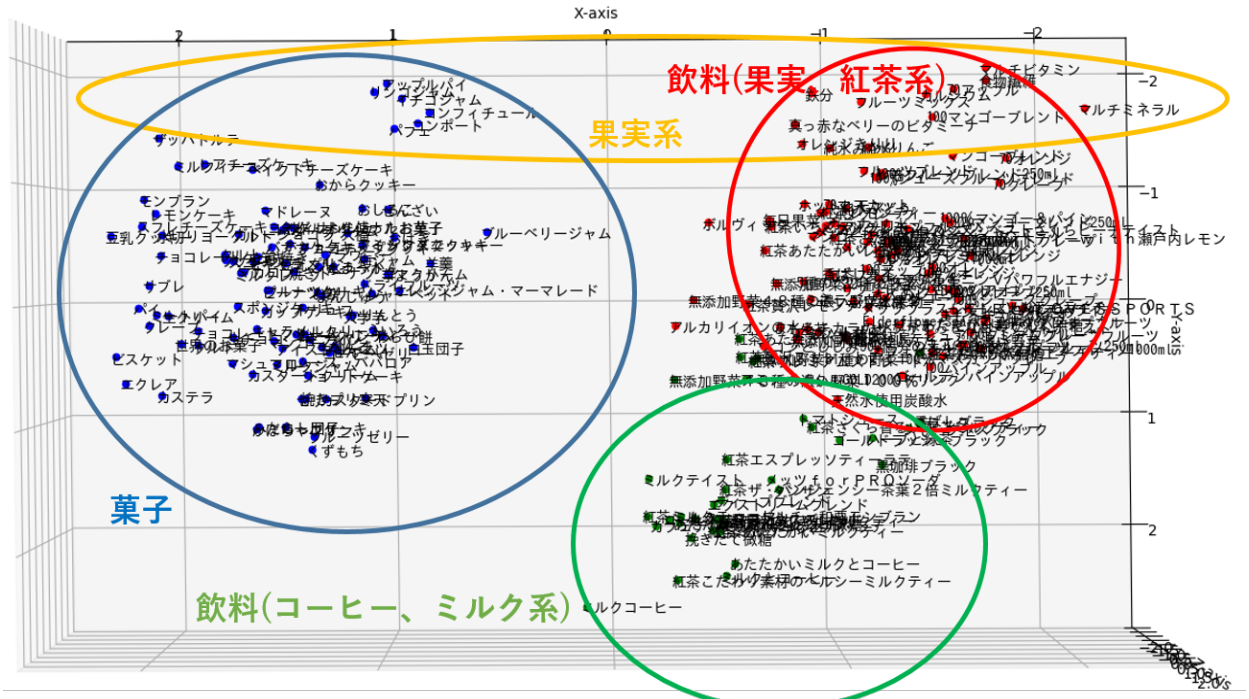

図 1. 食品単語ベクトル空間の可視化（コーパス 1）

する。菓子は楽天株式会社が運用する楽天レシピ(注 1)のレシピページから 100 品、飲料はキリン株式会社の製品一覧ページ(注 2)から 140 品を抽出し、それぞれ名前、カテゴリ（飲料または菓子）、原材料の情報を取得する。取得例として、チョコチップクッキーではカテゴリ（菓子）、原材料（薄力粉、バター、チョコ）となる。

3.3 文章生成

取得した食品情報を文章ジェネレータのテンプレート文に当てはめ、1 食品ごとに 1200 文の短文を生成する。チョコチップクッキーでの生成例を以下に示す。

1. [チョコチップクッキー] は [菓子]
2. [チョコチップクッキー] の 原料 は [バター]

第 1 報では、菓子、飲料の文章を別々にコーパスとして使用し,それぞれ独立したベクトル空間を作成した。第 2 報では菓子、飲料の 2 つを統合した文章をコーパスとして使用し、1 つのベクトル空間を構築する。テンプレート 1,2（カテゴリ、原材料）を用いて生成するコーパス 1 と、テンプレート 2（原材料）のみで生成するコーパス 2 を、word2vec の入力コーパスとして、ベクトル空間を構築する。

