地域住民から収集したひらがな料理名を漢字変換するツールの検討

A development of Japanese Kanji converter dictionary for the dietary recall survey

今枝奈保美[*1]，野崎浩成[*2]
Nahomi IMAEDA[*1], Hironari NOZAKI[*2]
[*1] 至学館大学健康科学部
[*1] Faculty of Wellness, Shigakkan University
[*2] 愛知教育大学
[*2] Aichi University of Education
Email: imaeda@sgk.ac.jp

あらまし：食事調査は，対象者に飲食物を想起させ，熟練栄養士が食品成分表の食品番号に関連付ける．本報は，この過程を自動化するために，住民が答えた表記揺れのある料理名（かな）を対象に，文字列の正規化，単語の区切りを検討し，漢字に変換するツールの開発を試みた．料理名は 404 人，4799 日分，延べ 80362 件（13209 種）の料理名である．結果，食品と調理法の間の助詞"と"，"の"は stop words，料理名の末尾"に"は動詞としての処理を要した．今後，食事調査に特異的な辞書の開発を検討したい．
キーワード：食事調査法，漢字変換ツール，地域住民，料理データベース

1. はじめに

生活習慣病の発症予防および重症化予防を目的に，食事摂取量の過不足や栄養素のバランスを評価する場合，摂取した料理や食品名をノートに文字として対象者が記録し，その後，熟練栄養士が面接して食事内容を確定していくプロセスがある．なぜ，栄養士に熟練が必要なのか，その理由は 2 点ある．1 つは，対象者の申告が不正確なので，一部の食品を忘れていたり，調味料が不明だったりするのを栄養士が補足するため，2 つめは，我が国の食生活が和洋折衷で，ジャンクフード，スイーツなど多種多彩な食品を，日本標準食品成分表に載っている概ね 1,850 種類に集約し，食材をすべて食品番号にコード化するスキルが必要なためである．私達の研究では，食事調査を簡便化するために，対象者が答えた話し言葉を，文字情報として認識させるツールを提案し検討する．本報では，実際の疫学調査で，地域住民が答えた表記揺れのある料理名（かな）を対象にして，既存の形態素解析がうまく認識できない事例を報告し，文字列の正規化，単語区切りの検討などの前処理を行い，漢字に変換するツールの開発を試みた．平仮名のデータを，一定のルールで漢字に変換すれば，食品成分表との関連付け（コード化）が可能になる．すべての飲食物を文字で記録する手間は，かなり面倒なので，このツールは食事調査の簡便化につながると考えている．

2. 検討対象になる食事データの概要

全国コーホート研究地区（山形市，山形鶴岡，千葉，静岡桜ヶ丘，浜松，名古屋，岡崎，滋賀高島，京都，佐賀，徳島）の男女 404 人を対象に 3 か月ごとに隔日 3 日間の食事記録調査を行って得た 4799 日分の料理名で，延べ 80362 件（13209 種）である．

3. 予防医学的に把握したい料理名とは

食事アセスメントを目的とした場合，料理名は，「アジのフライ」のように，素材名と調理法がわかる表記が望ましい．熟練栄養士は，調理法が判れば，油や調味料の種類を推定でき，エネルギーや食塩の量を見積もることが可能である．調理法は，生，漬ける，茹でる，和える，蒸す，煮物，汁物，焼き，炒める，揚げるを，動詞として認識したい．

4. 料理名を漢字に変換するツールの必要性

地域住民の申告は，「やさいにもの」，「やさいのにもの」といった揺らぎがあり，文字列を正規化し，標準成分表の食品番号に集約する必要がある．

自然言語処理分野では，自由回答式の評判調査などで得た書き言葉に関しては，形態素解析をして，構文解析，意味解析する手順が高精度に確立している．けれども，食事調査で得られた話し言葉は，既存の言語処理ソフトで形態素を解析すると，「握り寿司」の「すし」をサ行変格動詞と認識したり，「肉じゃが煮」の「に」を助詞と誤認識するなど，素材 X に調理操作 Y を施した物という構文解析には到達できない．

4.1 料理名の漢字変換するツールの作業手順

このツールは Microsoft Excel のマクロで作成し，た．変換の優先順は，a. 調味料，調理方法を一般の食品名より先に変換し，b.食品は五十音の順番に漢字変換した．調査データから同定した高頻度出現食品を優先した．c. 上記の変換で失敗した単語の優先順位を上げて変換ミスを修正した．d.食品成分表と関連付けるためには，複合名詞としての認識が必須

である単語を同定した．e. 商品名など平仮名を片仮名に変換する作業は，できるだけ省略した．f.食品と調理法をつなぐ助詞 "と", "の"は stop words とした．g. 食品成分表に載っていない食品は，類似の食品としてコード化した．

5. 結果

現状では，食事調査の出現料理 80,362 回のうち上位 75%をカバーする高頻出料理を正しく変換できた．品詞の認識の妥当性を，辞書 IPA，NAIST，UniDic 現代版，JUMAN の実演比較で検証した [1]．

5.1 表記揺れ

料理名に頻発する表記揺れは「すぱげってぃー，すぱげてぃー，すぱげてぃ」などが見受けられたが既存の辞書 JUMAN でも集約可能であった．酢は，「す」と濁点「ず」で揺らいでいたが，文字数が少ないので，画一的な変換ができなかった．そのため，長い文字数の言葉を，食事調査からリストアップして変換式を作り，「すづけ（酢漬け），あまず（甘酢），さんばいず（三杯酢）」を各々，変換した．

