DEIM Forum 2019 P2-145

網羅性を有する食材オントロジの自動構築法

窪田　茜† 馬場　睦也†† 楠　和馬†† 波多野賢治†††

† 同志社大学文化情報学部　〒 610-0394 京田辺市多々羅都谷 1-3
†† 同志社大学大学院文化情報学研究科　〒 610-0394 京田辺市多々羅都谷 1-3
††† 同志社大学文化情報学部　〒 610-0394 京田辺市多々羅都谷 1-3
E-mail: †bip0118@mail4.doshisha.ac.jp, ††{baba,kusu}@ilab.doshisha.ac.jp, †††khatano@mail.doshisha.ac.jp

あらまし　レシピ検索や代替食材の提案といった食材に着目した研究の増加に伴い，食材の表記ゆれや表記粒度の違いによって生じるレシピ研究の精度の低下という問題が浮き彫りになっている．この問題に対し，食材の表記を統一するための料理オントロジが構築されているが，既存のオントロジは半自動的に構築されるため人的コストがかかり，またそれに含まれる食材の種類も十分ではない．あらゆる食材を網羅した食材オントロジの自動構築を実現するため，本研究では幅広い利用目的に対応した日本食品標準成分表を用いて食材オントロジを自動構築する手法を提案し，その有用性を検証する．

キーワード　食材，オントロジ，網羅性，料理レシピ

1 はじめに

近年クックパッド[1]や楽天レシピ[2]のようなレシピ公開サイト（以下，レシピサイト）の普及により，Web 上に公開されているレシピの数が膨大になっている．そのため，ユーザの情報要求を満たすレシピの提示が困難であるという問題が生じている．この問題に対し，食材の優先度を考慮したレシピ検索に関する研究 [1] や代替食材の提案に関する研究 [2,3]，レシピの自動要約に関する研究 [4] などが行われている．しかし，レシピの材料や手順には「まあじ」や「葉だいこん」といったレシピ特有の表現や「豚肉」や「豚バラ」などの表記粒度の違いが存在するため，食材自体の認識が困難である [5]．また，ユーザ投稿型レシピサイトでは，投稿者によって食材の表記が異なるため，上記で述べた研究の精度が低下するという問題が浮き彫りになっている．食材の表記ゆれや表記粒度の違いに対応する方法としてオントロジの構築が挙げられる．Guarino et al. によるとオントロジとは，ある領域の論理構造および，その概念とそれらの間の関係を記述したものであると述べられている [6]．オントロジを構築している研究として，土居らの料理オントロジ [5] がある．土居らは楽天レシピと特許データ[3]を用いて料理オントロジを構築しているが，構築方法が半自動的であるため人的コストがかかるという問題点がある．また，オントロジに含まれる食材の種類が少ないことも問題点として挙げられる．

そこで本稿では，あらゆる食材を網羅したオントロジの自動構築を実現するため，教育や研究，行政など幅広い利用目的に対応した日本食品標準成分表（以下，成分表）[7–10] を用いて食材オントロジを構築する方法を提案する．また，本稿で提案する食材オントロジの有用性を確認するため，レシピサイトで実際に記載されている食材名が各オントロジにどの程度登録できているか確認するための評価実験を行う．

1：クックパッド，https://cookpad.com/（2019/3/22 閲覧）
2：楽天レシピ，https://recipe.rakuten.co.jp/（2019/3/22 閲覧）
3：特許情報プラットフォーム，https://www.j-platpat.inpit.go.jp/web/all/top/BTmTopPage（2019/3/22 閲覧）

2 関連研究

これまで，食材に着目したレシピシステムに関する研究が多くなされているが，それらの中では食材の表記に関する問題点が数多く指摘されている．例えば使用食材の優先度に基づいたレシピ検索や代替食材の発見，またレシピで使用される典型食材や調理手順の提示に関する研究が行われてきた [1–4]．これらの研究では，食材の表記ゆれや表記粒度の違いが提案手法の精度を低下させる要因になるため，これらに対応する必要があると述べられている．

上記で述べた表記ゆれや表記粒度の違いといった問題に対し，食材の表記を統一するための研究が行われている．土居らは，カテゴリ・エントリ・食材名の 3 階層を持つ料理オントロジを構築することで，食材の表記に関する問題に対応している [5]．料理オントロジを用いて，食材の上位・下位概念や同義語の関係性を示し，食材の表記統一を可能にすることで，上記の問題を解決することができる．料理オントロジを構築するため，以下の三つの処理を行っている．

