一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

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

土居 洋子† 辻田 美穂† 難波 英嗣† 竹澤 寿幸† 角谷 和俊‡

†広島市立大学情報科学部 〒731-3194 広島市安佐南区大塚東 3-4-1
‡兵庫県立大学環境人間学部 〒670-0092 兵庫県姫路市新在家本町 1-1-12
E-mail: †{doi, tsujita, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp, ‡sumiya@shse.u-hyogo.ac.jp

あらまし 料理オントロジーとは，料理レシピを言語処理するために必要な言語資源である．本研究では，統計的言語処理技術を用いて，料理レシピと特許データベースから，用語の上位・下位関係，同義語，属性，部分・全体関係を抽出し，人手で選定することで，料理オントロジーを構築した．

キーワード 料理オントロジー，料理レシピ，上位・下位関係，同義，属性，部分・全体関係

Construction of a Cooking Ontology from Cooking Recipes and Patents

Yoko DOI† Miho TSUJITA† Hidetsugu NANBA†
Toshiyuki TAKEZAWA† and Kazutoshi SUMIYA‡

†Faculty of Information Sciences, Hiroshima City University
3-4-1 Ozukahigashi, Asaminamiku, Hiroshima 731-3194 Japan
‡School of Human Science and Environment, University of Hyogo
1-1-12 Shinzaike–honcho, Himeji, Hyogo 670-0092 Japan
E-mail: †{doi, tsujita, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp, ‡sumiya@shse.u-hyogo.ac.jp

Abstract A cooking ontology is an inevitable language resource for language processing of cooking recipes. We constructed the cooking ontology using statistical natural language processing techniques for identifying hyponymy, synonymy, attributes, and meronymy.

Keyword Cooking Ontology, Cooking Recipe, Hyponymy, Synonymy, Attribute, Meronymy

1. はじめに

近年，料理レシピを対象にした様々な言語処理研究が増加している．料理レシピの要約や料理レシピを対象にした情報抽出などがその一例である．しかし，料理レシピでは固有の言い回しや表現の省略が多く存在することや，ユーザ投稿型レシピサイトでは料理レシピ中で用いる料理用語の表記が投稿者によって異なることなどが，料理レシピを十分な精度で解析できない大きな要因となっていた．そこで本研究ではこれらの問題を改善するために，言語処理を行う際の知識体系として利用される料理オントロジーの構築を試みる．

これまでに，自然言語処理分野では，テキストデータベースから同義語，関連語，用語の上位・下位関係などを抽出する手法が数多く提案されている．本研究では，テキストデータベースに，料理レシピと特許を用いる．近年，ユーザ投稿型レシピサイトなどで公開されている大量の料理レシピデータの一部は研究目的でも利用可能になっているため，料理オントロジーのための情報源のひとつとして，料理レシピデータを用いる．ここで，上述のとおり，料理レシピには表現の省略が多く存在するため，料理レシピデータのみを対象にオントロジーを構築すると，オントロジーとしての網羅性に欠ける可能性がある．この問題を回避するため，本研究では特許にも着目する．特許明細書には，権利の範囲をより明確にすることで特許権侵害訴訟を回避するために，自明なことであっても明示的に記述することが多い．そこで，特許データベースを，料理オントロジーを構築するためのもうひとつの情報源として利用することで，人間にとっては自明な知識も抽出できる可能性がある．本研究では，料理レシピと特許を情報源とし，同義語抽出手法や上位・下位関係の抽出手法などをこれらのテキストデータに適用することで，効率的な料理オントロジーの構築を目指す．

本論文の構成は以下のとおりである．2 節では関連研究，3 節では料理オントロジーの構築手法，4 節では実験について述べ，5 節で本稿をまとめる．

2. 関連研究

本節では，本研究に関連する研究として，英語版料理シソーラス,料理レシピを対象にした言語処理研究,テキストデータベースからの用語の抽出について，それぞれ 2.1 節，2.2 節，2.3 節で述べる．

