料理レシピの分散表現を用いた代替食材の発見手法

梅本 晴弥[1,a)]　豊田 哲也[1]　大原 剛三[1]

概要：近年，インターネット上で大量の料理レシピが共有されている．これらの料理レシピの中には，入手が困難であったり，調理時点で手元にない食材が用いられている場合がある．その場合，ユーザはその食材を使用せずに調理するか，代替可能な食材を用いて調理することになる．どのような食材によって代替可能かは必ずしも明らかではないため，これまでに大量のレシピをコーパスとして代替食材を推薦する手法が幾つか提案されているが，その精度はまだ十分なものとは言えない．これに対して，本研究では，word2vec によって獲得された食材名の分散表現に加えて，doc2vec によって獲得した料理レシピの分散表現を利用し，それら 2 つの分散表現から代替食材を発見する手法を提案する．評価実験では，word2vec のみを用いた代替食材の発見手法，および料理レシピのカテゴリを用いた代替食材の発見手法と提案手法を比較し，MAP，$GMAP$，$Recall$ の 3 つの評価指標を用いて提案手法の精度を評価する．

キーワード：代替食材検索，料理レシピ，分散表現

Finding Alternative Ingredients Based on Distributed Representation of Cooking Recepes

Haruya Umemoto[1,a)]　Tetsuya Toyota[1]　Kouzou Ohara[1]

Abstract: Recently, a large number of cooking recipes have been shared by many users on the Internet through specific sites such as Cookpad. Some of ingredients used in such a recipe may not be available at that moment for users who are going to cook dishes according to it since the ingredients are hard to get or simply the user do not have them at hand. In this case, the user would cook it either without using the ingredients or using alternative ingredients. As it is not always clear which ingredients can substitute for the unavailable ones, some methods of recommending alternative ingredients using a large amount recipes as a corpus have been proposed so far. But, their accuracy is not necessarily adequate. In this paper, we propose a method of finding alternative ingredients using both the distributed representation of cooking recipe obtained by doc2vec and that of ingredients obtained by word2vec. We experimentally show the usefulness of the proposed method by comparing it with two kinds of baseline methods, one that uses only the distributed representation of ingredients and one that uses categories of recipes for filtering out irrelevant ones, in terms of MAP, $GMAP$, and $Recall$.

Keywords: Alternative Ingredients，Cooking Recipes，Distributed Representation

1. はじめに

近年，インターネット上においてユーザが料理レシピを投稿するレシピ投稿型サイトにおいて，多数の料理レシピが共有されている．代表的なサイトとして，COOKPAD[*1] や楽天レシピ [*2]，ぺこり [*3] などが挙げられる．料理レシピには，料理の材料とその調理手順が記載され，料理の画像が付与されていること が多い．利用ユーザは，作成したい料理名や使用したい材料名で料理レシピを検索し，その結果から気に入った料理レシピを参考にしながら実際に料理を作成する．このとき，参考にする料理レシピにおいて

[1] 青山学院大学理工学部
[a)] a5814014@aoyama.jp
[*1] https://cookpad.com/
[*2] https://recipe.rakuten.co.jp/
[*3] http://pecolly.jp/

入手が困難であったり，調理する時点で手元にない食材がレシピ中で用いられている場合がある．その場合，ユーザはその食材を使用しなかったり，代替可能な食材を代わりに用いて料理を作成することになる．以下，ある食材に対して代替可能な食材を代替食材，元の食材を対象食材と呼ぶ．代替食材の候補には，基本的には対象食材の性質に近い食材が選ばれる．ここで，食材の性質とは食材自体の味や食感，栄養であったり，料理における食材の役割を指す．

どの食材をどの食材で代替できるかは必ずしも明らかではないため，レシピの情報を用いて食材間の類似度を求め，対象食材と代替食材の組み合わせを見つける手法がこれまでにも提案されている [1]．しかし，多くの場合，対象食材と代替食材の組み合わせは常に同じものを使用することができるが，特定の料理において用いることができなくなる組み合わせも存在する．たとえば，マヨネーズは肉料理や魚料理において広く用いられるバターの代替食材の 1 つである．一方，ケーキにおいてもバターは用いられるが，その代替食材としてマヨネーズはあまり用いられない．これは，肉料理や魚料理においては有塩バターが用いられるが，菓子やデザートにおいては無塩バターが用いられ，マヨネーズはより有塩バターに類似しているからである．この問題の解決方法としては，ある対象食材と代替食材の組み合わせを常に用いているのではなく，料理レシピの情報に基づいて組み合わせを変えることが考えられる．もし，ケーキのレシピにおいてバターではなくマヨネーズを用いる料理レシピが少なく，肉料理や魚料理においてバターの代わりにマヨネーズを用いる料理レシピが多いのであれば，類似するレシピを取得することで肉料理や魚料理ではバターの代替食材としてマヨネーズを抽出でき，ケーキにおいてはバターの代替食材としてマヨネーズを抽出することがなくなる．