- 楽天レシピの定義を利用したカテゴリの設定
- 特許データからのエントリ抽出
- 楽天レシピからの食材名抽出

カテゴリの設定は，楽天レシピのカテゴリを参考に一部を拡張し，「魚介」，「肉」，「野菜」，「その他」，「調味料」の 5 種類と定義している．エントリとは「あじ」や「ニンジン」といった食材名の代表表記のことを指す．特許データから抽出したカテゴリとエントリ間における上位・下位関係をエントリ候補とし，エントリの選定は人手で行っている．また食材名は，エントリに基づいて楽天レシピから関連語を収集し，関連語をエントリの

代替食材としたことがレシピの特徴となりうるかという基準を用いて人手で選定している．エントリと食材名は料理レシピから人手で抽出しているため，食材の表記ゆれやユーザによる表記の違いに対応している．

しかし，土居らが構築した料理オントロジの問題点として以下の四つがある．

（1） 食材名の選定基準が客観的ではない
（2） エントリと食材名の選定が人手である
（3） カテゴリの分類が適切でない
（4） エントリ数が十分でない

問題点 (1) は食材名候補がエントリの代替食材となりうるかどうかという主観的な基準で食材名の選定が行われているため，客観的な食材名の選定基準が必要である．問題点 (2) は，エントリと食材名の選定は人手で判別されているため，ヒューマンエラーが発生する可能性があり，また判別自体に多大なコストがかかる．問題点 (3)(4) については，料理オントロジのカテゴリに含まれるエントリ候補の決定は恣意的に行われており，全ての食材を網羅できるエントリ数であると言い難い．また，料理オントロジの構築方法は人的コストがかかり，ミスも発生しやすいためオントロジ構築の自動化が必要であると考える．

3 提 案 手 法

本稿では，文部科学省が公表している成分表の食品名をカテゴリ/エントリ/食材名の 3 階層に分け，食材オントロジを自動構築する．カテゴリについては，成分表で定められている食品群，および副分類を用いる．エントリと食材名の抽出には，成分表データから食材名を自動抽出する処理 1 を行った後，外部データを用いて食材名の表記を拡張する処理 2 を行う．本節では 3.1 節で使用するデータ，3.2 節と 3.3 節で処理 1 と処理 2 のそれぞれについて詳説する．

3.1 使用データ

本稿では，食材オントロジの構築のために，文部科学省科学技術・学術審議会資源調査分科会[4]が調査して公表している成分表を用いる．成分表は，栄養指導や生活習慣病予防などの観点から，日常的な食品の成分がまとめられたデータである．学校給食や病院などの給食の場や食事療法の問題を抱える一般家庭でも活用されているほか，教育・研究や行政においても広く活用されているデータであるため，あらゆる食材を網羅している．また成分表は，文部科学省が公表しているデータであることから，事前に校正がされているため，ヒューマンエラーが発生する可能性が低く，誤りがあった場合は修正が行われる．そのため，食材オントロジのカテゴリやそれに属する食材を抽出する情報源として信憑性が高い．

4：資源調査分科会：文部科学省，http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu3/index.htm（2019/3/22 閲覧）

表 1 成分表における食品名の分類

食品群	副分類	区分	大分類	中分類	小分類	細分
穀類			こむぎ	［小麦粉］	強力粉	一等
いも及びでん粉類	＜いも類＞		じゃがいも		塊根	蒸し
野菜類			あしたば		茎葉	ゆで
魚介類	＜魚類＞	（かに類）	がざみ		生	

3.2 成分表データを用いた食材名の抽出方法

成分表は表 1 に示す 7 階層で定義されている．しかし，成分表の食品名は 7 階層が明示されていないため，副分類や区分などを示す記号をもとに階層に分ける．また，成分表の備考欄には別名の記載があるため，食品名と対応さ食材オントロジを構築するデータとして使用する．本稿では，図の手順でカテゴリ/エントリ/食材名の 3 階層に分類する．

