Vol.2012-NL-208 No.7
2012/9/3

料理レシピに含まれる類似文字列の検索

安川　美智子[†1]

料理レシピには，料理名や素材名の表記揺れが多数含まれるため，料理レシピ検索の際に，検索クエリと料理レシピ中の文字列の間で不一致が生じるという問題がある．そこで，料理名や素材名を表す発音と意味が同じで，表記が異なる複数の文字列を料理レシピから抽出して特定し，検索漏れを防ぐことが必要となる．そこで，本稿では，日本語の発音照合と文字列編集距離を用いた類似文字列検索の手法を提案し，料理レシピに含まれる表記揺れの抽出に適用する．また，実際の料理レシピデータを用いた評価実験を行い，提案法の有効性を考察する．

Japanese Fuzzy String Matching in Cooking Recipes

Michiko Yasukawa[†1]

In this paper, we propose Japanese fuzzy string matching in cooking recipes. Cooking recipes contain spelling variants for recipe titles and ingredient names that cause mismatches between search queries and relevant recipe texts. In order to find these spelling variants, we use phonetic matching in Japanese and edit distance. We have evaluated the proposed methods using actual cooking recipes on the Internet. We report our findings based on the evaluation results.

1. はじめに

本稿では，料理レシピを対象とした情報検索システムにおいて，索引語 (indexing term) となる文字列の表記揺れを抽出する手法について検討する．

近年，料理に関する情報アクセス技術の重要性が高まっている．料理行動を対象としたメディア処理技術に関する先行研究として，献立を決める[1]，料理を作る[2]，料理を食べる[3] といった観点からのユーザ支援が提案されている．料理レシピの文書解析に取り組んだ先駆的な研究として，浜田ら[4] は，固有の辞書の構築を行い，料理テキスト教材に含まれるテキストデータ（非構造化データ）の構造解析を行う際に構築した辞書を用いることで，オペレーショングラフと呼ばれる形式（構造化データ）への変換を実現している．また，林ら[5] は，日本語で記述された料理レシピ文の時間構造の解析の手法を提案している．柴田ら[6] の研究では，料理番組映像を対象として，その発話内容のクローズドキャプションの構文・格・省略・談話構造解析を行い，作業構造を自動抽出する手法を提案している．高野ら[7] の研究では，料理レシピの構造解析に加えて，明示的に書かれていない情報の補完を行うことで，調理シナリオと呼ばれる，より構造化された形式へのデータ変換の手法を提案している．また吉川ら[8] は，料理レシピをデータフロープログラミング言語と呼ばれる形式に変換するためのデータ仕様と方法論を提案している．苅米ら[9] の研究では，調理動作における解析誤りを解決し，料理レシピからフローチャートを自動生成する際の精度向上を行っている．本研究では，料理レシピに対して品詞情報や係り受け関係などの言語分析を行うのではなく，料理レシピ検索を目的として，索引語となる部分文字列の抽出を行う．

また，本研究と同様に，料理レシピを検索するという観点からの研究として，塩澤ら[10] は，食材の優先度を考慮してインタラクティブに検索候補の絞り込みを行う手法を提案している．志土地ら[11] の研究では，料理レシピを検索する際に，ある食材の代替となる食材を考慮することを提案している．これらのような応用的な検索課題に取り組むことは将来課題とし，本研究では，料理名や材料名といった料理レシピ検索で基礎となる索引語の抽出を精度よく，かつ，漏れなく行うことに取り組む．

2. 関連研究

本研究に関連する発音照合と文字列編集距離についての従来研究について以下に述べる．

2.1 発音照合

類似文字列検索の手法のひとつで，文字列の綴りではなく，文字列が発音された音声の類似性に基づく類似文字列検索の手法は発音照合 (phonetic matching) と呼ばれている．英語の発音照合の手法として，よく知られているものに Soundex[12] と Metaphone[13] がある．これらの手法の目的は，人の名前のさまざまな変種すべてを 1 つのコードに変換することである[14]．Soundex と Metaphone では，先頭の文字を保持し，途中の母音を捨て，子音に対しては英語の発音を考慮した変換規則に従って変換することで同じ発音を持つ文字列が同じ文字列になるように符号化を行う．Soundex では先頭以外は 6 つの数字 (123456) に符号化されるのに対して，Metaphone では 16 文字の子音を表す文字 (0BFHJKLMNPRSTWXY) で符号

†1 群馬大学
Gunma University

表 1 Editex の文字列グループ
Table 1 Editex Letter Groups.

0	1	2	3	4	5	6	7	8	9
aeiouy	bp	ckq	dt	lr	mn	gj	fpv	sxz	csz

化される．たとえば，同じ発音で綴りが異なる英語の名字 SMITH と SMYTH は，Soundex の符号化により，同じ符号 S530 に変換される．また，Metaphone では，SMITH と SMYTH は，SM0 に変換される．発音照合は，元の文字列ではなく，符号化した文字列での完全な一致による文字列照合を行うことで，発音が同じで綴りが異なる文字列間の照合を実現している．

2.2 文字列編集距離

テキストの部分文字列とパターンとの完全な一致ではなく，近似的な一致を行う近似文字列照合 (approximate string matching) において，文字列間の距離を与える指標となるのが，編集距離 (edit distance)[15] である．文字列 x と y の編集距離 $d(x, y)$ は，x を y へ変換するのに必要な編集操作の最小のコストである．編集操作として，文字の挿入・削除・置換を扱い，挿入・削除・置換には任意のコストを割り当てることができる．たとえば，挿入・削除・置換のコストをすべて 2 とすることもできるし，挿入・削除のコストを 2，置換のコストを 1 としてもよい．

Zobel ら[16] は，文字列編集距離に発音照合の符号化のアイデアを取り入れた Editex と呼ばれる文字列編集距離の拡張版を提案している．Editex は，発音が類似する文字を表 1 のようにグループ化し，同じグループ内の文字での置換に対しては，編集操作のコストを小さくする．たとえば，通常の編集距離では，挿入・削除・置換のコストを 2 とした場合，編集距離 d(sip,zip) と d(sip,lip) はともに 2 となる．Editex では同じグループの文字の置換はコストを小さくすることができるため，同じグループの文字の置換のコストを 1，異なるグループの文字の置換のコストを 2 とすることにより，編集距離 d(sip,zip) は 1，d(sip,lip) は 2 となり，発音が類似する文字列の間の編集距離が小さくなる．

3. 提 案 法

発音照合と文字列編集距離を用いた提案法について以下に述べる．

3.1 日本語の発音照合を用いた類似文字列の検索

著者らの先行研究[17] では英語の発音照合の手法と日本語の五十音図 (表 2) を用いて日本語版の発音照合を行う手法を提案している．英語の発音照合と同様に日本語の発音照合でも，子音の発音の類似性に注目した文字のグループ化を行う．具体駅には，表 2 の列方向で

表 2 日本語の五十音
Table 2 The Japanese Syllabary (Fifty Sounds).