このような考えに基づき，本稿では，食材間の類似度に加え，レシピ間の類似度も考慮して代替食材を見つける手法を提案する．具体的には，既存研究 [1] と同様に，word2vec [2] によって食材の分散表現を獲得し，その類似度を利用することに加え，料理レシピ間の類似度も考慮して食材の代替食材としての重要度を定量化する．同様の発想の下，レシピのカテゴリを直接利用した既存の代替食材発見手法 [3] とは異なり，本研究における提案手法では，doc2vec [4] により獲得できる料理レシピ自体の分散表現を用いて料理レシピ間の類似度を計算することで，より高い精度での代替食材の発見を図る．

以下，2 節において関連研究について述べ，3 節にて提案手法の詳細について述べる．4 節では評価実験の設定について説明し，5 節において実験結果と考察を述べる．6 節で本稿をまとめる．

2. 関連研究

野沢らは, 料理レシピにおける調理手順に対して word2vec を適用することで食材名の分散表現を獲得し，その類似度を用いて代替食材を発見する手法を提案している [1]．この手法は，代替食材が対象食材に類似した食材であることが多いことに着目したものである．食材の類似度計算にその分散表現を用いる点に関しては，本稿における提案手法と同じであるが，提案手法とは異なり料理レシピの情報は利用していない．一方，本研究と同様に，対象食材と代替食材の組み合わせは常に使用できないという考えの下，志土地らは，料理カテゴリ毎の食材と調理手順の共起関係を用いて代替食材を発見する手法を提案している [3]．この研究では，料理カテゴリ名と食材名をそのまま用いているが，我々の提案手法では，料理レシピと食材の分散表現を使用している．

その他，事前に料理レシピを分析し，食材に対し代替可能か否かを付与し，それに基づいて料理レシピを推薦するシステム [5] や，健康効果を考慮しつつ，食材の食感情報を使用して代替食材を発見する手法 [6] などが提案されている．前者の研究は，料理カテゴリにおける食材の使用比率から代替食材の可否を決定してる点で本研究とは異なる．また，後者の研究は，食材の食感を用いた類似度計算を行っている点で本研究とは異なる．

また，料理レシピの内容を分散表現化する試みとしては，Salvador らの研究がある [7]．彼らは料理レシピの調理食材，調理手順と料理の画像を分散表現化することで，それらの間で類似度の計算を可能にし，料理の画像から調理食材，調理手順の推定を実現している．

3. 提案手法

提案手法は，対象食材と代替食材の組み合わせは料理のカテゴリによって変化するという考えに基づき，食材の類似度のみを用いるのではなく，料理レシピの類似度を用いることで，代替食材発見の精度向上を図るものである．それらの類似度には，それぞれ word2vec [2] と doc2vec [4] によって獲得された食材と料理レシピの分散表現を用いる．提案手法の流れは，以下の通りである．

(1) 対象食材の料理レシピと類似する料理レシピを，doc2vec によって獲得された分散表現を用いて計算し，その上位 N 件を抽出する．
(2) 類似料理レシピに使用されている食材と対象食材の類似度を，word2vec によって獲得された分散表現を用いて計算する．
(3) 類似料理レシピに使用されている食材すべてに対し，重要度を計算し，重要度が高い食材を代替食材候補とする．

(4) 代替食材候補から，非表記ゆれ条件を満たすものだけを出力する．

提案手法では，分散表現（ベクトル）間の類似度としてコサイン類似度を用いた．

以下，各手順の詳細について説明する．

3.1 doc2vec を用いた類似レシピの抽出

doc2vec [4] は Le と Mikolov により提案された文書の分散表現（ベクトル表現）を獲得する手法であり，同じく Mikolov らにより提案された単語の分散表現を獲得する word2vec [2] の拡張モデルである．word2vec が入力として単語列のみを受け取るのに対し，文章の ID も入力することで，単語の分散表現だけではなく文章の分散表現も獲得可能にしている．本稿においては，doc2vec のモデルの 1 つである Paragraph Vector with Distributed Memory（PV-DM）を用い，分散表現の次元数は 300，入力に与える単語列の長さである window size は 5 とした．また，料理の手順を結合し，1 つの料理レシピにおける料理手順のすべてを 1 入力として doc2vec に入力した．

提案手法では，このようにして獲得した料理レシピの分散表現のコサイン類似度を用い，対象食材が使用されている料理レシピの類似レシピの上位 N 件を抽出し，代替食材の発見に使用する．ここで，N は提案手法におけるハイパーパラメータである．