2.1. 英語版料理シソーラス

本研究で構築する料理オントロジーと類似したものとして,英語版料理シソーラス「Cook's Thesaurus[1]」がある．これは,17 種類のカテゴリで構成されており,各エントリには，食材の画像，同義語，発音，説明，代替材料，保存方法などが記載されている．この英語版料理シソーラスは，世界中の食材を網羅することを目的としているが，本研究では，言語処理精度改善のために，日本語の料理レシピに特化して料理オントロジーを構築する目的としている点で異なる．

2.2. 料理レシピを対象にした言語処理研究

難波ら[1]は，複数テキスト要約の技術を用いて，特定の料理に関する複数の料理レシピから，その料理で用いる典型的な材料と調理手順を出力する手法を提案している．難波らは，複数料理レシピ要約の作成における考慮すべき点として，表記の揺れと表現の省略を挙げている．これらについて，本研究で構築する料理オントロジーの同義語辞書，属性辞書，部分辞書を用いることで,言語処理精度を改善できると考えられる．

橘ら[2]は,レシピタイトルの特徴を表す「簡単」,「子供が喜ぶ」,「ヘルシー」といった修飾表現に着目し，それらの修飾表現の根拠をネーミングコンセプトと定義して,料理レシピから抽出する手法を提案している．橘らは，ネーミングコンセプトの抽出において材料と調理器具の抽出を行っている．この抽出において，本研究で構築する料理オントロジーの同義語辞書を用いることで,言語処理精度の改善ができると考えられる．

2.3. テキストデータベースからの用語の抽出

テキストデータベースから，上位・下位関係を抽出する代表的な手法として，Hearst[3]のものがある．Hearst は，「A や B などの(等の)C」といった定型表現に着目することで，「A と B の上位関係は C である」という用語の上位・下位関係を抽出する手法を提案している．本研究では，Hearst の手法を用いて，エントリ辞書の構築を行う．エントリ辞書の構築については 3.1 節で述べる．

テキストデータベースから，関連語を収集する代表的な手法として，分布類似度[4][5][6]がある．分布類似度とは，「2 つの用語 A と B が意味的に類似していれば，A と B の文脈に出現する語の傾向が似ている」という仮定に基づいた関連語収集手法である．文脈語の選定には，全単語を用いる手法，内容語のみを用いる手法などが考えられる．本研究では，相澤の手法に従い，対象となる語と係り受け関係にある動詞を文脈語として利用する．この手法は，ある動詞に着目し，その動詞と係り受け関係にある名詞を文脈語と考えれば，動詞の関連語の収集も可能になる．一方，料理レシピから関連語を収集する手法として，Chung[7]の手法がある．Chung は,「料理レシピでは，ある料理で使用する材料のうち，主要なものから順に材料リストに記載する傾向がある」という特徴を利用する関連語収集手法を提案している．例えば，楽天レシピ[2]のように各料理レシピが材料ごとに階層的に分類されている場合には,「エビ」カテゴリに分類されている料理レシピの材料リストの先頭に記載されている材料を収集することで,「エビ」の関連語が効率的に収集できる．本研究では，分布類似度と Chung の手法を用いて，同義語辞書の構築を行う．同義語辞書の構築については 3.2 節で述べる．

3. 料理オントロジーの構築

本節では，料理オントロジーの構築手法について述べる．料理オントロジーの構築は，以下の 5 つの Step から構成される．

- Step 1 概念辞書の構築
- Step 2 カテゴリの設定
- Step 3 エントリ辞書の構築
- Step 4 同義語辞書の構築
- Step 5 属性辞書，部分辞書の構築

Step 1 として，概念辞書の構築を行う．概念辞書とは，料理オントロジーの概念階層のことを指す．本研究で構築する概念辞書は,「カテゴリ–エントリ–同義語」の 3 階層とする．