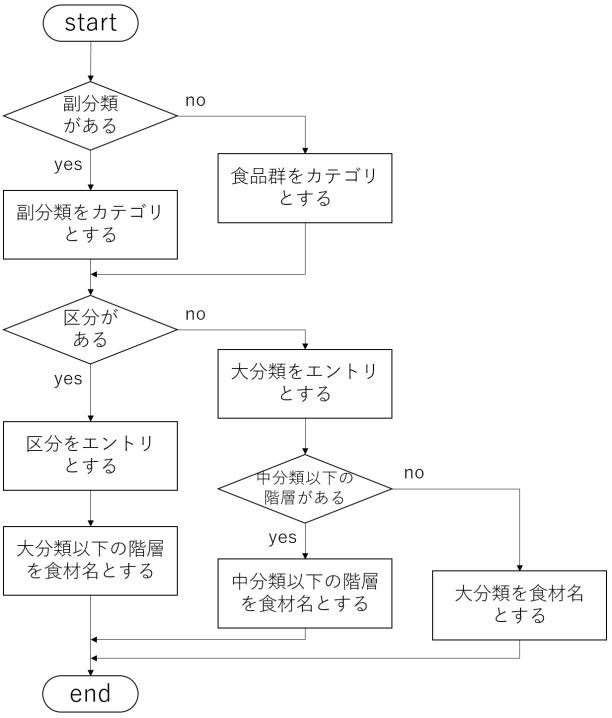

図 1 3 階層の分類手順

しかし，食材名となり得る小分類や細分には「塊根」や「生」などの食材以外の文字列も含まれるため，食材のみを抽出するため，以下の三つの手順を行う．

処理 1-1 成分表内の食品名に対する前処理
処理 1-2 成分表から最下層の重複文字列を抽出し，削除
処理 1-3 重複行を削除

処理 1-1 では，成分表データの食品名に対する前処理を行う．分類を示す＜＞や [] といった記号 (表 1 参照) や食品群や区分に頻出する「類」および，小分類や細分にある「塩分無調整タイプ」や「カスタードクリーム入り」などの食材ではない文字列の削除処理を行う．

処理 1-2 は，食品名の最下層から，重複している単語を削除する処理を指す．成分表の食品名の中には「生」や「蒸し」などの状態,「塊根」や「茎葉」といった可食部位の表記が存在する．また同階層の中に食材と状態や可食部位といった食材以外のものが混在しているため，食材オントロジの食材名を抽出す

るためには，食材となる階層を決める必要がある．成分表では，細分や小分類などの下層に「生」や「蒸し」といった状態が多く含まれる傾向がみられる.「塊根」や「茎葉」といった可食部位については，食品名が異なっても共通する場合がある．これらのことから成分表の食品名の最下層から，重複している単語を削除する処理を繰り返し，食材となる階層を決める．具体的な例を図 2〜図 6 に示す.

食品群	副分類	区分	大分類	中分類	小分類	細分
いも及びでん粉類	＜いも類＞	（さつまいも類）	さつまいも		塊根	皮つき
いも及びでん粉類	＜いも類＞	（さつまいも類）	さつまいも		塊根	皮むき
いも及びでん粉類	＜いも類＞	（さつまいも類）	さつまいも		蒸し切干	
いも及びでん粉類	＜いも類＞	（さつまいも類）	むらさきいも		塊根	皮つき
いも及びでん粉類	＜いも類＞	（さつまいも類）	むらさきいも		塊根	皮むき

図 2　いも及びでん粉類の例

図 2 はいも及びでん粉類の例である．処理 1-1 を行い，記号及び「類」の削除を行う.

食品群	副分類	区分	大分類	中分類	小分類	細分
いも及びでん粉	いも	さつまいも	さつまいも		塊根	皮つき
いも及びでん粉	いも	さつまいも	さつまいも		塊根	皮むき
いも及びでん粉	いも	さつまいも	さつまいも		蒸し切干	
いも及びでん粉	いも	さつまいも	むらさきいも		塊根	皮つき
いも及びでん粉	いも	さつまいも	むらさきいも		塊根	皮むき

図 3　処理 1-1 の前処理を行った食品名

処理 1-1 で記号と「類」を削除した成分表の食品名が図 3 となる．最下層の文字列を灰色で示している．処理 1-2 により最下層の「皮つき」,「皮むき」が重複するため，これらの文字列を削除し，図 4 のような中間結果を得る.