3.2 word2vec を用いた類似レシピの抽出

Mikolov らによって提案された word2vec [2] は，隠れ層 1 層のニューラルネットワークモデルで実現されており，単語列を入力し，単語間の共起関係を学習することで，任意の次元数の単語の分散表現を獲得できる．本研究では，word2vec のモデルの 1 つである Countinuous Bag-of-Words（CBOW）を用い，分散表現の次元数を 300，入力に与える単語列の長さである window size を 5 とした．入力文章に関しては，野沢らの研究 [1] に従い，料理レシピの調理手順における個々の文を 1 入力とした．提案手法では，対象食材が使用されている料理レシピの類似レシピ N 件で使用されている食材に対し，word2vec で獲得された分散表現を用いて対象食材との類似度を計算する．

なお，doc2vec においても単語の分散表現は獲得できるが，前節で述べた dco2vec の利用法では，料理手順すべてを結合して入力しているため，1 入力の単位が異なり，doc2vec と word2vec によって獲得された単語の分散表現は異なるものになる．そのため，それらの分散表現を用いて得られる類似語も異なる．例として，食材 “牛肉” に対して doc2vec と word2vec それぞれで獲得した分散表現により抽出した類似語を表 1 に示す．表 1 から，明らかに doc2vec では単語の分散表現をうまく学習することが出来なかったことが分かる．これは，料理手順のすべてを 1 入力としているため，手順と手順が連続した文章として認識される一方，実際には手順間において文章の傾向が大きく異なり，単語の共起関係をうまく学習できなかったためであると考えられる．以上の理由から，提案手法では，料理レシピの分散表現は doc2vec から獲得し，食材の分散表現については word2vec から獲得している．

表 1 “牛肉” に類似する単語上位 10 個

順位	word2vec	doc2vec
1	豚肉	旨み
2	豚バラ肉	あまい
3	豚ばら肉	取り分ける
4	肉	巻き寿司
5	豚こま	省き
6	豚バラ	アーモンドスライス
7	お肉	ブランデー
8	鶏肉	含み
9	豚もも肉	見計らい
10	豚こま肉	引っ張ら

3.3 食材と料理レシピの類似度を用いた食材重要度の算出

次に，代替食材としての食材の重要度の計算方法について述べる．提案手法では，料理レシピ r に含まれる対象食材 t に対して，N 件の料理レシピ集合 R に含まれる食材 f の t の代替食材としての重要度 $I_{t,f}$ は，次の式 (1) により計算する．

$$I_{t,f} = \max_{r' \in R_f} \left(\frac{S^W_{t,f} - S^W_{min}}{S^W_{max} - S^W_{min}} \times \frac{S^R_{r,r'} - S^R_{min}}{S^R_{max} - S^R_{min}} \right) \quad (1)$$

ここで，$S^W_{t,f}$ は対象食材 t に対する食材 f の類似度，S^W_{min} は t と料理レシピ r に類似する上位 N 件のレシピの集合 R に含まれる食材の最小の類似度，S^W_{max} は t と R 中の食材の最大の類似度，$S^R_{r,r'}$ は r と R に含まれるレシピのうち食材 f を含むレシピ集合 R_f の要素 r' の類似度，S^R_{min} は r と R 中のレシピの最小の類似度，R_{max} は r と R 中のレシピの最大の類似度である．式 (1) では，食材 t と f の類似度と料理レシピ r と f を含む料理レシピ r' の類似度の積の最大値を，食材 f の対象食材 t の代替食材としての重要度 $I_{t,f}$ としている．ただし，それぞれの類似度の計算の際には，それぞれの値域を 0 から 1 の範囲に正規化している．これは，食材の類似度の値域は比較的広い一方，料理レシピは類似している上位 N 件を抽出しているため，その類似度の値域はとても狭くなり，正規化せずに 2 つの類似度から同様にして重要度を算出した場合，料理レシピの類似度の影響がほとんどなくなってしまうためである．

提案法では，式 (1) によって類似レシピ集合 R に現れる食材すべてに対して重要度を付与し，重要度の高い食材から代替食材候補とする．

3.4 表記ゆれへの対処

料理レシピ中の食材名に関しては，表記ゆれが多数存在する．そのため，提示する代替食材に対象食材と同じものが含まれることや，同じ代替食材が複数含まれることを防ぐために，以下の条件を用いて代替食材候補を絞り込む．

- 代替食材候補のカタカナ読みに対して対象食材のカタカナ読みは部分一致しない．
- 代替食材候補のより上位に同一のカタカナ読みの食材は存在しない．

1 つ目の条件は対象食材の表記ゆれ単語を除外し，2 つ目の条件は代替食材の表記ゆれ単語を除外するものである．これら 2 つの条件を非表記ゆれ条件とし，提示手法では，前節で定義した食材重要度に基づいて抽出した代替食材候補のうち，非表記ゆれ条件を満たすもののみを最終的な代替食材候補として提示する．

