レシピサイトからの調理における技術的豆知識抽出手法

11471125　米田　吉希（灘本研究室）

あらまし：クックパッドや楽天レシピ等のユーザ投稿型レシピサイトのレシピには，美味しくなるコツや注意すべき点等の知っていると役に立つ情報が記載されているレシピもあれば，そうでないレシピも存在する．本研究では，知っているとその料理を作るのに役に立つ情報を技術的豆知識と呼び，レシピから機械学習を用いて技術的豆知識を自動で抽出する手法を提案する．

１．はじめに

近年，「クックパッド[1]」や「楽天レシピ[2]」等のユーザ投稿型レシピサイトの利用者が増加している．しかしながら，投稿されるレシピは投稿者自身が記載しているため，レシピに用いられる表現方法は統一されていない．その為，そのレシピの料理を作る上で，ユーザの知らない役に立つ情報が含まれたレシピページもあれば，これらが含まれていないレシピページもある．例えば，親子丼を作る際，「鶏肉を料理酒に浸けてよく揉んで，15 分ほどおくと柔らかくなる」という情報はおいしい親子丼を作る際に役に立つが，これらの情報が載っているレシピページ A もあれば載っていないレシピページ B もある．その為，もしもユーザがこのような役に立つ情報が載っていないレシピページ B を見て親子丼を作る場合，その役に立つ情報を知っていれば容易においしく親子丼が作れるのに，その役に立つ情報を知らないために堅い鶏肉の親子丼になってしまい失敗する場合もある．そこで，レシピページ B を見ているユーザに自動でその役に立つ情報を提示すると便利であると考えた．

しかしながら，レシピサイトのレシピの量が膨大である為，人手でその情報を探すのは困難である．そこで本研究では，ユーザ自身が調理についての知識の拡大や料理の幅を広げる事が可能であり，ユーザ自身の料理技術の向上が期待できる情報を豆知識と呼び，その豆知識を自動で抽出し提示する手法の提案を行う．豆知識には，「下ごしらえ」や「代替」，「切り方」等様々な豆知識がある．そこで本研究では図１に示すように料理の豆知識を分類した．この分類の中から，「より美味しくなる」や「失敗しない」，「時間短縮」のように，知っているとその料理を作るのに役に立つ情報を技術的豆知識と呼ぶ．本研究では，この技術的豆知識を自動でレシピサイトから抽出し提示する手法の提案を行う．具体的には，ルールベースによる技術的豆知識文抽出と SVM による技術的豆知識文の抽出を行い，技術的豆知識を抽出する．

2．提案手法

2．1 ルールベースによる技術的豆知識文抽出

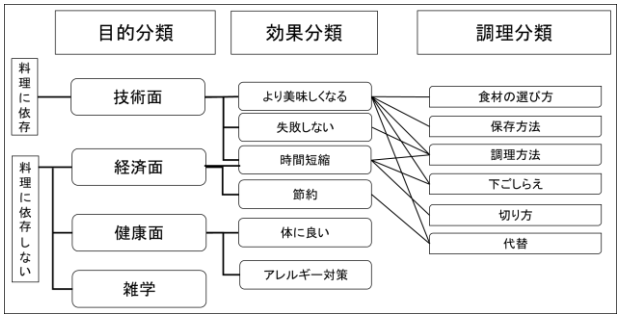

図１ 豆知識の分類

表１ コツ用語

下ごしらえ	下処理	代用	代替	代わり	無い
足りない	本格	仕上	きれい	キレイ	綺麗
コツ	こつ	大事	重要	美味	おいし
オススメ	アップ	風味	※	保存	おすすめ
ポイント	いい	良い	コク	引き立	グッド
気をつけ	冷蔵	冷凍	日持ち	注意	失敗
スピーディ	楽	簡単	カンタン	効率	時短
時間短縮	安く	安い	予防	防止	含む

技術的豆知識には「火をつけたまま、卵液を入れると一気に固まってしまうので注意♪」や「ベーコンをカリカリに焼くと、コクが出て美味しくなりますよ！」といったように「注意」や「コク」のような調理に関する特徴語が含まれている場合が多い．そこで本研究では，ここでいう「注意」や「コク」のような特徴語を抽出し，その特徴語を用いた文を技術的豆知識文として投稿型レシピサイトから抽出する．本研究では，この特徴語をコツ用語と呼び，1,000 件のレシピを分析し，表１に示すコツ用語を決定した．そして，クックパッドのコツ・ポイントの記載欄と楽天レシピのおいしくなるコツの記載欄からコツ用語を１つ以上含む文を，ルールベースによる技術的豆知識文として抽出する．また，1 つの料理のレシピに限らず，他の料理のレシピも同様に豆知識文を抽出できると考える．