食品群	副分類	区分	大分類	中分類	小分類	細分
いも及びでん粉	いも	さつまいも	さつまいも		塊根	
いも及びでん粉	いも	さつまいも	さつまいも		塊根	
いも及びでん粉	いも	さつまいも	さつまいも		蒸し切干	
いも及びでん粉	いも	さつまいも	むらさきいも		塊根	
いも及びでん粉	いも	さつまいも	むらさきいも		塊根	

図 4　重複語「皮つき」,「皮むき」を削除した結果

図 4 では「塊根」の行が重複するため，処理 1-3 を実行することで，重複行を削除する．これにより，得られた中間結果が図 5 となる.

食品群	副分類	区分	大分類	中分類	小分類	細分
いも及びでん粉	いも	さつまいも	さつまいも		塊根	
いも及びでん粉	いも	さつまいも	さつまいも		蒸し切干	
いも及びでん粉	いも	さつまいも	むらさきいも		塊根	

図 5　重複行を削除した結果

図 5 では最下層の中で「塊根」が重複するため，処理 1-2 を実行し，これらの文字列を削除する．重複語を削除した結果を図 6 に示す.

食品群	副分類	区分	大分類	中分類	小分類	細分
いも及びでん粉	いも	さつまいも	さつまいも			
いも及びでん粉	いも	さつまいも	さつまいも		蒸し切干	
いも及びでん粉	いも	さつまいも	むらさきいも			

図 6　重複語「塊根」を削除した結果

図 6 の最下層から「さつまいも」「蒸し切干」「むらさきいも」が抽出できるが，重複する文字列がないため，削除処理を終了する．図 6 の例では最終的に「いも/さつまいも/さつまいも」,「いも/さつまいも/さつまいも/蒸し切干」,「いも/さつまいも/むらさきいも」が抽出でき,「いも」がカテゴリ,「さつまいも」がエントリ,「さつまいも」「蒸し切干」「むらさきいも」が食材名となる.

3.3　外部データを用いた食材名表記の拡張

3.2 節で抽出した食材名の表記を拡張する処理 2 について説明する．成分表のデータは食品名の表記がほとんどひらがなであり,「食塩」や「食酢」,「かんきつ」などのレシピでは使われない表記が存在する．表記の種類が少なく，レシピでは使われない表記があるため，成分表の表記のみで食材オントロジを構築すると，表記の多様性に欠ける可能性がある．表記の多様性に対応するため，処理 2 では以下の 3 種類の処理を行う.

処理 2-1　ひらがな・カタカナ変換を行う Python モジュール pykakasi[5]を用いてひらがな⇔カタカナ変換

処理 2-2　Yahoo! かな漢字変換 API[6]を用いてかな⇒漢字変換

処理 2-3　形態素解析器 JUMAN++ [11] を用いた食材の一般表記の自動抽出

処理 2-2，2-3 は表記の種類を増やしてから行うことで，より多くの食材名表記を抽出できるため，処理 2-1 のひらがな・カタカナ変換を最初に行う．処理 2-2 の漢字変換は予測候補を出力するため，食材以外の表記も多く含まれている．そのため，処理 2-3 で JUMAN++を用いて形態素解析を行い,「人工―食べ物」もしくは「料理・食事」という情報が付与された単語のみを抽出する．JUMAN++を使用する理由は，他の形態素解析器と比べ精度が高く，食べ物や料理に関する情報を付与することができるためである [11, 12]．処理 2-1 のデータに対しても処理 2-3 を行うことで食材の一般表記を抽出することができる．一般表記の処理で抽出できた食材には「アイス」や「米」などがあった.

3.2 節の処理で，自動抽出された食材は 1,657 件であったが，食材名表記の拡張を行うことで最終的には 3,586 件になった.

5：pykakasi，https://github.com/miurahr/pykakasi（2019/3/22 閲覧）
6：テキスト解析：かな漢字変換，https://developer.yahoo.co.jp/webapi/jlp/jim/v1/conversion.html（2019/3/22 閲覧）

4 評価実験

本節では，3.2 節で提案した食材を自動抽出する手法を評価するため，食材の抽出精度の算出を行う．また，食材名の網羅性を評価するため，先行研究と本研究で提案したそれぞれのオントロジがクックパッドの食材をどの程度含むことができたか比較実験を行い，その結果から考察を行う．

4.1 食材の自動抽出手法の精度評価

食材を自動抽出する手法の精度評価を行うため，適合率，再現率，F 値を計算する．適合率は自動抽出した食材の中で，どの程度正しい成分表の食材が含まれているかどうかを判断する指標であり，再現率は正しい成分表の食材の中で，どの程度正しく食材を自動抽出できたかどうかを判断する指標である．適合率と再現率は一方が高ければ，もう一方が低くなるというトレードオフの関係にあるため，適合率と再現率の両方を総合的に評価する指標として F 値という指標がある．表 2 に自動抽出食材と成分表の食材における正解・不正解の関係を示す．