3.4 Skip-gram による学習

word2vec には学習モデルとして Skip-gram と Countinuous Bag-of-Words があり、本研究では Skip-gram を採用した。Skip-gram は、文章中の単語を入力とし、その前後の単語を推定する学習を行う。前後何単語を関係性のある単語とするかは window パラメータで設定する。出力層では、ソフトマックス関数を用いて window パラメータで設定した前後の単語の出現確率を出力する。中間層では、出力層の周辺単語の出現確率のエラー率が最小となるように学習を行う。word2vec では、この中間層を単語の特徴ベクトルとし、ベクトル空間として使用する。今回は、文章ジェネレータで生成した文章を word2vec のコーパスとして使用する。window パラメータは初期値の 5 として学習を行う。また、word2vec は初期値として 100 次元の単語ベクトルを生成するが、今回は菓子の語彙数が 282 個、飲料の語彙数が 384 個と少数であるため、生成する単語ベクトルの次元数は 30 とした。

4．ベクトル空間の可視化

word2vec により構築した 30 次元の食品単語ベクトル空間に主成分分析を施し、3 次元上に可視化を行う。また、分布を確認するためクラスタ数を 3 に設定し、k-means 法を用いたクラスタリングを行う。コーパス 1 で生成したベクトル空間を図 1、コーパス 2 で生成したベクトル空間を図 2 に示す。

図 1 を見ると、コーパス 1 では、テンプレート 1 を用いて菓子か飲料かを明示的に学習させているため、菓子が左側に、飲料が右側に大きく 2 つに分離している。さらに飲料の中でも、野菜、果物、紅茶系飲料は右上に、ミルク、コーヒー系飲料が左下の 2 つに分かれて分布していることが確認できる。図 1 の上部では、原材料として果物が含まれる菓子、飲料が同じ軸に並んでいる。これらの結果より、コーパス 1 で構成したベクトル空間では、菓子か飲料のカテゴリ軸と原材料の軸によるベクトル空間が構築できていることが確認できる。

図 2 は、原材料のみを学習させたベクトル空間である。カテゴリ分類を陽に学習させていないため、図 1 のように分布が 2 極化することはないが、

(注 1) : https://recipe.rakuten.co.jp/
(注 2) : http://www.kirin.co.jp/products/list/nutrition/softdrink/