2．2 SVM による技術的豆知識文の抽出

ルールベースにより抽出された豆知識文だけでは豆知識でない情報も含まれている．そこで，

本研究では取得したレシピに対してSVMを用いて，その料理に対しての技術的豆知識であるか，そうでないかの二値分類を行う．機械学習にはSVMのライブラリであるLIBSVM[3]を用い，カーネルにはRBFカーネルを用いる．教師データは人手で作成した正例負例データを用いる．それらを用いて抽出された豆知識文を判定し，技術的豆知識を決定する．

3. 評価実験

本研究ではレシピサイトから技術的豆知識を抽出する事を目的とし，ルールベースとSVMを用いた手法を提案した．これら2つの手法の有用性を計るために，各々の評価実験を行った．

3. 1 データセット

本研究で用いるレシピデータには，クックパッドと楽天レシピ各々に投稿された，カルボナーラ，カレーライス，親子丼の3種類のレシピ合計10,500件を用いた．

3. 2 ルールベースによる技術的豆知識文抽出

データセットのコツ・ポイント欄からコツ用語54語のうち1つ以上含まれている文を豆知識文として抽出した．また，人手により技術的豆知識文であるかどうかを判定した．

結果と考察

表２ ルールベースによる豆知識文抽出結果

料理名	適合率	再現率	F値
カルボナーラ	0.292	0.671	0.407
カレーライス	0.203	0.638	0.308
親子丼	0.206	0.410	0.274

表２に適合率と再現率，F値の結果を示す．適合率，再現率は低い結果となった．これはルールベースでは技術的豆知識が取り切れていないこと及び，それに反して豆知識ではない文を豆知識として抽出していることがわかる．悪い例で「コツやポイントがない」といったように否定語が「コツ」や「ポイント」のコツ用語にかかっているものがあったので，コツ用語を否定している文に対しては考慮しなければならないことがわかった．また，技術的豆知識文ではあるが，対象料理に対する豆知識ではない文もあったため，判定が必要であることがわかった．さらに，他のページを参照している文もあったため，今後考慮する必要があることがわかった．このように，ルールベースではコツ用語だけでなく，種々のルールがさらに必要であることがわかった．また，適合率が低いことにより，機械学習を用いる必要があることがわかった．

3. 3 SVMによる技術的豆知識文抽出

教師データには人手で作成した，正例（+1）が対象レシピの技術的豆知識文，負例（-1）が料理やレシピに関係はあるが，技術的豆知識ではない文を用いた．素性は，料理ごとの教師データに形態素解析を行い，得られた名詞・動詞・形容詞全てを対象とし，各々の単語が豆知識文内に含まれているか否かとする．また，SVMの精度を調べるために性能評価を行った．評価には交差検定を用い，分割数は10とする．他に，未知のレシピデータに対する分類も行った．

結果と考察

表３ SVMによる技術的豆知識文抽出結果

料理名	評価手法	適合率	再現率	F値	精度
カルボナーラ	10 交差検定				0.998
	未知データ	0.5	1.00	0.667	0.500
カレーライス	10 交差検定				0.995
	未知データ	0.486	1.00	0.654	0.486
親子丼	10 交差検定				0.741
	未知データ	0.673	0.760	0.714	0.695

表２に適合率，再現率，F値，精度の結果を示す．未知データの分類での適合率が低く，再現率が高い結果となった．これは未知データの正解データと失敗データの出現単語にあまり差が無かったためだと考えられる．本研究の評価実験では素性の単語の重みが全て同じであったため，正解データの重要単語の重みを付け，正解データと失敗データとの差をつける必要があると考えられる．

4. まとめと今後の課題

本研究では，ルールベースと機械学習による技術的豆知識抽出手法の提案を行い，その有用性を測った．また，料理ごとに技術的豆知識の種類や量が異なることから，精度が安定しないことから，料理の種類ごとに分類する必要があると考えられる．

今後の課題として，SVMの精度向上，コツ用語の再考察，別の手法による技術的豆知識抽出などが挙げられる．

参考文献

[1] クックパッド，https://cookpad.com/
[2] 楽天レシピ，https://recipe.rakuten.co.jp/
[3] Chih-Chung Chang and Chih-Jen Lin, ¥”LIBSVM {A Li-brary for Support Vector Machines”″, 2014, https://www.csie.ntu.edu.tw/cjlin/libsvm/