Step 2 として,カテゴリの設定を行う.カテゴリは,楽天レシピのカテゴリを参考に一部拡張して「材料–魚介」,「材料–肉」,「材料–野菜」,「材料–その他」,「調味料」,「調理器具」,「動作」の 7 種類とする．Step 3, Step 4，Step 5 について，それぞれ 3.1 節，3.2 節，3.3 節で述べる．

[1] http://www.foodsubs.com/

[2] http://recipe.rakuten.co.jp/

3.1. エントリ辞書の構築

本節では，エントリ辞書の構築について述べる．エントリ辞書の構築は,次の 2 つの Step から構成される．

- Step 3-1 上位・下位関係の抽出
- Step 3-2 エントリの選定

Step 3-1 では,2.3 節で述べた Hearst の手法を用いて，特許から上位・下位関係の抽出を行う．例えば,「材料−魚介」カテゴリの下位関係を抽出するとき,「A や B 等の魚類」や「C 等の魚介類」という表現から，A，B，C といった用語をエントリの候補として抽出する.「魚類」や「魚介類」に相当する用語は，カテゴリごとに次の用語を利用する．

- 材料−魚介：魚類，魚介類，海産物，水産物
- 材料−肉　：肉類，食肉，食肉類，原料肉
- 材料−野菜：野菜，果菜類，野菜類，果菜物，農産物
- 調味料　：調味料，香辛料，薬味，スパイス類
- 調理器具：調理器具，調理容器，調理器，調理具，調理道具

Step 3-2 では，Step 3-1 で抽出された用語を頻度順に並べ，頻度の高い用語から順にエントリを選定する．この際,「材料−その他」については，Step 3-1 で挙げた用語を使って収集した結果，いずれにも当てはまらなかった用語をエントリの候補とした.「動作」については，料理レシピに出現する動詞を頻度順に並べ，頻度の高い動詞から順にエントリを選定する．

3.2. 同義語辞書の構築

本節では，同義語辞書の構築について述べる．同義語辞書の構築は，以下の 2 つの Step から構成される．

- Step 4-1 関連語の収集
- Step 4-2 同義語の選定

Step 4-1 では，2.3 節で述べた分布類似度を用いて幅広く関連語を収集する手法と 2.3 節で述べた Chung の料理レシピの記載傾向を用いて効率的に収集する手法を組み合わせて関連語の収集を行う．分布類似度による関連語の収集について述べる．係り受け解析器 CaboCha[3]を用い，すべての料理レシピを構文解析する．得られた解析木から，係り受け関係にある名詞と動詞の対を抽出する．次に，名詞ごとに，係り受け関係にある動詞の頻度を数え，共起語ベクトルを作成する．与えられた名詞に対し，共起語ベクトル間の類似度を計算する尺度として，コサイン距離を利用する．

Step 4-2 では，Step 4-1 で収集した関連語を人手で選定する．人手での同義語の選定は，代替したことが料理レシピの特徴となりうるかを基準とする．例えば,「ピーマン」と「パプリカ」について,「ピーマンの肉詰め」を作る際,「ピーマン」の代わりに「パプリカ」を使用しても，料理レシピの特徴になるとは考えにくい．よって,「ピーマン」と「パプリカ」は同義語とする．一方,「豚肉」と「鶏肉」について,「酢豚」を作る際,「豚肉」の代わりに「鶏肉」を使用することは，料理レシピの特徴になりうる．よって,「豚肉」と「鶏肉」は同義語としない．以上の基準に基づいて同義語の選定を行う．