	Hiragana Symbol					Katakana Symbol					
	A	I	U	E	O	A	I	U	E	O	
φ	あ	い	う	え	お	ア	イ	ウ	エ	オ	1
	E38182	E38184	E38186	E38188	E3818A	E382A2	E382A4	E382A6	E382A8	E382AA	
	a	i	u	e	o	a	i	u	e	o	
K	か	き	く	け	こ	カ	キ	ク	ケ	コ	2
	E3818B	E3818D	E3818F	E38191	E38193	E382AB	E382A4	E382A6	E382A8	E382AA	
	ka	ki	ku	ke	ko	ka	ki	ku	ke	ko	
S	さ	し	す	せ	そ	サ	シ	ス	セ	ソ	3
	E38195	E38197	E38199	E3819B	E3819D	E382B5	E382B7	E382B9	E382BB	E382BD	
	sa	si	su	se	so	sa	si	su	se	so	
T	た	ち	つ	て	と	タ	チ	ツ	テ	ト	4
	E3819F	E381A1	E381A4	E381A6	E381A8	E382BF	E38381	E38384	E38386	E38388	
	ta	ti	tu	te	to	ta	ti	tu	te	to	
N	な	に	ぬ	ね	の	ナ	ニ	ヌ	ネ	ノ	5
	E381AA	E381AB	E381AC	E381AD	E381AE	E3838A	E3838B	E3838C	E38838D	E3838E	
	na	ni	nu	ne	no	na	ni	nu	ne	no	
H	は	ひ	ふ	へ	ほ	ハ	ヒ	フ	ヘ	ホ	6
	E381AF	E381B2	E381B5	E381B8	E381BB	E3838F	E388392	E38395	E38398	E3839B	
	ha	hi	hu	he	ho	ha	hi	hu	he	ho	
M	ま	み	む	め	も	マ	ミ	ム	メ	モ	7
	E381BE	E381BF	E38280	E38281	E38282	E3839E	E3839F	E383A0	E383A1	E383A2	
	ma	mi	mu	me	mo	ma	mi	mu	me	mo	
Y	や		ゆ		よ	ヤ		ユ		ヨ	8
	E38284		E38286		E38288	E383A4		E383A6		E383A8	
	ya		yu		yo	ya		yu		yo	
R	ら	り	る	れ	ろ	ラ	リ	ル	レ	ロ	9
	E38289	E3828A	E3828B	E3828C	E3828D	E383A9	E383AA	E383AB	E383AC	E383AD	
	ra	ri	ru	re	ro	ra	ri	ru	re	ro	
W	わ	ゐ		ゑ	を	ワ	ヰ		ヱ	ヲ	10
	E3828F	E38290		E38291	E38292	E383AF	E383B0		E383B1	E383B2	
	wa	wi		we	wo	wa	wi		we	wo	
	1	2	3	4	5	1	2	3	4	5	

はなく，行方向で文字をグループ化し，近似文字列照合のための符号化を行う．発音照合により得られる文字列検索の結果は，文字をどのようにグループ化するかによって大きく変わる．著者らの先行研究では，表 3，表 4，表 5，表 6 に示す文字のグループ化の規則を用いた発音照合の 4 つの異なる手法を提案している．本稿ではこれら 4 つの手法 (jppm1，jppm2，jppm3，jppm4) を用いて料理レシピに含まれる料理名と材料名の表記漏れの抽出を行う．

発音照合を用いた表記揺れの手順を次に説明する．まず，検索対象となる全ての文字列を表 3，表 4，表 5，表 6 に示す発音照合の符号化の規則に従って符号化する．符号化の際，文字列の先頭には特徴が表れやすいことから，英語版の発音照合において，先頭文字は常に保持されるため，日本語版の発音照合でも先頭文字はそのまま保持することとする．符号化をした後，同じ符号を持つ文字列で集合を作り，集合の要素が 2 つ以上である集合をすべて出力する．この集合に含まれる各要素のことを「表記揺れ文字列」，集合のことを「表記揺れ文字列集合」と呼ぶこととする．符号化前のもとの表記揺れ文字列のうち，もっとも多くの料理レシピに含まれるもの，つまり，文書頻度 (DF 値) が最大の文字列を,「代表文字列」と呼ぶこととする．

発音照合の提案法 jppm1 は，文字列の発音の特徴をできるだけ残して，厳密に文字列の照合を行う手法である．これに対して，提案法 jppm2 は，文字列の発音の特徴をできるだけ削除して，曖昧な文字列の照合を行う手法である．提案法 jppm3 は，提案法 jppm1 の文字列のグループ化に加えて，濁音・清音など他の文字についてもグループ化を行い，提案法 jppm1 よりも曖昧性の高い文字列照合を行う．一方，提案法 jppm4 は，提案法 jppm2 の文字のグループ化の一部を抑制して，提案法 jppm2 よりも曖昧性の低い文字列照合を行う．

提案法 jppm1〜jppm4 の中では，提案法 jppm2 がもっとも曖昧性の高い文字列照合を行う．料理名「キウイジャム」と「チンジャオロース」の発音照合の具体例を表 7 に示す．表において太字の文字は，それぞれの表記揺れ文字列集合の代表文字列である．

発音照合では，文字が追加されているもの，文字が削除されているもの，文字の順序が入れ替わっているものは検索結果に含まれない．たとえば,「キウイノジャム」のように間に助詞「の」が入ったものは，検索漏れがおきにくい手法である提案法 jppm2 でも検索漏れとなってしまうという問題がある．この問題に対処するためには，発音照合を文字列編集距離と組み合わせることが有効であると考えられる．発音照合を用いた文字列編集距離の手法について次に説明する．