5.2 同音異義語の正規化

さけ（鮭，酒）は，同音異義語と表記揺れの両方が複合してしまい，「しおさけ，しおざけ」が区別できなかった．「かき」は，果物の柿と貝類の牡蠣とを区別できなかったし，検証した辞書によっては，「動詞：書く」と認識されることもあった．

5.3 文字数の長い単語の優先

漢字変換の手順は文字数の多い単語を優先すべきである等のルールを見いだした．具体的には「どん：2 文字」を「丼」に変換する場合は「うどん：3 文字」を優先して変換する．さらに「ぎゅうどん:5 文字」は「牛丼」に変換したいので，「うどん：3 文字」よりも先に変換するルールが必要であった．

5.4 助詞「と」「の」は変換保留，「に」は煮物

「と」「の」はストップワードと考えて，変換処理はしなかった．「に」は煮物として自動変換させるために，「「に」で終わるを条件」にして，住民食事調査から出現例を抽出し，90 種類の煮物を得た．

「炒め物」「煮物」などの「もの」は，出現頻度が高かったので，一括変換を試みたが，「いもの煮物」で変換ミスが起きた．

5.5 複合名詞としての認識

素材食品に対して何らかの加工をした食品，例えば「ドクダミ」と「茶」の 2 つの名詞ではなく，1 つの複合名詞「ドクダミ茶」と認識させる必要があった．そのためには，「茶」や「漬け」を接尾詞と認識させると便利であった．

5.6 カタカナに変換すべき例

カタカナ名詞は，料理専用の漢字変換ツールでなくても，汎用の形態素解析器で認識可能な例が多かった.しかし，平仮名のままだと「やくると：商品名」は，「や：助詞」「くる：動詞」「と：助詞」，「ばじる風味」は．「ば：名詞」「じ：未定義」「る：接尾語」と誤分類された．

6. 考察

汎用の形態素解析器は，最近 2-3 年で大きく進歩し，Wikipedia から自動的に語彙を獲得して，辞書を整備している．自動獲得が可能になったのは，開発当初に人手で整理した基本語彙があったからである [3]．本研究が取り組んでいる「料理のかな漢字変換ツール」は初歩的で人手を駆使した状況であるが，音声で記録された食事内容を，栄養士がいない場面でも，機械的に認識できれば，自動的に日本食品標準成分表と関連付けて栄養価計算が可能と考えている．

表 1. 住民食事調査で出現した「に」で終わる調理法

甘辛煮，甘酢煮，甘煮，アラ煮，イカナゴ釘煮，炒め煮，炒め煮，イトコ煮，田舎煮，炒り煮，旨煮，梅煮，オイスターソース煮，オイスターソース煮，おかか煮，親子煮，オランダ煮，角煮，重ね煮，南瓜煮，カレースープ煮，カレー煮，カレー和風煮，甘露煮，キムチ煮，キンピラ煮，釘煮，葛煮，クリーム煮，ケチャップ煮，ケンチン煮，昆布巻き煮，胡麻味噌煮，五目煮，コンソメスープ煮，コンソメ煮，サット煮，サッパリ煮，山椒煮，時雨煮，渋皮煮，シロップ煮，スープ煮，すき煮，すき焼き風煮，雑煮，ソース煮，ソボロ煮，大根煮，タラコいり煮，筑前煮，チリソース煮，佃煮，照り煮，土佐煮，トマトソース煮，トマト煮，トロトロ煮，蜂蜜煮，ピーマンの炒め煮，ヒジキ煮，ヒジキの炒り煮，ピリ辛煮，ブイヨン煮，フキ寄せ煮，含め煮，鰤のあら煮，ポッポ煮，マーマレード煮，ミゾレ煮，ミルク煮，大和煮，レモン煮，若竹煮，和風煮，海老煮，昆布煮，酒粕煮，醤油煮，蒸し煮，水煮，生姜煮，赤ワイン煮，中華スープ煮，豆乳煮，味噌煮，野菜煮，揚げ煮，卵とじ煮

表 2 「ちゃ」を含むデータの抽出と日本標準食品成分表のコード化（例）

平仮名〝ちゃ〟	漢字表記	食品成分表*の食品番号
茶以外	南瓜	6048
	チャンプル，チャンポン，チャウダー，叉焼，チャイ，巾着，チャンチャン焼き，きゅうりのQちゃん	適宜，調味料をコード化する
茶を同定	日本茶，緑茶，お茶，茶，煎茶	16037
	抹茶	16038
	番茶	16039
	ほうじ茶	16040
	ウーロン茶，マテ茶，ルイボスティ	16042
	紅茶	16044
	麦茶，薬草茶，ドクダミ茶，サクランボ茶，爽健美茶	16055

*文部科学省;日本食品標準成分表2015年版（七訂）

7. 参考文献

(1) Masato Watanabe: IPA，NAIST，UniDic，JUMAN の辞書実演比較, http://www.mwsoft.jp/programming/munou/mecab_dic_perform.html (2016/05/30 閲覧)
(2) 黒橋禎夫，河原大輔: 日本語形態素解析システム JUMAN, http://nlp.ist.i.kyoto-u.ac.jp/ (2016/05/30 閲覧)
(3) 柴田知秀，村脇有吾ら，実テキスト解析をささえる語彙知識の自動獲得，言語処理学会第 18 回発表論文集 p81-84 (2012)

付記：本研究は，科研費新学術領域研究 221S001，基盤 C 24501007, 15K00856 の助成を得た．