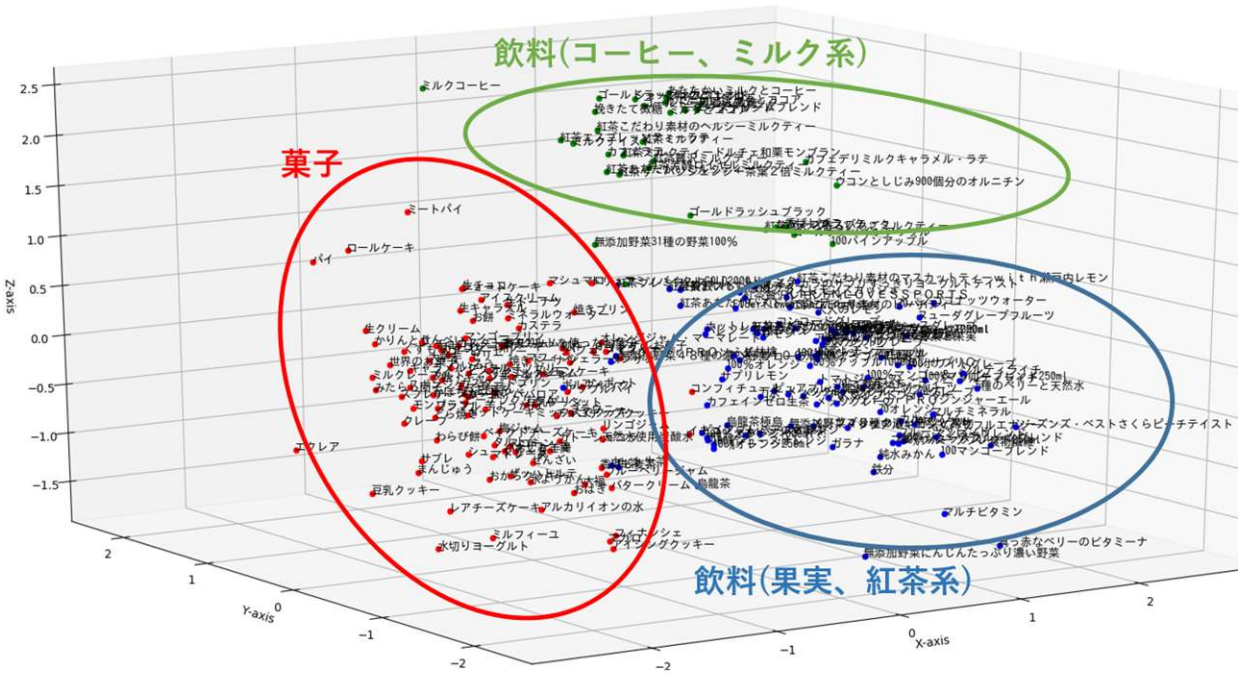

図 2. 食品単語ベクトル空間の可視化（コーパス 2）

原材料の観点から、菓子と飲料が分かれていることが確認できる。左側に菓子、右側に飲料が分離し、飲料の中でも原材料の観点から 2 つに分離している。また水系の飲料は、菓子、飲料の原材料として菓子、飲料どちらにも含まれるため中央付近に分布している。これらの結果より、コーパス 2 で構成したベクトル空間では、原材料のみでベクトル空間を構築していることから菓子、飲料の区別を考慮しない分類ができていることが確認できた。

5. ベクトル空間の評価

構築したベクトル空間をクラスタリングした後、各クラスタ内で類似語検索を行いベクトル空間の評価をする。クラスタリングには k-means を用いる。クラスタ数は菓子(14 分類)、飲料(7 分類)のカテゴリ数を考慮し 30 とする。

5.1 食品単語ベクトル空間のクラスタリング

ベクトル空間にクラスタ数 2~30 で k-means を適用し、以下の条件でその変化を確認する。

①クラスタリングの結果、和菓子が最も多く含まれるクラスタを和菓子のクラスタ、果実飲料が最も多く含まれるクラスタを果物飲料のクラスタとする。

③2 つのクラスタ間の級間分散、級内分散を算出、加えて各クラスタに含まれる和菓子、果実飲料のデータ数から含有率を算出し各クラスタでの変化を評価の指標とする。級間分散/級内分散、クラスタ内の含有率結果はクラスタ数ごとに 20 回のクラスタリングを施しその平均とする。

コーパス 1 ベクトル空間でのクラスタリング結果を図 3 に示す。また、実際の分布を 2 次元上に可視化した図を図 4,5,6 に示す。それぞれの図には果実飲料、和菓子が多く含まれる上位 3 つのクラスタの果実飲料、和菓子を表示している。果実飲料が含まれる上位 3 クラスタには赤系の 3 色、和菓子系が含まれる上位 3 クラスタには青系の 3 色、それ以外は灰色でプロットしている。同様にコーパス 2 ベクトル空間でのクラスタリング結果を図 7 に、2 次元上に可視化した図を図 8,9,10 に示す。

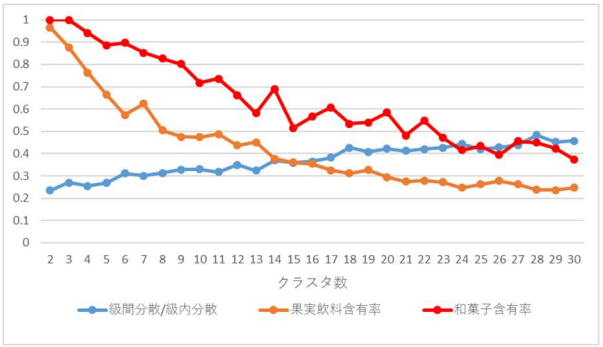

図 3. コーパス 1 ベクトル空間クラスタリング

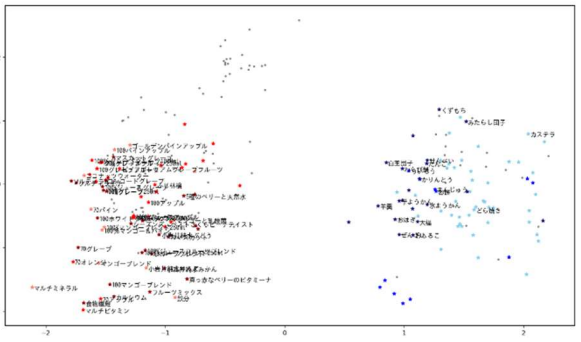

図 4. コーパス 1 ベクトル空間のクラスタリング可視化(k=10)

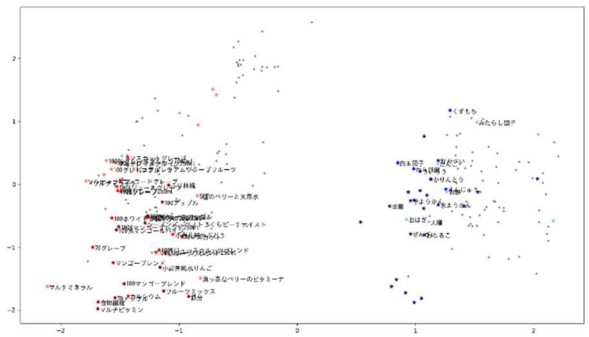

図 5. コーパス 1 ベクトル空間のクラスタリング可視化(k=20)