4. 評価実験

4.1 実験データ

本実験では，COOKPAD 株式会社が公開する料理レシピシピデータ [*4] を用いた．料理レシピの件数は 1,716,252 件であり，料理レシピのタイトル，食材，調理方法，つくれぽの情報が含まれている．ここでつくれぽとは，料理レシピに対する利用ユーザのコメントである．

4.2 正解データ

本実験では，野沢らの研究 [1] に従い，料理レシピに対するコメントから正規表現を用いて代替食材を抽出し，それを正解データとして用いる．料理レシピに対するコメントの一部には，食材を別の食材で置き換えて調理した報告が存在するため，それを正解データとすることで被験者実験を必要とせずに手法の評価が可能になる．正規表現は文献 [1] に従い，/きらして | (が—は)(なかった—無かった) — (の (代わり—かわり—替わり—換わり—代り—替り))/ を用い，正規表現でマッチしたパターンの前後 1 名詞を食材と代替食材の組み合わせとし，正解データとした．実験データにおけるつくれぽの総数は 805,187 件で，そこから 9,504 件の正解データを正規表現で抽出することができた．この 9,504 件のデータを正解データセット 1 とする．ここで，この件数は，抽出した対象食材と代替食材の組み合わせの頻度の合計を意味する．表 2 に正解データセット 1 における出現頻度上位 10 件の対象食材と代替食材の組み合わせを示す．

一方，正解データセット 1 には誤抽出した組み合わせが存在したため，出現回数が 2 回未満の組み合わせは正解データから除外した．その結果，正解データ数は 3,389 件となり，これを正解データセット 2 とする．

[*4] https://www.nii.ac.jp/dsc/idr/cookpad/cookpad.html

表 2 正解データセット 1 における対象食材と代替食材の組み合わせの出現頻度上位 10 件

食材	代替食材	件数
牛乳	豆乳	52
ベーコン	ウインナー	43
ベーコン	ハム	40
ベーコン	ウィンナー	25
大葉	ネギ	24
ネギ	玉ねぎ	23
ハム	ベーコン	21
ネギ	大葉	21
豆乳	牛乳	20
生クリーム	牛乳	20

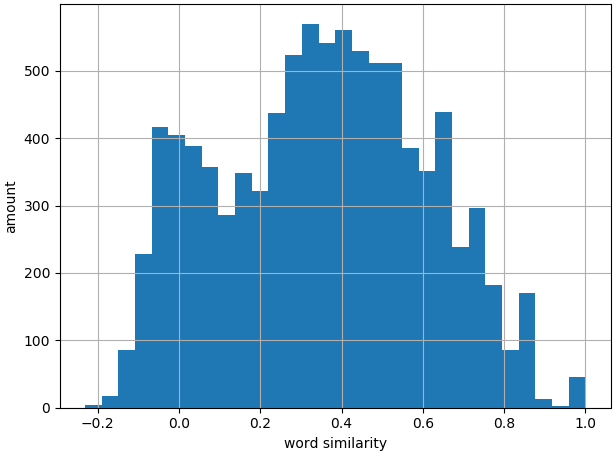

図 1 正解データセット 1 における対象食材と代替食材の類似度

正解データセット 1 と正解データセット 2 のそれぞれにおける対象食材と代替食材の単語の類似度の分布をそれぞれ図 1 と図 2 に示す．図 1 と図 2 を比較すると，出現回数が 2 回未満の組み合わせを正解データセットから除外することで，類似度 0 から 0.2 付近の組み合わせが大きく減少していることがわかる．これは，類似度が小さい組み合わせには誤抽出の組み合わせが多く存在していることを示しており，実際，類似度 0 付近の組み合わせについては誤抽出の組み合わせが多かった．そこで，正解データセット 2 の対象食材と代替食材の類似度が 0.2 未満の組み合わせを除外した．この処理によって正解データ数は 2,951 件となり，これを正解データセット 3 とする．