3.2 日本語の発音照合と文字列編集距離を用いた類似文字列の検索

文字列編集距離では，文字の削除や挿入がある文字列に対しても文字列間の距離を計算して，近似文字列照合を行えることから，発音照合を文字列編集距離と組み合わせることで，表記揺れ抽出の精度を高めつつ，抽出漏れを削減できると考えられる．

英語の近似文字列照合を行う Editex は，表 1 のように，英語のアルファベットに対して 0〜9 のグループ化を行っている．そこで，日本語版の Editex（以下，jpeditex と呼ぶ）でも同様に，日本語のカタカナの文字に対して表 8 に示すような「あ，か，さ，た，な，は，ま，や，ら，わ，ん」の 11 のグループを定義し，編集距離のコストの計算において，同じグループの文字の置換には低い編集コストを与えることとする．具体的には，発音照合を考慮しない編集距離では編集操作（削除，挿入，置換）のコストをすべて 2 とするのに対して，発音照合を考慮した編集距離では，削除と挿入のコストを 2 とし，異なるグループの文字の置換のコストを 2，同じグループの文字の置換のコストを 1 とする．発音照合を考慮しない日本語版の編集距離（ベースライン法）を jpedit，発音照合を考慮した日本語版の編集距離（提案法）を jpeditex と呼ぶこととする．ベースライン法 jpedit と提案法 jpeditex による表記揺れの抽出の具体例を表 9 に示す．ベースライン法 jpedit では類似する音声を持つ文字のグループ化を行わないため,「キウイジャム」に対して「ウメジャム」という誤った表記揺れの抽出が行われている．

表 3 日本語の発音照合 (jppm1) のための符号化表
Table 3 Encoding Table for Japanese Phonetic Matching (jppm1).

Fifty Sounds [in]	Code [out]	Voiced Sounds [in]	Code [out]	Additional Symbols [in]	Code [out]
アイウエオ (ϕ) → ヰエヲ (obs., ϕ)	E38182 あ			ァィゥェォ (lower-case, ϕ) → ー (macron, ϕ)	E38182 あ
カキクケコ (K) →	E3818B か	ガギグゲゴ (G) →	E3818C が	ヵヶ (lower-case, K) →	E3818B か
サシスセソ (S) →	E38195 さ	ザジズゼゾ (Z) → ヂヅ (obs., Z)	E38196 ざ		
タチツテト (T) →	E3819F た	ダデド (D) →	E381A0 だ	ッ (lower-case, T) →	E381A3 っ
ナニヌネノ (N) →	E381AA な			ン (syllabic nasal, N) →	E38293 ん
ハヒフヘホ (H) →	E381AF は	バビブベボ (B) → ヴ (V)	E381B0 ば		
		パピプペポ (P) →	E381B1 ぱ		
マミムメモ (M) →	E381BE ま				
ヤユヨ (Y) →	E38284 や			ャュョ (lower-case, Y) →	E38283 ゃ
ラリルレロ (R) →	E38289 ら				
ワ (W) →	E3828F わ			ヮ (lower-case, W) →	E3828F わ

表 4 日本語の発音照合 (jppm2) のための符号化表
Table 4 Encoding Table for Japanese Phonetic Matching (jppm2).

Fifty Sounds [in]	Code [out]	Voiced Sounds [in]	Code [out]	Additional Symbols [in]	Code [out]
アイウエオ (ϕ) → ヰヱヲ (obs., ϕ)	削除			ァィゥェォ (lower-case, ϕ) → ー (macron, ϕ)	削除
カキクケコ (K) →	E3818B か	ガギグゲゴ (G) →	E3818C が	ヵヶ (lower-case, K) →	削除
サシスセソ (S) →	E38195 さ	ザジズゼゾ (Z) → ヂヅ (obs., Z)	E38196 ざ		
タチツテト (T) →	E3819F た	ダデド (D) →	E381A0 だ	ッ (lower-case, T) →	削除
ナニヌネノ (N) →	E381AA な			ン (syllabic nasal, N) →	削除
ハヒフヘホ (H) →	E381AF は	バビブベボ (B) → ヴ (V)	E381B0 ば		
		パピプペポ (P) →	E381B1 ぱ		
マミムメモ (M) →	E381BE ま				
ヤユヨ (Y) →	E38284 や			ャュョ (lower-case, Y) →	削除
ラリルレロ (R) →	E38289 ら				
ワ (W) →	E3828F わ			ヮ (lower-case, W) →	削除

表 5 日本語の発音照合 (jppm3) のための符号化表
Table 5 Encoding Table for Japanese Phonetic Matching (jppm3).

Fifty Sounds [in]	Code [out]	Voiced Sounds [in]	Code [out]	Additional Symbols [in]	Code [out]
アイウエオ (ϕ) → ヰヱヲ (obs., ϕ)	E38182 あ			ァィゥェォ (lower-case, ϕ) → ー (macron, ϕ)	E38182 あ
カキクケコ (K) →	E3818B か	ガギグゲゴ (G) →	E3818B か	ヵヶ (lower-case, K) →	E3818B か
サシスセソ (S) →	E38195 さ	ザジズゼゾ (Z) → ヂヅ (obs., Z)	E38195 さ		
タチツテト (T) →	E3819F た	ダデド (D) →	E3819F た	ッ (lower-case, T) →	E3819F た
ナニヌネノ (N) →	E381AA な			ン (syllabic nasal, N) →	E381AA な
ハヒフヘホ (H) →	E381AF は	バビブベボ (B) → ヴ (V)	E381AF は		
		パピプペポ (P) →	E381AF は		
マミムメモ (M) →	E381BE ま				
ヤユヨ (Y) →	E38284 や			ャュョ (lower-case, Y) →	E38284 や
ラリルレロ (R) →	E38289 ら				
ワ (W) →	E3828F わ			ヮ (lower-case, W) →	E3828F わ

表 6 日本語の発音照合 (jppm4) のための符号化表
Table 6 Encoding Table for Japanese Phonetic Matching (jppm4).

