Vol.2016-MPS-108 No.9
Vol.2016-BIO-46 No.9
2016/7/4

レシピの材料表における語の出現頻度とその特徴

但馬 康宏[1,a)]　菊井 玄一郎[1]

概要：投稿型のレシピサイトでは多種多様な入力が行われ，材料表の記述も一貫していない．本研究では，レシピにおける材料表の中でも特に分量の部分に注目し，出現する語の分布と数値表現との関係を考察した．その結果，助数詞となる単語は数値表現の直後に，計測の道具を表す単語は数値表現の直前に出現することが確認できた．さらに，位置関係の分布のエントロピーから単語の出現位置から役割を決定できる可能性を示した．

Words distribution and its characteristics on the material list of cocking recipes

Tajima Yasuhiro[1,a)]　Kikui Genichiro[1]

1. はじめに

近年，投稿型のレシピサービスがウェブサービスの中でも人気を集めている．しかし，その記述には投稿者ごとの癖や違いが多く，サイト全体として一貫性のある記述とするには難しいことが多い．

一方，レシピの計算機による理解は，作業手順の理解と解析 [1][2]，オントロジーの構築 [3]，料理アレンジの提案 [4] などにとって必要な技術である．本研究では，投稿型レシピサイトにおけるレシピの中でも，特に材料表の分量の表現に注目して，そこでの単語の出現頻度を調査する．特に数値の表現に前後して出現する単語に注目すると，分量を表す単位や分量を測るための調理器具などが頻出する．これらの出現位置と出現頻度，さらにそれらの分布のエントロピを用いると単位を表す単語と分量のための調理器具との分離が可能となった．

2. 投稿型レシピサイト

投稿型のレシピサイトは現在著名なものだけでも「楽天レシピ」「クックパッド」などがあり，それらのサイトで取り扱っているレシピ数も数百万におよぶ．サイトの種類は違っても，ひとつのレシピが保持する情報およびその表示形式はある程度一般化している．図 1 に代表的なレシピの表示を示す．代表的な構成要素は以下のとおりである．

- タイトル
- レシピが属するカテゴリ
- 完成写真
- 材料表（材料名とその分量）
- 調理手順

これらの記述において，材料表，調理手順はテキスト形式で利用者が作成する．記述のルールは多くない場合がほとんどで，書き手のセンスに依存している点が特徴である．

3. レシピにおける材料表

材料表は，材料名と分量がひとつの行となり，そのレシピで使われる材料や調味料のすべてをリストしている．材料名の記述は，レシピ執筆者がその料理に合わせて作成するため，同一の食材でも複数の名称を持つ食材や，執筆者の好み，料理内容により異なる表記がなされる場合がある．さらに，同一食材の同一名に関しても表記ゆれの問題があり，投稿型レシピサイトにおける材料名からの食材同定は容易ではない．

材料表におけるもうひとつのカラムは分量であり，材料名にある食材の使用分量を表す．この表記についても投稿型レシピの場合は，投稿者の自由記述となるため，多彩な

[1] 岡山県立大学
岡山県総社市窪木１１１
[a)] tajima@cse.oka-pu.ac.jp

Vol.2016-MPS-108 No.9
Vol.2016-BIO-46 No.9
2016/7/4

図 1 レシピ表示の例

表現が存在する．

(1) 助数詞の多様性によるもの:魚の切り身一切れに対して，「一つ」と表現したり「1」のみの場合などがある．また，分量を正確に理解するためには，一切れが何グラムであるかを理解する必要がある．これにはデータをあらかじめ蓄積した辞書やデータベースが必要である．

(2) 投稿者の注記:「５個小玉なら７個」「茶碗１杯か 200g」など複数の選択肢を記述する場合などである．

(3) 表記ゆれ、記述ミス:「グラム」「g」などは異なる表記であり，特に英文字や外来語が入ると記述者に依存する違いが多く存在する．

この中で助数詞の多様性による問題以外は，料理学校が運営するレシピサイトなど一貫した管理がなされているレシピでは起こりにくい問題である．

また，栄養価計算を行う場合などは，日本食品標準成分表 [5] による計算が必要となる場合が多いが，そこでは多くの食材はグラム単位で栄養価が示されており，換算の問題は大きな影響をおよぼす．

4. 材料表における単語の出現頻度

楽天株式会社より提供されている「楽天データ」に対して，レシピの材料表から分量表現の部分を取り出し，以下の前処理を行った．

- 漢数字の置き換えと文字の半角化
- +,-,/,(,) 以外の記号の空白文字化
- 小数，分数の統一

前処理の後，字種の変化点および空白を単語の区切りとした．楽天データの材料表は，材料名と分量のペアを 1 行と数えると，およそ 270 万行である．数値の出現回数ごとに分類すると，表 1 のような分布となる．今回は，数値の出現回数がちょうど 1 回である行の分量に対して単語を抽出した．例えば，「小¡数値¿個」「大さじ¡数値¿強」のような行である．その結果，以下のような分量表現に関する単語辞書を得た．

表 1 コーパスの各行における数値の出現回数

0 回出現	495,752 行
1 回出現	2,085,378 行
2 回出現	109,415 行
3 回出現	3,451 行
4 回出現	245 行
5 回出現	5 行
6 回出現	1 行