4.3 評価指標

評価指標としては MAP（Mean Average Precision），$GMAP$（Geometric Mean Average Precision），および $Recall$ を用いる．それぞれの定義を以下に示す．

$$MAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \qquad (2)$$

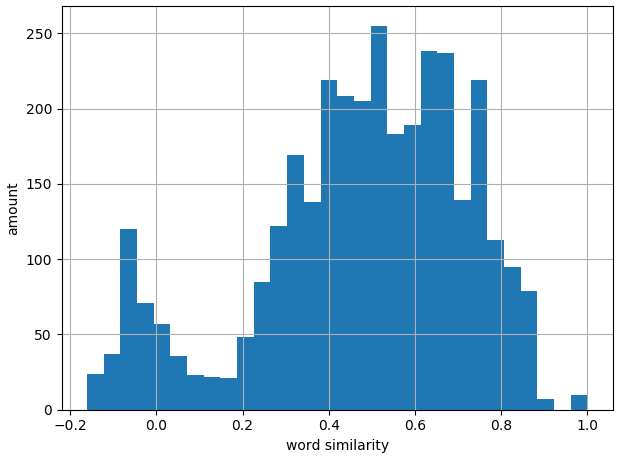

図 2 正解データセット 2 における対象食材と代替食材の類似度

$$GMAP = \left(\prod_{i=1}^{n} AP_i \right)^{\frac{1}{n}} \tag{3}$$

$$Recall(r) = \frac{1}{N} \sum_{i=1}^{N} |\{t_i\} \cap B_i(r)| \tag{4}$$

ここで，N は正解データ数を表し，AP_i は i 番目の正解データ（対象食材 t_i と代替食材 f_i^* の組み合わせ）に対して，f_i^* がこのデータに対して提示された代替食材候補集合 B_i において r 番目の候補として提示された場合に $1/r$ という値を取るものであり，以下のように定義される．

$$AP_i = \sum_{r=1}^{|B_i|} Prec(i,r) I(i,r) \tag{5}$$

ただし，$Prec(i,r)$ は，B_i 中の上位 r 件の代替食材候補を $B_i(r)$ とした場合に次式のように定義される．

$$Prec(i,r) = \frac{|\{t_i\} \cap B_i(r)|}{|B_i(r)|} \tag{6}$$

また，$I(i,r)$ は，B_i の r 番目の代替食材候補が正解 f_i^* に一致する場合に 1，そうでない場合は 0 を返す関数である．

MAP は推薦の文脈ではモデルの平均的な適合率を評価するものであり，本実験では，代替食材発見手法が正解代替食材を上位に提示できるほどその値が高くなる．$GMAP$ も同様に平均的な適合率を評価するものであるが，正例が推薦候補に存在しない場合，その値が大幅に減少する．本実験においては，モデルが正解代替食材を候補に含めることができるかどうかを評価する指標となる．$Recall(r)$ は，推薦数における正解代替食材の含有率を示し，上位 r 個の代替食材候補の中で，どの程度正解代替食材を提示できたかを表す．

4.4 比較手法

4.4.1 評価実験における比較手法

本実験では，次の 4 つの手法を比較する．1 つ目の手法は 4.4.2 節で述べる word2vec を用いた代替食材の発見手法であり，以下，word2vec 法と表記する．2 つ目の手法は 4.4.3 節で述べる料理レシピのカテゴリと word2vec を用いた代替食材の発見手法であり，以下，カテゴリ抽出法と表記する．3 つ目の手法は 3 節で述べた提案手法において，料理レシピの類似度を使用しない手法であり，以下，doc2vec 法と表記する．4 つ目の手法は 3 節で述べた提案手法である．

表 3 料理レシピから抽出した料理カテゴリ

カテゴリ名	件数
サラダ	24,526
煮物	11,111
パスタ	8,990
スープ	8,334
チーズ焼き	4,969
炊き込みご飯	4,929
ケーキ	4,850
カレー	4,358
クッキー	4,342
パウンドケーキ	4,304

4.4.2 word2vec 法

本実験では，野沢らの手法 [1] を食材の分散表現を用いる word2vec 法として用いた．単語の分散表現の獲得においては，料理レシピの調理手順を形態素解析し，1 行を 1 入力として word2vec に対して入力し，食材の分散表現を獲得した．このモデルを代替食材の発見に使用する場合は，対象食材と他の単語の類似度を計算し，類似度が高い単語を代替食材候補とする．このとき，word2vec への入力単語は食材に限定していないため，代替食材として食材単語以外が抽出されてしまう可能性があり，後述するモデルと比較し不利であるため，代替食材候補は料理レシピに食材として使用されている単語に限定するという処理を加えた．さらに，3.3 節で述べた非表記ゆれ条件を適用した．以下に，word2vec 法の処理の流れを示す．

(1) 対象食材に対して高い類似度を持つ単語を，word2vec によって獲得された分散表現を用いて計算する．
(2) 対象食材に対して高い類似度を持つ単語のうち，料理レシピの食材として使用されていない単語を除外し，代替食材候補とする．
(3) 代替食材候補から，非表記ゆれ条件を満たすもののみを出力する．