表 2　自動抽出食材と成分表の食材における正解・不正解の関係

		成分表の食材	
		正解	不正解
自動抽出食材	正解	tp	fp
	不正解	fn	tn

適合率 P，再現率 R はそれぞれ式 (1)，(2) で算出する．

$$P = \frac{tp}{tp + fp} \tag{1}$$

$$R = \frac{tp}{tp + fn} \tag{2}$$

F 値は適合率と再現率の調和平均により，式 (3) から算出する．

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

重複語削除後のデータ 1,657 件と，そのうち実際に食材であった 1,266 件，成分表の正解データ 1,317 件を用いて計算を行った結果，再現率，適合率，F 値はそれぞれ 0.96, 0.76, 0.85 であった．F 値から高い精度で食材の自動抽出を行えたことが確認できたが，適合率から自動抽出した食材のうち 24%が食材でなかったことがわかる．食材ではない文字列には「長期熟成」や「土耕栽培」などがあり，これらの文字列は重複しないため，削除できなかった．

4.2 各オントロジの網羅性評価

料理オントロジと食材オントロジの食材名の網羅性を評価するため，クックパッドの食材名を各オントロジがどの程度含んでいるかを調べる．

クックパッドは国内最大のレシピサイトであり，レシピ投稿数が最も多い．そのため，クックパッドデータの食材名を網羅できれば，他のサイトの食材も網羅できると考え，食材含有率を算出するデータとしてクックパッドデータセットを用いる．しかし，クックパッドデータの食材には,「大根おろし」や「溶き卵」などの加工名と食材からなる複合名詞や「ズッキーニまたは茄子」といった代替食材を指定する表現が存在する．上記で述べた問題点があるため，形態素解析器を用いて形態素解析を行った上で，食材オントロジの食材に含まれるかを調べる．形態素解析を行うことで，複合名詞からなる食材は「大根/おろし」や「溶き/卵」，代替食材を指定する表現は「ズッキーニ/または/茄子」に分けることができる．これにより「大根」や「卵」,「ズッキーニ」,「茄子」といった食材を認識可能になる．方法としては，食材オントロジの食材名を JUMAN++のユーザ辞書に登録し，食材オントロジに登録されていることが区別できるラベルが付与されるようにする．料理オントロジに対しても 3.3 節で述べた食材名表記の拡張を行い，ユーザ辞書への登録を行う．各オントロジの網羅性は式 (4) に示す食材含有率 C で評価する．

$$C = \frac{n}{N} \tag{4}$$

ただし，n はクックパッドデータ [13] 内に含まれる各オントロジの食材名数（食材含有数），N はクックパッドデータ内の全食材名数 103,753 件である．

表記の拡張処理の効果も調べるため，処理ごとに食材含有率を算出する．表記の拡張処理によるオントロジごとの食材含有率の変化を表 3 に示す．

表 3　表記の拡張処理による食材含有率の変化

処理順序		–	1	2	3	4
処理内容		処理なし	ひらがな	カタカナ	漢字変換	一般表記
既存	食材含有数	44,044	44,111	44,290	50,739	53,697
	食材含有率	0.425	0.425	0.427	0.489	0.518
提案	食材含有数	31,105	31,175	35,822	51,000	54,468
	食材含有率	0.300	0.300	0.345	0.492	0.525

料理オントロジはひらがなやカタカナ，漢字変換を行ってもそれほど食材含有率に変化はなかった．これは，料理オントロジでは人手で食材名の抽出が行われているため，表記の種類への対応は行われていたことが理由として考えられる．食材オントロジもひらがな変換では食材含有率に変化はなかったが，カタカナや漢字変換を行うことで大幅に食材含有率が向上した．ひらがな変換が食材含有率の向上に効果的でなかった理由は，カタカナで書かれている食材数が少なかったことと，レシピ内でカタカナの食材をひらがなで書くことが少ないためであると考える．