- 単語総数：2,228,909
- 単語の異なり数：8,360
- 最多出現単語と出現数：「大さじ」336,217 回
- 出現数上位 100 単語による単語総数のカバー率：96.20%
- 出現回数が 1000 回を上回る単語：86 単語
- 出現回数が 100 回を上回る単語：276 単語

出現数上位 10 単語は，大さじ，g，個，小さじ，本，枚，cc，大，小，大匙である．

次にそれぞれの単語について，その出現位置と数値表現の出現位置との関係を調べ統計をとった．すなわち，「250 cc 程度」ならば「cc」は数値表現の一つ後ろの単語として出現し，「程度」は数値表現の二単語後ろに出現している．また，「小さじ 3 弱」ならば「小さじ」は数値表現の一つ前に出現し，「弱」は数値表現の一つ後ろに出現している．各単語について，数値表現の５つ前から５つ後ろまでの出現数を計測し，その単語の総出現数で割って割合の分布を求めた．表 2 にいくつかの単語の分布を示す．

助数詞にあたる単語は数値表現の１つ後ろに出現しており，計量の道具にあたる単語は数値表現の１つ前に出現していることがわかる．「くらい」「程度」などは数値表現の１つ後と２つ後に出現する割合が半々であり，「大さじ 3 程度」「20 g 程度」など表現が同程度の頻度で出現していることがわかる．また，助数詞としてしか解釈されない「リットル」が数値表現の直後にしか出現しないのに対して，大きいサイズを表すこともある「L」は，数値表現の前後に出現がばらついている．

各単語の出現分布について，エントロピーをとると，以

表 2 単語と数値表現の位置関係の分布

単語	5つ前	4つ前	3つ前	2つ前	1つ前	1つ後	2つ後	3つ後	4つ後	5つ後
大さじ	0	0	0	0	1.000	0	0	0	0	0
小さじ	0	0	0	0	1.000	0	0	0	0	0
茶碗	0	0	0	0.002	0.998	0	0	0	0	0
g	0	0	0	0	0	1.000	0	0	0	0
グラム	0	0	0	0	0	1.000	0	0	0	0
個	0	0	0	0	0	1.000	0	0	0	0
cc	0	0	0	0	0	1.000	0	0	0	0
腹	0	0	0	0	0	1.000	0	0	0	0
滴	0	0	0	0	0	1.000	0	0	0	0
カップ	0	0	0	0	0.135	0.865	0	0	0	0
缶	0	0	0	0	0.004	0.994	0.002	0	0	0
くらい	0	0	0	0	0	0.422	0.578	0	0	0
程度	0	0	0	0	0	0.528	0.471	0.001	0	0
程	0	0	0	0	0	0.540	0.459	0	0	0
)	0	0	0	0.010	0.045	0.030	0.254	0.157	0.456	0.039
(	0.001	0.010	0.045	0.154	0.146	0.128	0.491	0.024	0.001	0
L	0	0.001	0.010	0.085	0.313	0.566	0.004	0.019	0.002	0.002
リットル	0	0	0	0	0	1.000	0	0	0	0
少々	0.003	0.006	0.296	0.281	0.078	0.006	0.183	0.117	0.024	0.006
ひとつまみ	0.027	0	0.324	0.125	0.162	0.027	0.027	0.243	0	0.054

下のような特徴が得られた.

- 全 8360 単語中，エントロピーが 0 である単語：7689 単語
- エントロピーの最大値とその単語：2.488「ひとつまみ」(計測対象は分量表現の中に数値表現が 1 つだけの行であることに注意)

エントロピーが 0 であることは，その単語の数値表現に対する位置関係が固定的であることを示している．したがって，出現単語の 9 割以上が，その出現位置から役割を決定できる可能性がある．

数値表現がない場合に, 分量として頻繁に使われる「少々」「適宜」に関しては，今回の数値表現が 1 つのみの行にも出現し，出現関数はそれぞれ，334 回と 108 回であった．出現分布のエントロピーは「少々」が 2.419,「適宜」が 2.267 であった.

5. おわりに

投稿型レシピデータの材料表の中の分量を記述する部分について，出現する単語を取り出し出現頻度に関する統計をとった．また，数値表現との出現位置の関係を調べ，計測の道具に使われるものは数値表現の直前に，助数詞となる単語は数値表現の直後に出現することが確認できた．このような分布を用いることにより，未知の単語に対してどのような役割の単語であるかを判定する手法の開発が，今後の課題となる.

謝辞 本研究で用いられたデータは，楽天株式会社より提供されています「楽天データ」を利用しています．ここに感謝の意を表します.

参考文献

[1] 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, 情報処理学会自然言語処理研究会, 2013-NL-214(13), (2013).
[2] 山崎 健史, 吉野 幸一郎, 前田 浩邦, 笹田 鉄郎, 橋本 敦史, 舩冨 卓哉, 山肩 洋子, 森 信介, フローグラフからの手順書の生成, 情報処理学会論文誌, vol.57, no.3, pp.849–862, (2016).
[3] 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, 信学技報, vol.113, no.470, MVE2013-68, pp.37–42, (2014).
[4] 工藤 貴徳, 北山 大輔, レシピ間の対応度と相違性に基づく料理アレンジナビゲーション, DEIM Forum 2016, E2-5, (2016).
[5] 日本食品標準成分表 (七訂), 文部科学省, (2015).