Fifty Sounds [in]	Code [out]	Voiced Sounds [in]	Code [out]	Additional Symbols [in]	Code [out]
アイウエオ (ϕ) → ヰヱヲ (obs., ϕ)	E38182 あ			ァィゥェォ (lower-case, ϕ) → ー (macron, ϕ)	削除
カキクケコ (K) →	E3818B か	ガギグゲゴ (G) →	E3818C が	ヵヶ (lower-case, K) →	E3818B か
サシスセソ (S) →	E38195 さ	ザジズゼゾ (Z) → ヂヅ (obs., Z)	E38196 ざ		
タチツテト (T) →	E3819F た	ダデド (D) →	E381A0 だ	ッ (lower-case, T) →	削除
ナニヌネノ (N) →	E381AA な			ン (syllabic nasal, N) →	E38293 ん
ハヒフヘホ (H) →	E381AF は	バビブベボ (B) → ヴ (V)	E381B0 ば		
		パピプペポ (P) →	E381B1 ぱ		
マミムメモ (M) →	E381BE ま				
ヤユヨ (Y) →	E38284 や			ャュョ (lower-case, Y) →	削除
ラリルレロ (R) →	E38289 ら				
ワ (W) →	E3828F わ			ヮ (lower-case, W) →	E3828F わ

表 7 日本語の発音照合を用いて抽出した表記揺れ文字列集合の例
Table 7 Example of Spelling Variant Sets using Phonetic Matching.

表記揺れ文字列	DF 値	提案法 jppm1	提案法 jppm2	提案法 jppm3	提案法 jppm4
キウイジャム	12	きああざゃま	きざま	きああさやま	きああざま
キーウィージャム	1	きあああああざゃま	きざま	きあああああさやま	きあざま
キイウィジャム	1	きあああざゃま	きざま	きあああさやま	きあざま
キウィジャム	1	きあああざゃま	きざま	きあああさやま	きあざま
チンジャオロース	33	ちんざゃあらあさ	ちざらさ	ちなさやあらあさ	ちんざあらさ
チンジャオロースー	16	ちんざゃあらあさあ	ちざらさ	ちなさやあらあさあ	ちんざあらさ
チンジャオロウスウ	2	ちんざゃあらあさあ	ちざらさ	ちなさやあらあさあ	ちんざあらあさあ
チンジャオロウスー	1	ちんざゃあらあさあ	ちざらさ	ちなさやあらあさあ	ちんざあらあさ

4. 評価実験

実際の料理レシピデータを用いた評価実験を行い，提案法の有効性を考察する．

4.1 実験データ

提案法の評価実験を行うために 2 種類の実験用のデータセット (Dataset-A と Dataset-B) を作成した．Dataset-A は,「キューピー 3 分クッキング*1」の Web サイトからダウンロード

⋆1 http://www.ntv.co.jp/3min/

表 8 日本語版 Editex のための符号化表
Table 8 Encoding Table for Japanese Editex.

Fifty Sounds [in]	Code [out]	Voiced Sounds [in]	Code [out]	Additional Symbols [in]	Code [out]
アイウエオ (ϕ) → ヰヱヲ (obs., ϕ)	E38182 あ			ァィゥェォ (lower-case, ϕ) → ー (macron, ϕ)	E38182 あ
カキクケコ (K) →	E3818B か	ガギグゲゴ (G) →	E3818B か	ヵヶ (lower-case, K) →	E3818B か
サシスセソ (S) →	E38195 さ	ザジズゼゾ (Z) → ヂヅ (obs., Z)	E38195 さ		
タチツテト (T) →	E3819F た	ダヂド (D) →	E3819F た	ッ (lower-case, T) →	E3819F た
ナニヌネノ (N) →	E381AA な			ン (syllabic nasal, N) →	E38293 ん
ハヒフヘホ (H) →	E381AF は	バビブベボ (B) → ヴ (V) パピプペポ (P)	E381AF は		
マミムメモ (M) →	E381BE ま				
ヤユヨ (Y) →	E38284 や			ャュョ (lower-case, Y) →	E38284 や
ラリルレロ (R) →	E38289 ら				
ワ (W) →	E3828F わ			ヮ (lower-case, W) →	E3828F わ

表 9 編集距離と拡張版の編集距離 jpeditex を用いて抽出した表記揺れ文字列集合の例
Table 9 Example of Spelling Variant Sets using jpedit and jpeditex.

順位	ベースライン法 (jpedit)	距離	順位	提案法 (jpeditex)	距離
1	キウイジャム	0	1	キウイジャム	0
2	キウィジャム	2	2	キウィジャム	1
3	キウイノジャム	2	3	キウイノジャム	2
4	ウメジャム	4	4	キイウィジャム	3
5	キイウィジャム	4	5	キウィノジャム	3

した 1990 年 1 月 20 日から 2012 年 7 月 11 日までの 5000 件のレシピデータから作成した．レシピデータに含まれる HTML タグを手掛かりとしたパターンマッチにより，料理名と材料名の抽出を行った．Dataset-A に含まれる料理名と材料名のうち文書頻度 (DF 値) が高い上位 10 件を付録の表 13 に示す．また，料理名の文書頻度 (DF 値) を縦軸に，料理名を横軸にして，頻度が高いものを左から順に並べると，付録の図 1 のようになる．

Dataset-B は,「COOKPAD$^{\star 1}$」の Web サイトからダウンロードした公開日が 1998 年 4 月 21 日から 2010 年 7 月 17 日までの 80 万件のレシピデータから作成した．Dataset-A と同様に，レシピデータに含まれる HTML タグを手掛かりとしたパターンマッチにより，料理名と材料名の抽出を行った．Dataset-B に含まれる料理名と材料名のうち，文書頻度 (DF 値) が高い上位 10 件を付録の表 14 に示す．また，料理名の文書頻度 (DF 値) を縦軸に，料理名を横軸にして，頻度が高いもの 1 万件を左から順に並べると，付録の図 2 のようになる．

Dataset-A と Dataset-B の両方に対して，評価実験のための前処理として，抽出した料理名と材料名に含まれる全てのひらがなをカタカナに変換する処理を行い，変換後の文字列がカタカナのみから構成される文字列を表記揺れ検出の評価実験に用いることとした．

4.2 発音照合を用いた表記揺れの抽出の評価

3.1 節で説明した日本語の発音照合の手法 (jppm1〜jppm4) を，料理名と材料名の表記揺れの抽出に適用する実験を行った．提案法ごとの得られた文字列集合の数を表 10 に示す．