3.3. 属性辞書，部分辞書の構築

本節では，属性辞書と部分辞書の構築手法について述べる．各辞書の構築は，以下の 3 つの Step から構成される．

- Step 5-1 定型表現「A の B」の収集
- Step 5-2 属性，部分に関する用語の収集
- Step 5-3 属性，部分の選定

Step 5-1 では，料理レシピから定型表現「A の B」の収集を行う．本研究では,「A の B」という定型表現に着目する．例えば,「サバの色」という表現で「色」はサバの属性,「サバの皮」は部分・全体関係である．本研究では，定型表現「A の B」を収集すれば，効率的に属性，部分・全体関係を収集できると仮定し，料理レシピから定型表現「A の B」を収集する．

Step 5-2 では,Step 5-1 で収集した定型表現「A の B」から属性，部分に関する用語の収集を行う．例えば,「サバ」の属性，部分に関する用語を収集する際,「サバの B」という表現から，B の用語を属性，部分の候補として収集する．

Step 5-3 では，Step 5-2 で収集した B の用語を人手で選定する．人手での属性，部分の選定は以下の定義に基づいて行う．まず，属性の定義について述べる．本研究では,属性を「色が変わったら鍋から取り出す」のように，変化することで次の手順にうつる基準となる用語，もしくは,「形を崩さないように煮る」のように，料理におけるコツやポイントを示す用語と定義する．次に，部分の定義について述べる．本研究では，部分を，全体を構成する要素，パーツと定義する．例えば,「サバ」を全体としたとき,「皮」,「骨」,「身」が部分となる．

提案手法

属性，部分の選定を効率的に行うため，Step 5-2 で収集した用語を，テキストデータに出現する定型表現「A の B」の頻度順に並べて選定する手法を提案する．

[3] http://code.google.com/p/cabocha/

テキストデータは，特許と料理レシピを用いる．特許は表現の省略や表記の揺れが少ないことから，より定量的な選定が可能であると考えられる.属性選定手法，部分選定手法について，それぞれ次の 4 つの手法を提案する.

- **特許 A**：特許に出現する定型表現「A の B」の A に該当する表現を頻度順に並べて選定する手法
- **特許 B**：特許に出現する定型表現「A の B」の B に該当する表現を頻度順に並べて選定する手法
- **料理レシピ A**:料理レシピに出現する定型表現「A の B」の A に該当する表現を頻度順に並べて選定する手法
- **料理レシピ B**:料理レシピに出現する定型表現「A の B」の B に該当する表現を頻度順に並べて選定する手法

4. 実験

本節では，本研究で行った実験について述べる．4.1 節では，3.2 節で述べた同義語辞書の構築とその結果について述べる．4.2 節では，3.3 節で提案した属性，部分の選定手法の有効性を調べるために行った実験について述べ，実験結果を考察する.

4.1. 同義語辞書

データセット

まず，エントリ辞書の構築を行った．3.1 節で述べた手法を用いて，3 節で設定したカテゴリ 7 種類について，特許公開公報(1993～2011)から上位・下位関係の抽出を行った．抽出した用語を頻度順に並べ，頻度の高い用語から順にエントリを選定した．エントリ辞書の構築結果を表 1 に示す.

表 1　エントリ辞書の構築結果

カテゴリ	エントリ数(語)
材料−魚介	61
材料−肉	6
材料−野菜	122
材料−その他	55
調味料	51
調理器具	48
動作	131
合計	474

次に，選定したエントリをクエリとして，関連語の収集を行った．2.3 節，3.2 節で述べた Chung の手法と分布類似度を用いて，楽天レシピに投稿された料理レシピ約 44 万件から関連語を収集した.収集した関連語数を表 2 に示す.

表 2　Chung の手法と分布類似度で収集した関連語数

カテゴリ	Chung の手法(語)	分布類似度(語)
材料−魚介	216	1,247,767
材料−肉	265	208,390
材料−野菜	379	3,485,494
材料−その他	241	1,542,022
調味料	24	1,665,538
調理器具	0	431,382
動作	0	88,822

構築結果

収集した関連語のうち，類似度の高い関連語を，3.2 節で述べた選定基準に基づき，人手で選定した．同義語辞書の構築結果を表 3 に示す.