本稿で構築した食材オントロジと既存の料理オントロジを比較した結果とを表 4 に示す．登録食材数とは表 3 で示した処理を行っていない食材名の総数を表し，拡張食材数は一般表記の抽出までの処理を行った食材名の総数を表す．2 節で問題点に挙げていたエントリ数は食材オントロジの方が多くなったが，食材オントロジはエントリに伴う食材名の数が少ないため，登録食材数は料理オントロジの方が多くなった．また，表記の拡張を行った拡張食材数も料理オントロジの方が多くなり，最終的な食材含有率は料理オントロジの方が高くなった．

食材オントロジに含まれていない食材名は,「オニオン」や「ガーリック」といった外来語があった．料理オントロジに含

表 4　料理オントロジと食材オントロジの比較結果

オントロジ	エントリ数	登録食材数	拡張食材数
既存	295	3,375	5,372
提案	409	1,657	3,586

まれており，食材オントロジに含まれていない食材は 7,538 件あり，そのうち外来語が含まれている食材は 3,502 件であった．どちらのオントロジにも含まれなかった食材は,「日清フラワー」や「雪印ネオソフト」などの商品名であった．

5 おわりに

本稿では，レシピの食材表記に関する問題を解決できる食材を網羅したオントロジを構築するため，成分表を用いて食材オントロジを自動構築する方法を提案した．また，成分表から食材を自動抽出する方法の抽出精度評価と，クックパッドデータセットの食材名を用いて各オントロジごとに比較実験を行った．その結果，本研究で提案した最下層文字列を削除する方法を用いることで，F 値が 8 割を越える高い精度で食材を抽出することができた．食材名の網羅性評価においては表記の拡張を行うことで，食材オントロジの食材含有率が既存の料理オントロジの食材含有率に近づくこと確認した．しかし，成分表から抽出した食材名は外来語や商品名に対応できていないため，食材含有率は低くなった．

今後の課題として，外来語や商品名に対応することが挙げられる．これにより，食材含有率を更に向上させることができると考える．また，本稿で取り扱わなかった状態や可食部位についても関係性を構築し，食材オントロジの適用可能範囲を広げる必要がある．

謝　辞

本研究は JSPS 科研費 JP16H02908 の助成を受けたものである．また本研究では，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．ここに記して謹んで謝意を表す．

文　献

[1] 塩澤秀和, 三田村祐介. 食材の優先度を考慮した料理レシピの検索. 情報処理学会研究報告ヒューマンコンピュータインタラクション (HCI), Vol. 2007, No. 41, pp. 51–57, 2007.

[2] 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. *DEIM Forum 2015*, pp. G6–6, 2015.

[3] 志土地由香, 井手一郎, 高橋友也, 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, Vol. 94, No. 7, pp. 532–535, 2011.

[4] 難波英嗣, 土居洋子, 辻田美穂, 竹澤寿幸, 角谷和俊. 複数料理レシピの自動要約. 電子情報通信学科技術研究報告, Vol. 113, No. 338, pp. 39–44, 2013.

[5] 土居洋子, 辻田美穂, 難波英嗣, 竹澤寿幸, 角谷和俊. 料理レシピと特許データベースからの料理オントロジーの構築. 電子情報通信学会技術報告, Vol. 113, No. 468, pp. 37–42, 2014.

[6] Nicola Guarino and Pierdaniele Giaretta. Ontologies and Knowledge Bases: Towards a Terminological Clarification. *Towards Very Large Knowledge Bases*, pp. 25–32, 1995.

[7] 文部科学省科学技術・学術審議会資源調査分科会. 日本食品標準成分表 2015 年版（七訂）. 全国官報販売協同組合, 2015.

[8] 文部科学省科学技術・学術審議会資源調査分科会. 日本食品標準成分表 2015 年版（七訂）追補 2016 年. 全国官報販売協同組合, 2016.

[9] 文部科学省科学技術・学術審議会資源調査分科会. 日本食品標準成分表 2015 年版（七訂）追補 2017 年. 全国官報販売協同組合, 2017.

[10] 文部科学省科学技術・学術審議会資源調査分科会. 日本食品標準成分表 2015 年版（七訂）追補 2018 年. 全国官報販売協同組合, 2018.

[11] *Morphological analysis for unsegmented languages using recurrent neural network language model*. Conference on Empirical Methods in Natural Language Processing, 2015.

[12] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN++ version 1.02, 2016.

[13] クックパッド株式会社. クックパッドデータ. 国立情報学研究所情報学研究データリポジトリ（データセット）, 2015. https://doi.org/10.32130/idr.5.1.