4.4.3 カテゴリ抽出法

カテゴリ抽出法では，料理レシピの類似度を用いるのではなく，料理レシピのカテゴリを用いて代替食材を絞り込む．料理レシピのカテゴリ情報は，今回用いたデータセットには殆ど含まれていなかったため，料理レシピのタイトルからその料理のジャンルを抽出し，それを料理レシピのカテゴリとして用いた．抽出方法は，料理レシピのタイト

表 4 正解データセット 3 における各手法のスコア

	word2vec 法	カテゴリ抽出法	doc2vec 法	提案手法
MAP	0.094	0.199	0.238	0.248
$GMAP$	$0.376e^{-10}$	$0.165e^{-6}$	$0.466e^{-5}$	$0.525e^{-5}$
R@1	0.117	0.129	0.146	0.153
R@5	0.294	0.331	0.390	0.399
R@10	0.421	0.426	0.508	0.518
R@20	0.539	0.524	0.611	0.621
R@50	0.682	0.625	0.728	0.730

ルに対して形態素解析を行い，最後の名詞または最後とその直前が名詞であれば 2 つの名詞を結合し，それをその料理のジャンルとした．表 3 に料理レシピ数上位 10 件のカテゴリを示す．料理レシピに対しカテゴリを付与することで，同一カテゴリである類似した料理レシピを抽出することができる．以下に，カテゴリ抽出法の処理手順を示す．ただし，N は本手法におけるハイパーパラメータとする．

(1) 対象食材の料理レシピからカテゴリを抽出する．
(2) 抽出されたカテゴリと同一のカテゴリに属する料理レシピをランダムに N 件抽出する．
(3) 同一のカテゴリに属する料理レシピに使用されている対象食材と対象食材の類似度を，word2vec によって獲得された分散表現を用いて計算する．
(4) 対象食材に対して類似度の高い食材を抽出し，代替食材候補とする．
(5) 代替食材候補から，非表記ゆれ条件を満たすもののみを出力する．

4.5 実験方法

本実験では，作成した正解データセット 3 を用いて実験を行った．正解データセットは，対象食材と代替食材の組み合わせ，および対象食材が用いられている料理レシピの情報が 1 組のデータとなっている．モデルが word2vec のみの場合は，料理レシピの情報は用いず，対象食材との類似度が高い単語を計算し代替食材候補を抽出する．代替食材候補数は 50 とし（$|B_i| = 50$），同一の対象食材と代替食材の組み合わせが複数存在するが，重複を許し評価をした．料理レシピのデータを使用する手法の場合は，料理レシピの情報を用いて代替食材候補を抽出する．この時，レシピの分散表現を用いる場合は，対象食材が使用されている料理レシピとの類似度が高いレシピ上位 100 件を抽出した．レシピのタイトルを用いる場合は，同一カテゴリの料理レシピをランダムに 100 件を抽出した．ただし，毎回結果が異なるため，各評価値は 10 回実行した結果の平均を取った．代替食材候補数は word2vec と同様の 50 とした．

5. 実験結果と考察

5.1 評価指標による各手法の評価

正解データセット 3 における各手法の実験結果を表 4 に示す．ただし，R@r は提示代替食材の上位 r 件における $Recall$ を表す．

表 4 において，MAP と $GMAP$ は word2vec 法とカテゴリ抽出法を比較すると，大幅にスコアが向上していることが分かる．次に $Recall$ において比較すると，上位 1 件，5 件，10 件ではカテゴリ抽出が word2vec 法よりもスコアが高いものの，上位 20 件，50 件においてはスコアが低くなっている．これは，カテゴリの絞り込みによって得られた類似料理レシピのみから代替食材を発見することで，対象食材に対して高い類似度を持つものの代替食材となり得ない食材を代替食材候補とせずに，正しい代替食材をより上位の代替食材候補とすることができたため，上位における $Recall$ が向上したと考えられる．一方，上位 20 件，50 件では $Recall$ は低下は上記理由により正しい代替食材が省かれたことを表している．

次に，カテゴリ抽出法と doc2vec 法を比較する．MAP，$GMAP$，$Recall$ の全ての評価指標において doc2vec のみ手法はスコアを向上させていることが分かる．カテゴリ抽出法と同様に，doc2vec 法は類似料理レシピのみから代替食材を発見する手法であるが，結果には大きな差が生じた．その理由としては，doc2vec 法では料理レシピ間の類似度の計算が可能であり，最も類似する N 件の料理レシピを抽出することが可能であるため，カテゴリ抽出法において省かれていた正しい代替食材が，doc2vec 法においては省かれづらくなったことが考えられる．また，doc2vec 法の上位 20 件，50 件における $Recall$ が word2vec 法と比較し，大幅に向上している結果もこの考察を支持する．