Dataset-A は小規模なデータであり，料理研究家などの専門家が記述したレシピデータであるため，レシピデータに含まれる表記はおおむね統一されており，表記揺れは料理名，材料名ともに，比較的少数である．提案法 jppm1 と提案法 jppm3 では，符号化の際にもとの文字列をあまり削除しないため，類似文字列検索の曖昧性が低く，抽出される類似文字列は符号化前の文字列との類似度がかなり高いものに限定される．具体的には，提案法 jppm1 と提案法 jppm3 により抽出された料理名の表記揺れ文字列集合はそれぞれ 1 つで，結果はともに,{ラタトゥイユ，ラタトゥーユ} のみであった．提案法 jppm2 と提案法 jppm4 は長音や拗音などに対する曖昧文字列検索を行う手法であり，具体的には,{ビフカツ，ビーフカツ} や {ムースオマロン，ムースオーマロン} などが表記揺れとして正しく抽出された．しかし，提案法 jppm2 は符号化の際に文字を過剰に削除する傾向があるため文字列検索の曖昧性が高まり，結果として,「サバラン」と「サブレ」,「サムパ$^{\star 2}$」と「サーモンパイ」,「ポテトグラタン」と「ポテトガレット」の 3 つの文字列集合が誤って表記揺れとして抽出された．

材料名の表記揺れの抽出では，提案法 jppm1 により {キーウィフルーツ，キーウイフルーツ} ,{メープルシロップ，メイプルシロップ} などが抽出された．また，提案法 jppm3 により {トッポギ，トッポキ} ,{バゲット，バケット} ,{ムキクルミ，ムキグルミ} などが抽出された．

提案法 jppm2 と提案法 jppm4 により {オレンジママレード，オレンジマーマレード} ,{カルヴァドス，カルバドス} ,{キルシュ，キルッシュ} などが抽出されたが，提案法 jppm2 は提案法 jppm4 が削除しない文字も削除して符号化を行うため，文字列検索の曖昧性が高ま

⋆1 http://cookpad.com/

⋆2 韓国料理の名称

り，結果として,「キンカン*1」と「キンキ*2」,「キュウリ」と「キャラウェイ」,「コショウ」と「コシアン」,「タチウオ」と「ターツァイ」など，他の手法では抽出されなかった誤った表記揺れが抽出された．提案法 jppm1〜jppm4 のすべてで抽出された誤った表記揺れは,「オカラ」と「オクラ」であった．発音が類似している短い文字列が誤って表記揺れとして抽出される場合にどう対処するかを今後，検討していく必要があるが，典型的な誤りは，あらかじめ辞書に文字列を定義しておき，類似文字列検索の際に辞書を参照するなどの対処法が考えられる．

Dataset-B は大規模なデータであり，料理を趣味とする一般ユーザなどが自由に記述したレシピデータであるため，外国語の音訳 (transliteration) の多様性や，文字入力の誤りなど，多数の表記揺れが含まれている．得られた結果を確認すると，提案法 jppm1〜jppm4 で，多くの表記揺れが正確に抽出されている一方で，Dataset-A の場合と同様に，発音が類似する短い文字列が誤って表記揺れとして抽出されていることが分かった．たとえば，料理名では「クッキー」と「ケーキ」,「ウドン」と「オデン」，材料名では「キナコ」と「キノコ」,「ハム」と「ハモ」などである．また，Dataset-A に含まれなかった「略語の表記揺れ」における誤りも Dataset-B の結果からは観察された．たとえば，卵（たまご）の略語としての「タマ」(例 「ツナたまサラダ」) とトマトの略語としての「トマ」が（例「ツナトマサラダ」）を含む料理名が，表記揺れとして誤って抽出された．

表記揺れの抽出は，正確であることと同時に，抽出の漏れがないこと，つまり，どれだけ多くの異なる表記揺れを抽出できているかという点からも評価する必要がある．そこで，Dataset-B から各手法で抽出された文字列集合のうち,「表記揺れ文字列の異なり数」が大きい表記揺れ文字列集合の上位 3 件を次に詳しく見ていくこととする．料理名に提案法 jppm1 を適用した場合の上位 3 件は, {ラタトゥイユ，ラタトゥーユ，ラタトゥウユ，ラタトゥィユ，ラタトウィユ}（表記揺れの異なり数 5）, {ラタトゥユ，ラタトイユ，ラタツィユ，ラタトィユ，ラタトウユ}（表記揺れの異なり数 5）, {チキンパルメジャーナ，チキンパルメジャーノ，チキンパルミジャーノ}（表記揺れの異なり数 4）であったが，最初の二つの集合は，ひとつの同じ集合にまとまっているべきである．提案法 jppm2 の表記揺れ異なり数最大の集合は上位 3 件が {ラタトゥイユ，ラタトゥユ，ラタトゥーユ，ラタトュイユ，ラタトイユ，ラタトューユ，ラタトゥウユ，ラタトウユ，ラタトィユ，ラタツィユ，ラタトウィユ，ラタトゥィユ}（表記揺れの異なり数 12）, {ゴーヤチャンプル，ゴーヤチャンプルー，ゴーヤーチャンプルー，ゴーヤーチャンプル，ゴーヤチャンプルゥー，ゴーヤチャンプルゥ，ゴーヤーチャンプルゥ，ゴーヤタンプルー}（表記揺れの異なり数 9）, {ソウメンチャンプルー，ソーメンチャンプルー，ソーミンチャンプルー，ソーメンチャンプル，ソーミンチャンプル，ソウメンチャンプルゥ}（表記揺れの異なり数 7）となっており，提案法 jppm1 では別の集合になっていた「ラタトゥイユ」の表記揺れが一つの集合にうまくまとめられた．提案法 jppm3 は,「濁点・半濁点」の処理以外は，提案法 jppm1 と共通であるため,「ラタトゥイユ」の表記揺れは提案法 jppm1 と同様の結果となり，3 つめの集合として, {ネギダレ，ネギタレ，ネギドリ，ネギトリ} が抽出されたが，この表記揺れ文字列集合では,「たれ」と「鶏（とり）」という異なる意味の文字列が表記揺れとして誤って抽出されている．提案法 jppm4 では, {ゴーヤチャンプル，ゴーヤチャンプルー，ゴーヤーチャンプルー，ゴーヤーチャンプル，ゴーヤチャンプルゥー，ゴーヤタンプルー，ゴーヤチャンプルゥ，ゴーヤーチャンプルゥ}（表記揺れの異なり数 8）, {タラコスパゲッティ，タラコスパゲティ，タラコスパゲティー，タラコスパゲッティー，タラコスパッゲティ，ターラコースパゲッチィ}（表記揺れの異なり数 6）, {ラタトゥイユ，ラタトイユ，ラタトュイユ，ラタトゥウユ，ラタトウユ，ラタトウィユ}（表記揺れの異なり数 6）が上位 3 件の集合となった．提案法 jppm4 は提案法 jppm2 と共通の特徴があるため,「ラタトゥイユ」と「ゴーヤチャンプル」の表記揺れ集合の要素に，提案法 jppm2 の結果と共通のものがあったが，提案法 jppm2 で抽出できている表記揺れの一部が提案法 jppm4 では検索漏れとなっていた．材料名からの表記揺れの抽出でも同様の傾向が見られ，提案法 jppm1 と提案法 jppm3 の結果は一部共通であり，提案法 jppm3 では，濁点と半濁点のある文字を同一視する曖昧文字列検索が行われた結果,「オリーブオイル」の表記揺れ（入力誤り）である「オリープオイル」が正しく抽出された．提案法 jppm2 と提案法 jppm4 は共通の特徴があるため,「オリーブオイル」の表記揺れ集合の要素には共通の文字列が含まれるが，提案法 jppm2 だけで抽出できている文字列があり，その中には誤りが含まれていた．具体的には以下の文字列が提案法 jppm2 と提案法 jppm4 で抽出された．提案法 jppm2 のみで抽出された文字列のうち，下線を付したものは「オリーブオイル」の表記揺れではない，誤って抽出された文字列である．