表 3　同義語辞書の構築結果

カテゴリ	同義語数(語)
材料−魚介	453
材料−肉	383
材料−野菜	947
材料−その他	732
調味料	909
調理器具	643
動作	956
合計	5,023

4.2. 属性，部分の選定

データセット

まず，楽天レシピに投稿された料理レシピ約 44 万件から，定型表現「A の B」の収集を行った．次に，収集した定型表現から，カテゴリ「材料−魚介」の同義語 453 語をクエリとして，属性，部分に関する用語を 717 語収集した.さらに,収集した用語 717 語から，料理名を除外した 453 語を人手で選定した．人手での選定結果を表 4 に示す．表 4 の選定結果に対し，属性選定手法，部分選定手法について，それぞれ表 5，表 6 のデータを用いた.

3.3 節で述べた特許と料理レシピについて述べる．特許データは，国際特許分類(IPC) のサブクラスレベルで A23L(食品、食料品)、A47J(台所用具)、H05B(電気加熱)が筆頭 IPC(ひとつの特許明細書に付与される複数の分類コードの中で一番重要なもの)として付与された料理分野の特許明細書(1993～2012 年公開特許公報) 91,736 件を用いた．料理レシピデータは，楽天レシピに投稿された料理レシピ約 44 万件を用いた.特許，料理レシピ，それぞれに出現する定型表現「A の B」とその頻度を用いた.

評価ツールには，評価ワークショップ TREC(Text REtrieval Conference)で使われる trec_eval を用いた．このツールを用い，再現率が 1 になったときの精度の値を評価値として算出した．これは，用語を漏れなく収集したとき(再現率が 1)に，どの程度収集した用語に正しいものが含まれているのかを評価するためである．

表 4　人手で属性か部分か該当なしか選定した結果

属性(語)	部分(語)	該当なし(語)	合計(語)
146	144	163	453

表 5　属性選定手法の実験データ

正解(語)	不正解(語)	合計(語)
146	307	453

表 6　部分選定手法の実験データ

正解(語)	不正解(語)	合計(語)
144	309	453

実験結果と考察

属性選定手法，部分選定手法について，実験結果をそれぞれ表 7，表 8 に示す．表 7 の実験結果より，属性選定手法について，属性，部分に関する用語を，特許に出現する定型表現「A の B」の B における頻度順に並べて選定する手法で最も高い精度 0.493 を獲得した．よって，属性選定手法において，特許 B 手法が最も有効であることがわかった．表 8 の実験結果より，部分選定手法について，属性，部分に関する用語を，料理レシピに出現する定型表現「A の B」の A における頻度順に並べて選定する手法で高い精度 0.257 を獲得した．よって，部分選定手法において，料理レシピ A 手法が有効であることがわかった．

表 7　属性選定手法の実験結果

手法	精度
特許 A	0.452
特許 B	0.493
料理レシピ A	0.411
料理レシピ B	0.452

表 8　部分選定手法の実験結果

手法	精度
特許 A	0.222
特許 B	0.201
料理レシピ A	0.257
料理レシピ B	0.215

まず，属性選定手法において，最も高い精度を獲得した特許 B 手法について考察を行う．表 9 に特許に出現する定型表現「A の B」の B における頻度の高い上位 10 語と人手による属性，部分選定結果を示す．表 9 より，頻度の高い上位 10 語のうち，3 語は属性，4 語は部分であることがわかる．このことから，あらかじめ部分の用語を除外することで，より選定の精度を改善できると考えられる．一方，特許に出現する定型表現「A の B」の A における頻度の高い用語を調べたところ，6 位が「状態」，8 位が「量」であることがわかった．このように，定型表現「A の B」の A，B どちらにも出現する理由として，「水の量」，「量の比率」といった階層的な表現が成り立つためと考えられる．この問題について，Hearst の上位・下位関係を抽出する手法を用いることで，階層的な表現を考慮できるようになると考えられる．