最後に，doc2vec 法と提案手法を比較する．MAP，$GMAP$ においては，提案手法は幅は小さいもののスコアが向上していることが分かる．これは，代替食材を発見する際に料理レシピの類似度を使用することによって，料理レシピに合致した代替食材をより上位の代替食材候補とすることができたためと考えられる．$Recall$ に関しても，料理レシピの類似度を考慮することによってスコアが向上していることがわかるが，上位 50 件においてはそのスコア差がほぼ同一のものとなっている．理由として，料理レシピ類似度を使用によって上位の代替食材候補の順位は入れ替わる一方，上位 50 件の代替食材候補の集合として考えると，内容はほとんど変化していないからだと考えられる．

表 5 word2vec 法のスコアが悪かった料理レシピタイトルと正解データ組み合わせの例

料理レシピカテゴリ	ひじきの煮もの
対象食材	竹輪
代替食材	油揚げ

表 6 word2vec 法のスコアが悪かった例の各手法による代替食材の提示内容

提示順位	word2vec のみ	カテゴリ抽出	doc2vec のみ	doc2vec+類似度
1	魚肉ソーセージ	さつま揚げ	ウィンナー	油揚げ
2	さつま揚げ	薄揚げ	薄揚げ	薄揚げ
3	お魚ソーセージ	油揚げ	油揚げ	ウィンナー
4	ウインナー	厚揚げ	蒲鉾	茄子
5	薄揚げ	エリンギ	厚揚げ	かまぼこ
6	小揚げ	長ネギ	蓮根	厚揚げ
7	ソーセージ	茄子	ハム	蓮根
8	ごぼう天	ピーマン	茄子	いんげん
9	笹かま	レンコン	ピーマン	ピーマン
10	油揚げ	椎茸	レンコン	人参

表 7 カテゴリ抽出法のスコアが悪かった料理レシピタイトルと正解データ組み合わせの例

料理レシピカテゴリ	のり酢あえ
対象食材	ツナ
代替食材	ハム

表 8 カテゴリ抽出法のスコアが悪かった例の各手法による代替食材の提示内容

提示順位	word2vec のみ	カテゴリ抽出	doc2vec のみ	doc2vec+類似度
1	シーチキン	きゅうり	シーチキン	シーチキン
2	シーチキン缶	もやし	コーン	コーン
3	オイルサーディン	ほうれんそう	ハム	ハム
4	コーン	酢	ミックスビーンズ	ミックスビーンズ
5	鮭缶	切干大根	しらす	しらす
6	ハム	のり	塩昆布	塩昆布
7	コンビーフ	三杯酢	なめたけ	なめたけ
8	サーディン	白醤油	かにかま	かにかま
9	カニ缶	ソルト	じゃこ	マヨネーズ
10	ホタテ缶	マザー	マヨネーズ	じゃこ

5.2 代替食材の提示内容に対する考察

各手法の提示した代替食材の内容について考察する．まず，word2vec 法のスコアが悪かった例を示す．料理レシピタイトルと正解データの組み合わせを表 5 に，各手法による代替食材の提示内容を表 6 に示す．word2vec 法では，竹輪の代替食材として魚肉ソーセージやウィンナーなどの棒系の食材が代替食材として提示されているが，これらの食材は煮ものにおいてはあまり使用されない食材である．一方，料理レシピの情報を用いた他手法のうち，カテゴリ抽出法では揚げ系の食材が上位に提示されており，料理カテゴリの絞り込みによって候補食材も同様に大きく絞りこまれていることが分かる．doc2vec 法では，揚げ系の食材以外にウィンナーやかまぼこが上位に提示されており，料理カテゴリでは除外された料理レシピが類似レシピとして抽出されていることが分かる．また，提案手法では，煮ものではあまり用いられないウィンナーやハムの順位が下がっており，料理レシピの類似度を用いることで料理レシピにより適合した代替食材候補の提示順序になっていることが分かる．

次に，カテゴリ抽出法のスコアが悪かった例を示す．料理レシピタイトルと正解データ組み合わせを表 7 に，各手法による代替食材の提示内容を表 8 に示す．word2vec 法では主に缶詰の食材が提示されており，正解食材も上位 10 個に提示されている．ただし，酢の物の代替食材としてはあまり使用されないコンビーフが提示されている．カテゴリ抽出法では，一般的な酢の物に使用される食材は提示出来ているが，代替食材がほとんど抽出できておらず，正解食材も提示できなかった．これは，料理カテゴリの絞り込みによって得られた料理レシピの多様性が小さく，ツナに対応する食材を発見できなかったためであると考えられる．doc2vec 法と提案手法では，正解食材も提示できており，しらすやかにかまが word2vec 法と比較し上位に提示されている．また，食材提示順位の変化は小さく，この例において料理レシピの類似度を用いた効果は小さかった．