- 「オリーブオイル」の表記揺れのうち jppm2 と jppm4 で共通のもの：オリブオイル，オリヴオイル，オリービオイル，オリーブイオル，オリーブオイリ，オリーブオイル，オリーブーオイル，オリーヴオイル，オルーブオイル，オレーブオイル，オローブオイ

★1 金柑
★2 魚の名前

表 10 日本語の発音照合を用いて抽出した表記揺れ文字列集合の数
Table 10 Number of Spelling Variant Sets using Phonetic Matching.

	提案法 jppm1	提案法 jppm2	提案法 jppm3	提案法 jppm4
Dataset-A (料理名)	1	6	1	2
Dataset-A (材料名)	7	2	11	16
Dataset-B (料理名)	335	1695	580	952
Dataset-B (材料名)	781	1845	1173	1270

ル，オーリブオイル，オーリーブオイル

- 「オリーブオイル」の表記揺れのうち jppm2 のみで抽出されたもの：オイリーブオイル，オリーブイオイル，オリーブイル，オリーブオィル，オリーブオイルウ，オリーブオオイル，オリーブオル，オールブラウン，オールブラン

提案法 jppm1〜jppm4 に共通する問題点としては，短い文字列は，表記揺れではない文字列をを大量に集めてしまうということがある．提案法 2 は，長い文字列の中に，わずかな異なりが複数にあるような表記揺れを多く集めることができるが，表記揺れではない文字列（たとえば,「オリーブオイル」に対して「オールブラン」など）を検索結果に含むことがあるため，検索結果の精度を高めるためには，検索された表記揺れ文字列を後処理でフィルタリングする必要がある．

4.3 編集距離を用いた表記揺れの抽出の評価

3.2 節で説明した文字列編集距離の手法を用いて，Dataset-B の料理名と材料名の表記揺れ抽出の実験を行った．具体的には，上述の 4.2 節の実験で，提案法 jppm2 の表記揺れ異なり数が最大となった上位 3 件の表記揺れ文字列集合の代表文字列[⋆1]を検索クエリとして使用することとした．具体的には，料理名の類似文字列検索では「ラタトゥイユ」「ゴーヤチャンプル」「ソウメンチャンプルー」を検索クエリとして使用した．また，素材名の類似文字列検索では,「モッツァレラチーズ 」「スパゲティ」「オリーブオイル」を検索クエリとして使用した．

編集距離を用いた類似文字列検索の手法であるベースライン法 jpedit と，発音照合を考慮した編集距離を用いる提案法 jpeditex のそれぞれで，料理名の表記揺れと，素材名の表記揺れを抽出した．

⋆1 表記揺れ文字列集合の中で文書頻度 (DF 値) が最大となる表記揺れ文字列

料理名の表記揺れは，検索結果文字列の上位 15 件までを，また，素材名の表記揺れは，検索結果文字列の上位 30 件までを，ベースライン法 jpedit と提案法 jpeditex で比較したところ，結果の上位はいずれの手法も共通であり，結果の下位に手法ごとの特徴が表れていることが確認できた．また，ベースライン法 jpedit と提案法 jpeditex ともに，提案法 jppm2 で抽出された表記揺れの文字列を含み，さらに，提案法 jppm2 では抽出されなかった文字列が検索結果に含まれていた．料理名「ラタトゥイユ」と材料名「モッツァレラチーズ」の検索結果の例を表 11 と表 12 に示す．表中の太字は提案法 jppm2 では抽出されなかった文字列である．提案法 jppm2 では，検索クエリに含まれない特徴を含む文字列は検索されないが，ベースライン法 jpedit と提案法 jpeditex では,「挿入」という操作が考慮されるため，検索クエリ「ラタトゥイユ」の先頭に「オム」が追加された文字列が検索結果に含まれている．ただし，このような文字列は，表記揺れ（同じ意味を持ち，綴りが異なる文字列）というよりも，類義語・関連語（同じ意味ではないが，類似する，あるいは，関連する意味を持つ文字列）として扱うべきであり，表記揺れではないと考える．

下線を付した文字は，ベースライン法 jpedit と提案法 jpeditex の各手法の特徴が反映された表記揺れ文字列である．表 11 を見ると「ラタツィユ」の「ツ」を「ラタトゥイユ」の「ト」に置換すると，同じ文字のグループであるためコストは 1 となり,「ィ」を「ゥ」に置換すると，同じ文字のグループであるためコストは 1 となり,「イ」を追加するとコストは 2 となり，合計すると距離は 4 となるため,「オムラタトゥーユ」よりも類似度が高いという結果になっている．これは，発音照合を考慮した編集距離が有効である例である．しかし，表 12 に示すように，提案法 jpedit では「モッツァレラチーズナド」，提案法 jpeditex では「モッツァレラダイス」が検索されており，これらも「モッツァレラチーズ」の表記揺れではない．これらの文字列は類義語，もしくは，関連語として扱うべきものであると考える．