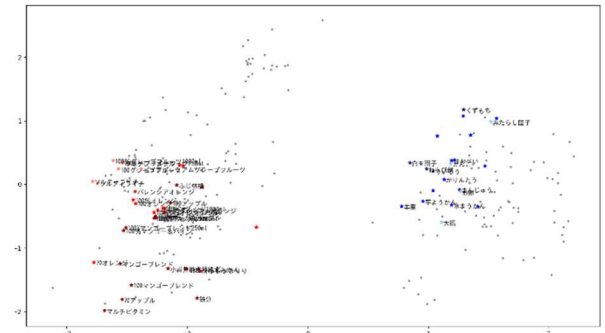

図 6. コーパス 1 ベクトル空間のクラスタリング可視化(k=30)

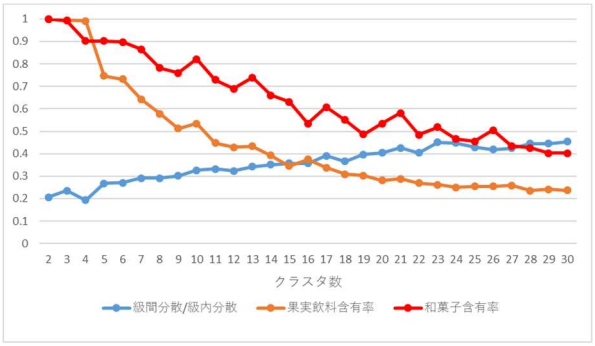

図 7. コーパス 2 ベクトル空間クラスタリング

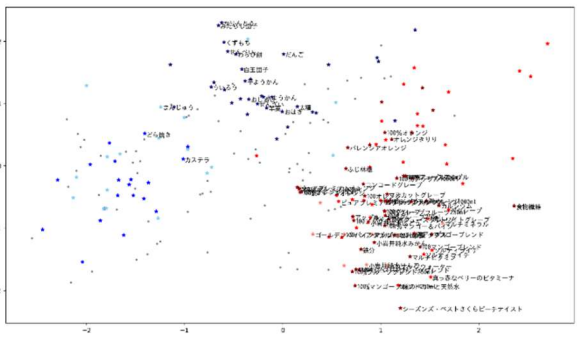

図 8. コーパス 2 ベクトル空間のクラスタリング可視化(k=10)

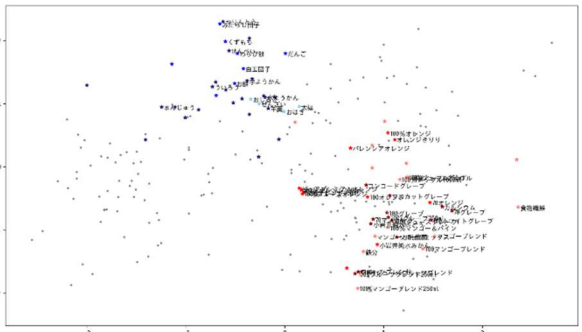

図 9. コーパス 2 ベクトル空間のクラスタリング可視化(k=20)

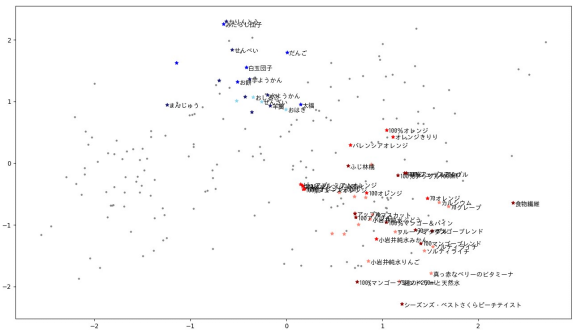

図 10.コーパス 2 ベクトル空間のクラスタリング可視化(k=10)

図 3、図 7 の結果を見ると、クラスタ数を増やすごとに果実飲料クラスタの果実飲料含有率、和菓子クラスタの和菓子含有率が減少し、級間分散/級内分散が増加している。クラスタ数を増やすことで、より細かい基準で分類されてクラスタが細分化されるため妥当な結果だといえる。図 3 ではカテゴリを学習させているため、クラスタ数が少ない段階において分類度が高いが、クラスタ数を増やすと図 4 の各値に大きな差異はなかった。2 次元プロット図では、クラスタ数の増加に伴いデータの細分化が視覚的に確認できる。また、コーパス 1 は菓子、飲料を区別しているため、コーパス 2 に比べてデータの分散が小さく、クラスタがコンパクトにまとまっている。