表 9　特許の B における頻度の高い上位 10 語

属性，部分に関する用語	属性選定結果	部分選定結果
場合	不正解	不正解
表面	不正解	正解
間	不正解	不正解
量	正解	不正解
状態	正解	不正解
水	不正解	不正解
上面	不正解	正解
種類	正解	不正解
部分	不正解	正解
面	不正解	正解

次に，部分選定手法において，高い精度を獲得した料理レシピ A 手法について考察を行う．部分の用語について調べたところ，「背わた」，「背ワタ」，「背綿」といった表記の揺れが多いことがわかった．これらの用語の料理レシピに出現する定型表現「A の B」の A における頻度はそれぞれ 1.060e-05，0，0 であった．このことから，あらかじめ分布類似度を用いて用語の表記の揺れに対応した上で頻度を求めることで，より選定の精度を改善できると考えられる．

5. おわりに

本研究では，統計的言語処理技術を用いることにより，効率的な料理オントロジー[4]の構築を試みた．分布類似度と Chung の手法を組み合わせることにより収集した関連語を人手で選定することで，同義語辞書を構築した．この結果，同義語 5,023 語を獲得した．属性，部分の選定では，定型表現「A の B」とその頻度に着

[4] http://www.ls.info.hiroshima-cu.ac.jp/cooking/ontology.html

目した．実験の結果，最も高い精度が得られた，特許に出現する定型表現「A の B」の B における頻度順に属性を選定する手法が最も有効であることがわかった．

今後の課題として，まず，辞書構築について，規模の拡大が挙げられる．一方，人手で用語を選定するコストの削減が挙げられる．森ら[8]は，材料などの固有表現認識に機械学習を用いている．本研究でも同様に，機械学習による用語の選定を検討する必要があると考えられる．次に，属性，部分の選定について，上位・下位関係を考慮する必要があると考えられる．一方，表記の揺れを考慮した上で頻度を求めることで，より選定の精度を改善できると考えられる．さらに，構築した料理オントロジーを用いて，料理レシピの言語処理を行うことで，精度が改善されるか実験し，本研究で構築した料理オントロジーの有効性を確認する必要がある．

謝辞

本研究を遂行するにあたり，解析対象となるレシピデータを楽天株式会社よりご提供いただいた．ここに記して謹んで感謝の意を表する．

文　献

[1] 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，”電子情報通信学会技術研究報告，Vol.113，No.338，NLC2013-41, pp.39-44, 2013.

[2] 橘明穂，若宮翔子，難波英嗣，角谷和俊，“料理名の修飾表現の関係性に基づくレシピのネーミングコンセプト抽出，”電子情報通信学会技術研究報告，Vol.113，No.214，DE2013-36，pp.19-24，2013.

[3] M. A. Hearst, “Automatic Acquisition of Hyponyms from Large Text Corpora,” Proc. 14th International Conference on Computational Linguistics, pp.539-545, 1992.

[4] D. Lin, “Automatic Retrieval and Clustering of Similar Words,” Proc. COLING/ACL1998, pp.768-774, 1998.

[5] L. Lee, “Measures of Distributional Similarity,” Proc. 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pp.25-32, 1999.

[6] 相澤彰子，“大規模テキストコーパスを用いた語の類似度計算に関する考察，”情報処理学会論文誌，Vol.49, No.3, pp.1426‐1436, 2008.

[7] Y. Chung, “Finding Food Entity Relationships Using User-generated Data in Recipe Service,” Proc. 21st ACM International Conference on Information and Knowledge Management (CIKM2012), pp.2611-2614, 2012.

[8] 森信介，山肩洋子，笹田鉄郎，前田浩邦，“レシピテキストのためのフローグラフの定義，”情報処理学会研究報告, Vol.2013-NL-214, No.13, pp.1-7, 2013.