編集距離を用いる表記揺れの抽出では，短い文字列，違いが小さいものに対しても適用可能であり，追加や削除があるものも見つけることができるが，ノイズも拾ってしまうという問題もある．編集距離の計算における追加，削除，置換のコストの検討と，文字のグループ分けの詳細化などが必要である．

5. おわりに

本稿では，料理レシピに含まれる料理名や素材名の表記揺れを抽出することを目的として，日本語の発音照合と文字列の編集距離を用いた類似文字列検索の手法を提案した．また，実際の料理レシピデータを用いて，表記揺れの少ない小規模なデータセットと，表記揺

表 11 編集距離を用いた料理名の表記揺れの検出例
Table 11 Spelling Variants of Recipe Titles with Edit/Editex.

順位	ベースライン法 (jpedit)	距離	順位	提案法 (jpeditex)	距離
1	ラタトゥイユ	0	1	ラタトゥイユ	0
2	ラタトイユ	2	2	ラタトゥィユ	1
3	ラタトゥィユ	2	3	ラタトゥウユ	1
4	ラタトゥウユ	2	4	ラタトゥーユ	1
5	ラタトゥユ	2	5	ラタトイユ	2
6	ラタトゥーユ	2	6	ラタトゥユ	2
7	ラタトュイユ	2	7	ラタトウィユ	2
8	ラタトィユ	4	8	ラタトュイユ	2
9	ラタトゥーイ	4	9	ラタトィユ	3
10	ラタトウィユ	4	10	ラタトゥーイ	3
11	ラタトウユ	4	11	ラタトウユ	3
12	ラタトューユ	4	12	ラタトューユ	3
13	**オムラタトゥーユ**	6	13	ラタツィユ	4
14	カブノラタトゥイユ	6	14	ラタトウィル	4
15	チキンラタトゥイユ	6	15	**オムラタトゥーユ**	5

れを大量に含む大規模なデータセットを作成し，提案法を用いた表記揺れ抽出の評価実験を行った．発音照合を用いる手法では，もとの文字列の詳細な特徴を削除して，曖昧性の高い文字列検索を行う手法 (提案法 jppm2) により，多くの表記揺れを集めることができるが，表記揺れではない文字列も検索されるため，検索結果の精度を高める後処理が必要である．発音照合を考慮した編集距離による表記揺れ抽出は，検索結果の上位に精度よく，表記揺れを集めることができるが，文字の挿入や追加，発音照合を考慮した置換の副作用により，表記揺れではない文字列も検索結果に含まれるため，得られた文字列を検証する後処理が必要である．これらの後処理の検討に加えて，今後の課題としては，提案法により得られた文字列集合を，料理レシピの検索に活用することを検討していく予定である．

謝辞 本研究は科研費 (課題番号: 21700273) の助成を受けたものである．

参 考 文 献

1) 井手一郎，上田真由美，間瀬健二，上田博唯，土屋誠司，小林亮博：2. 献立を決める (< 小特集 > 生活に役立つメディア処理 – 料理行動を科学する –)，電子情報通信学会誌， Vol.93, No.1, pp.33–38 (2010).
2) 山肩洋子，舩冨卓哉，上田博唯，辻 秀典，美濃導彦，中内 靖，宮脇健三郎，中村裕一，椎尾一郎：3. 料理を作る (< 小特集 > 生活に役立つメディア処理 – 料理行動を科学する –)，電子情報通信学会誌， Vol.93, No.1, pp.39–47 (2010).

表 12 編集距離を用いた材料名の表記揺れの抽出例
Table 12 Spelling Variants of Ingredients with Edit/Editex.

順位	ベースライン法 (jpedit)	距離	順位	提案法 (jpeditex)	距離
1	モッツァレラチーズ	0	1	モッツァレラチーズ	0
2	**モッァレラチーズ**	2	2	モッチァレラチーズ	1
3	**モッサァレラチーズ**	2	3	モッツアレラチーズ	1
4	モッチァレラチーズ	2	4	モッツェレラチーズ	1
5	**モッツァラレラチーズ**	2	5	モッツエレラチーズ	1
6	**モッツァレアチーズ**	2	6	**モッァレラチーズ**	2
7	**モッツァレチーズ**	2	7	**モッサァレラチーズ**	2
8	モッツァレッラチーズ	2	8	モッチェレラチーズ	2
9	**モッツァレラーズ**	2	9	**モッッアレラチーズ**	2
10	モッツァレーラチーズ	2	10	**モッツァラレラチーズ**	2
11	モッツアレラチーズ	2	11	**モッツァレアチーズ**	2
12	モッツェレラチーズ	2	12	**モッツァレチーズ**	2
13	モッツエレラチーズ	2	13	モッツァレッラチーズ	2
14	モッツッァレラチーズ	2	14	**モッツァレラーズ**	2
15	モッツャレラチーズ	2	15	モッツァレーラチーズ	2
16	**モッツラレラチーズ**	2	16	**モッツアレラチース**	2
17	モッツレラチーズ	2	17	モッツアレルチーズ	2
18	モツァレラチーズ	2	18	モッツッァレラチーズ	2
19	**マッツレラチーズ**	4	19	モッツャレラチーズ	2
20	モォツレラチーズ	4	20	**モッツラレラチーズ**	2
21	**モッアレラチーズ**	4	21	モッツレラチーズ	2
22	モッザレラチーズ	4	22	モツァレラチーズ	2
23	モッチェレラチーズ	4	23	**マッツレラチーズ**	3
24	モッチッァレラチーズ	4	24	**モッアレラチーズ**	3
25	**モッッアレラチーズ**	4	25	モッチッァレラチーズ	3
26	モッッツアレラチーズ	4	26	モッッツアレラチーズ	3
27	モッツァレラチーズナド	4	27	モッツァレラダイス	3
28	**モッツアレアチーズ**	4	28	**モッツアレアチーズ**	3
29	モッツアレッラチーズ	4	29	モッツアレッラチーズ	3
30	**モッツアレラチース**	4	30	モッツェレーラチーズ	3