5.2 クラスタ内類似語検索

構築したベクトル空間に k=30 で k-means を施し、各クラスタ内での類似語検索を行う。類似度の評価にはユークリッド距離(E-dist)とコサイン類似度(C-dist)を用いる。表 1,2 にコーパス 1,2、ベクトル空間における「羊羹」が属するクラスタ内での「羊羹」の類似語検索結果を示す。同様に、表 3,4 ではコーパス 1, 2 ベクトル空間における「トロピカーナ 100%ジュースグレープ」の類似語検索結果を示す。表は、コサイン類似度を軸に類似度が高い順に並べている。ユークリッド距離と相違がある順位には赤で記している。表 4 を除くこれらの結果では、両距離計において、類似度順に違いはなかった。

食品（コーパス 1）では、羊羹の類似度が高いものとして、水羊羹や寒天といった菓子以外に、梅ジャムや、シャーベットといった菓子も取得していることが確認できる。上位 5 件以外にも同じクラスタには和菓子やプリン菓子が属していた。食品（コーパス 2）では、羊羹と同じクラスタには 4 つのみの単語が属し、原材料から類似度が高いものを取得できている。

飲料（コーパス 1）、100%ジュースグレープの類似度が高いものとして、同じグレープ系飲料が取得できている。上位 5 件以外には、ブレンド系の果実飲料や、果実飲料以外にもトマトジュースなどが属していた。飲料(コーパス 2)では、類似度の上位 5 件以外にもグレープ系の果実飲料が同じクラスタに属していた。

表 1. 「羊羹」類似語結果（コーパス 1）

単語名	C-dist		E-dist	
水ようかん	1	0.985	1	1.308
寒天	2	0.893	2	1.974
梅ジャム	3	0.849	3	2.314
シャーベット	4	0.779	4	2.749
コーヒーゼリー	5	0.758	5	3.144

(C-dist: コサイン類似度, E-dist: ユークリッド距離, 距離値の左側にある番号は類似度順を表す,以下同)

表 2. 「羊羹」類似語結果（コーパス 2）

単語名	C-dist		E-dist	
水ようかん	1	0.960	1	1.268
寒天	2	0.901	2	1.905
芋羊羹	3	0.895	3	2.040
コーヒーゼリー	4	0.787	4	3.013

表 3. 「100%ジュースグレープ」類似語結果（コーパス 1）

単語名	C-dist		E-dist	
トロピカル 100 グレープ	1	0.996	1	0.317
ハイパー100 グレープ	2	0.995	2	0.347
ホワイトグレープ	3	0.914	3	1.564
コンコードグレープ	4	0.866	4	2.428
マスカットグレープ	5	0.861	5	2.436

表 4. 「100%ジュースグレープ」類似語結果（コーパス 2）

単語名	C-dist		E-dist	
トロピカーナ 100 グレープ	1	0.996	1	0.306
ハイパー100 グレープ	2	0.995	2	0.346
ホワイトグレープ	3	0.914	3	1.536
コンコードグレープ	4	0.859	6	2.535
マスカットグレープ	5	0.898	7	2.584

5.3 食品単語ベクトル演算

word2vec ではそれぞれの単語を分散表現としてベクトル化しているため、単語間でのベクトル演算が可能となる。ベクトル演算の例としては、「King」-「Man」+「Woman」=「Queen」のようになる。単語のベクトル演算から、ベクトル空間の評価を行う。ここでは、3 つの単語を入力し、コサイン類似度、ユークリッド距離により 4 単語目を推測する比較演算を行う。比較演算結果を表 5,6 に示す。表 5 は、「ファイアブラック」「ファイアカフェラテ」「午後の紅茶おいしい無糖」を入力した結果である。表 6 は「生クリーム」「ロールケーキ」「生チョコ」を入力した結果である。比較演算にはコーパス 1 で構成したベクトル空間を用いた。前節と同様に、コサイン距離を軸に類似度が高い順に並べている、