表 9 提案手法のスコアが悪かった料理レシピタイトルと正解データ組み合わせの例

料理レシピカテゴリ	キャベツの辛子マヨネーズ和え
対象食材	ハム
代替食材	カニカマ

表 10 提案手法のスコアが悪かった例の各手法による代替食材の提示内容

提示順位	word2vec のみ	カテゴリ抽出	doc2vec のみ	doc2vec+類似度
1	魚肉ソーセージ	カニカマ	コーン	コーン
2	ソーセージ	コーン	ツナ	ツナ
3	ウインナー	ツナ	レタス	レタス
4	ベーコン	ちくわ	人参	キャベツ
5	お魚ソーセージ	レタス	シーチキン	マヨネーズ
6	カニカマ	海苔	キャベツ	人参
7	ウィンナー	人参	マヨネーズ	シーチキン
8	スモークサーモン	プロセスチーズ	キュウリ	大葉
9	スライスチーズ	シーチキン	大葉	チーズ
10	コーン	キャベツ	チーズ	ピーマン

最後に，doc2vec 法と提案手法のスコアが悪かった例を示す．料理レシピタイトルと正解データ組み合わせを表 9 に，各手法による代替食材の提示内容を表 10 に示す．word2vec 法とカテゴリ抽出法では正解食材が提示出来ている一方，doc2vec 法と提案手法では正解食材を提示できなかったことがわかる．これは，doc2vec に基づく類似レシピが正しく抽出できていないことを示している．そこで，料理レシピの調理手順を見てみると，調理手順の文章量が少なく，“☆を混ぜて〜” と使用食材が記号で省略されていることがわかった．つまり，調理手順文章から doc2vec を用いて料理レシピの分散表現を獲得しているため，調理手順が短かったり，使用食材名が記号で置き換えられていた場合，正しい料理レシピの分散表現が獲得できなかったと考えられる．

6. おわりに

本研究では，料理レシピと食材の分散表現を用いた代替食材の発見手法を提案した．提案手法では，料理レシピの分散表現を用いて類似レシピを抽出し，その類似度と食材自体の類似度を掛け合わせることでより料理レシピに適合した代替食材の発見を可能にした．word2vec のみ用いた代替食材の発見手法と，料理レシピのカテゴリを用いた代替食材の発見手法を比較すると，後者の方が高い結果となった．これは，食材の検索範囲を料理レシピのカテゴリで絞り込むことで，類似しているが特定のカテゴリでは使用されない代替食材を省くことが出来ているからである．また，料理レシピのカテゴリを用いた代替食材の発見手法と提案手法を比較すると，後者の方が高い結果となった．これは，料理レシピの分散表現を用いることで，カテゴリを用いるよりもより類似する料理レシピを抽出することが出来たからである．

今後の課題としては，料理レシピのより正確な分散表現化が挙げられる．評価実験によって提案手法には欠点が存在することが明らかになり，料理レシピにおける調理手順において，文章が短かったり食材が記号に置き換えられている場合，正しい料理レシピの分散表現が獲得できず，それによって代替食材も正しく提示することが出来なかった．これは調理手順の情報のみから料理レシピの分散表現を獲得していることに起因する．よって，食材と調理手順，料理画像を用いたより料理レシピを正しく表現できる分散表現を獲得することで，代替食材の発見精度が向上する可能性がある．

本研究では、クックパッド株式会社と国立情報学研究所が提供する「クックパッドデータ」を利用した。

参考文献

[1] 野沢健人，中岡義貴，山本修平，佐藤哲司：word2vec を用いた代替食材の発見手法の提案 (データ工学)，電子情報通信学会技術研究報告= IEICE technical report: 信学技報，Vol. 114, No. 204, pp. 41–46 (2014).
[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119 (2013).
[3] 志土地由香，井手一郎，高橋友和， 村瀬洋：料理レシピマイニングによる代替可能食材の発見，電子情報通信学会論文誌 A， Vol. 94, No. 7, pp. 532–535 (2011).
[4] Le, Q. and Mikolov, T.: Distributed representations of sentences and documents, *International Conference on Machine Learning*, pp. 1188–1196 (2014).
[5] 眞喜子平野，朋文植竹：代替食材を考慮した料理レシピ推薦システムの提案，第 79 回全国大会講演論文集，Vol. 2017, No. 1, pp. 371–372 (2017).
[6] 花井俊介，難波英嗣，灘本明代：健康を意識した代替食材の発見手法，第 7 回データ工学と情報マネジメントに関するフォーラム，G6-6 (2015).
[7] Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I. and Torralba, A.: Learning cross-modal embeddings for cooking recipes and food images, *Training*, Vol. 720, pp. 619–508 (2017).