3) 宮脇健三郎，尾関基行，木村 穣，相澤清晴，北村圭吾，山崎俊彦，森 麻紀，武川直樹：4. 食べる (< 小特集 > 生活に役立つメディア処理 – 料理行動を科学する –)，電子情報通信学会誌， Vol.93, No.1, pp.48–54 (2010).
4) 浜田玲子，井手一郎，坂井修一，田中英彦：料理テキスト教材における調理手順の構造化，電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理， Vol.85, No.1, pp.79–89 (2002-01-01).
5) 林 絵梨，吉岡 卓，東条 敏：日本語レシピ文における時間的関係構造の自動生成，

自然言語処理， Vol.10, No.2, pp.3–17 (2003).
6) 知秀柴田，禎夫黒橋：料理レシピテキストの構造解析とその応用，情報処理学会研究報告. 自然言語処理研究会報告， Vol.2004-NL-164, No.108, pp.117–122 (2004).
7) 高野哲郎，上島紳一：段取りの導出を行う調理支援システムの提案 (特集矢島脩三教授定年退職記念)，情報研究 : 関西大学総合情報学部紀要， Vol.22, pp.117–142 (2005).
8) 吉川祐輔，宮下芳明：グラフィカルデータフローによる調理レシピプログラミング言語の提案，情報処理学会研究報告. HCI, ヒューマンコンピュータインタラクション研究会報告， Vol.2010, No.4, pp.1–7 (2010).
9) 苅米志帆乃，藤井　敦：料理レシピテキストの構造解析とその応用，言語処理学会第18回年次大会発表論文集，pp.839–842 (2012).
10) 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 (セッション 3:インタラクションデザイン:理論と実践 (3))，情報処理学会研究報告. HCI, ヒューマンコンピュータインタラクション研究会報告， Vol.2007, No.41, pp.51–57 (2007).
11) 志土地由香，高橋友和，井手一郎，村瀬　洋：調理レシピテキストからの代替素材の発見，人工知能学会全国大会論文集， Vol.22, No.1B1-02, pp.1347–9881 (2009).
12) The U.S. National Archives and Records Administration: *The Soundex Indexing System*, (online), available from ⟨ http://www.archives.gov/research/census/soundex.html ⟩ (2007).
13) Philips, L.: The Double Metaphone Search Algorithm, *C/C++ Users Journal*, (online), available from ⟨ http://drdobbs.com/cpp/184401251 ⟩ (2000).
14) ドナルド・E. クヌース：*The Art of Computer Programming Volume 3 Sorting and Searching Second Edition* 日本語版，pp.375–376, アスキー (2004).
15) 言語処理学会：言語処理学事典，共立出版株式会社 (2009).
16) Zobel, J. and Dart, P.W.: Phonetic String Matching: Lessons from Information Retrieval, *SIGIR '96 Proceedings*, pp.166–172 (1996).
17) Yasukawa, M., Culpepper, J.S. and Scholer, F.: Phonetic Matching in Japanese, *Proceedings of SIGIR 2012 Workshop on Open Source Information Retrieval (OSIR 2012)*, Portland, Oregon, USA., pp.68–71 (online), available from ⟨http://opensearchlab.otago.ac.nz/⟩ (2012).

付　　録

A.1 実験データに含まれる料理名と素材名

表 13 Dataset-A に含まれる料理名・材料名の文書頻度 (DF 値) の高い上位 10 件
Table 13 Top 10 Titles/Ingredients in Dataset-A.

順位	料理名	文書頻度 (DF 値)	順位	材料名	文書頻度 (DF 値)
1	肉豆腐	6	1	しょうゆ	1882
2	若竹煮	7	2	塩	1833
3	生春巻き	5	3	酒	1486
4	たけのこごはん	5	4	砂糖	1307
5	豚肉のしょうが焼き	4	5	玉ねぎ	943
6	肉じゃが	4	6	こしょう	916
7	ロールキャベツ	4	7	にんにく	830
8	冷や汁	4	8	卵	822
9	さつま芋ごはん	4	9	水	737
10	だて巻き	4	10	みりん	737

図 1 Dataset-A に含まれる料理名の文書頻度 (DF 値)
Fig. 1 Document Freuqency (DF value) of Recipe Titles in Dataset-A.

Vol.2012-NL-208 No.7
2012/9/3

表 14 Dataset-B に含まれる料理名・材料名の文書頻度 (DF 値) の高い上位 10 件
Table 14 Top 10 Titles/Ingredients in Dataset-B.

順位	料理名	文書頻度 (DF 値)	順位	材料名	文書頻度 (DF 値)
1	ガトーショコラ	125	1	砂糖	145661
2	麻婆豆腐	123	2	塩	136400
3	バナナケーキ	117	3	卵	110129
4	豚の角煮	117	4	水	100246
5	ポテトサラダ	115	5	牛乳	73506
6	ひじきの煮物	107	6	醤油	71064
7	大根サラダ	105	7	バター	70784
8	ベイクドチーズケーキ	105	8	玉ねぎ	58388
9	カルボナーラ	105	9	酒	57340
10	お弁当	102	10	マヨネーズ	49950

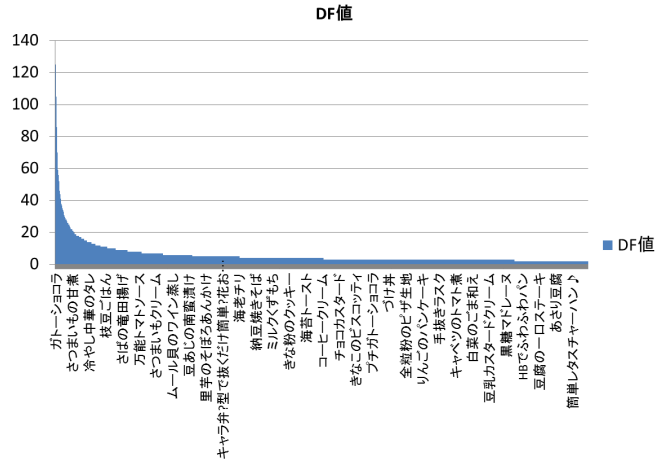

図 2 Dataset-B に含まれる料理名の文書頻度 (DF 値)
Fig. 2 Document Freuqency (DF value) of Recipe Titles in Dataset-B.