表 5 では、飲料単語での比較演算を行っている。無糖のコーヒーと加糖のコーヒー、無糖の紅茶を入力としたため出力には加糖の紅茶が期待できる。コサイン類似度の結果を見ると、ミルクティー、エスプレッソティーといった加糖の紅茶が取得できた。同様にユークリッド距離の結果においても、加糖の紅茶が取得できた。また、両距離計を用いた結果、上位 5 単語中、4 単語は同じ単語が取得できた。

表 6 では、菓子単語での比較演算を行っている。こちらは材料とそれから作れる菓子の関係を入力としたため、出力としてチョコ系の菓子が期待できる。結果を見ると生チョコを使った菓子が取得できた。また、両距離計での結果上位 5 単語中、3 単語は同じ単語が取得できた。前節のクラスタ内類似語検索結果と合わせて、30 次元ベクトル空間において、ユークリッド距離系の有用性が確認できた。

表 5,6 のように、カテゴリ内でのベクトル演算は期待した結果を得ることができたが、菓子、飲料単語を跨いだベクトル演算を行うと期待した結果は得られなかった。これには、菓子、飲料に使用させる原材料の表記が異なるためである。例えばレモンという材料は、飲料ではレモンで表記されるが、菓子ではレモン果汁で表記されることが多い。カテゴリ間でのベクトル演算を可能にするには、お互いの表記のゆれをなくし、統一する必要があると考えられる。

「ファイアブラック」:「ファイアカフェラテ」
=「午後の紅茶おいしい無糖」 : ?
表 5. ベクトル演算結果 1

単語名	C-dist		E-dist	
午後の紅茶ミルクティー	1	0.739	1	3.797
午後の紅茶 茶葉 2 倍ミルクティー	2	0.724	8	4.240
午後の紅茶 エスプレッソティー	3	0.719	4	4.003
午後の紅茶 あたたかいミルクティー	4	0.718	3	3.985
午後の紅茶 ストレートティー	5	0.715	2	3.893

「生クリーム」:「ロールケーキ」=「生チョコ」 :?
表 6. ベクトル演算結果 2

単語名	C-dist		E-dist	
トリュフ	1	0.684	1	5.515
ガトーショコラ	2	0.673	5	5.581
ジェラート	3	0.661	20	5.951
カップケーキ	4	0.659	6	5.597
チョコレートケーキ	5	0.658	2	5.524

次に、表 5 における比較演算の位置関係について 2 次元上に可視化した図を図 11 に示す。名前は入力の 3 単語と類似度が最も高いものを表示しプロットには、入力単語と、類似度上位 5 単語を赤で表示した。30 次元のデータを 2 次元まで圧縮したため、綺麗な平行線とはならないが、コーヒーとカフェオレの関係性が視覚的に確認できる。

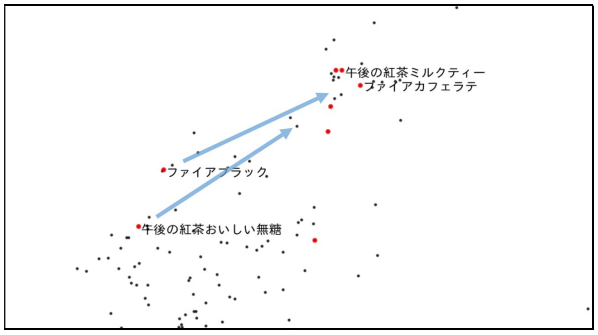

図 11. 比較演算 2 次元可視化

6. まとめ

本研究では、食品単語の類似性に関して密なベクトル空間の構築を目的とし、ジェネレータから文章を生成し、word2vec を用いたベクトル空間の構築を行った。また、構築した 2 つのベクトル空間にクラスタリング、主成分分析を用いてその評価を行った。カテゴリを明示的に学習させたコーパス 1 では、菓子、飲料が大きく分離するため各カテゴリ内での類似語分類となり、コーパス 2 で構築したベクトル空間では、カテゴリの分類なく食品単語全体を原材料から分類できた。また、様々なベクトル空間評価方法により、ジェネレータで生成する文章、学習させる内容を変えることで期待するベクトル空間の構築が可能であると分かった。

今後の課題として、菓子、飲料間でもベクトル演算が可能となるようなベクトル空間の構築が挙げられる。

参考文献

[1]Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean “Efficient Estimation of Word Representations in Vector Space," ICLR, 12pages, (2013)

[2]根本太晴,岡田浩之「word2vec による雑談対話システムの検討」,第 2 回 iHR 研究会,2 頁,(2015)

[3]矢野達也,林 豊洋,大橋 健「食品単語ベクトル空間とその評価」,第 6 回 iHR 研究会,(2017)