レシピに含まれる不使用な材料等に関する記述の抽出

山口泰弘[1]　染谷大河[2]　深澤祐援[1]　原島純[1]
[1] クックパッド株式会社　[2] 東京大学大学院 総合文化研究科
{yasuhiro-yamaguchi,yusuke-fukasawa,jun-harashima}@cookpad.com
taiga98-0809@g.ecc.u-tokyo.ac.jp

概要

レシピ中のテキストには，しばしば「卵不使用」のように特定の材料等について不使用である旨を示す記述が含まれる．こうした記述は，例えば検索において「卵」で検索した際に「卵不使用」のレシピが表示される等，不都合な場合がある．そこで，本研究では系列ラベリングモデルを用いてレシピに含まれるテキストから不使用な材料／調理器具／調理工程を抽出する手法を提案する．また，レシピ特有の表現に着目したデータ拡張手法も併せて検討した．実験の結果，F1 85.1% の抽出精度を達成し，データ拡張手法についてもその有効性が示された．

1　はじめに

クックパッドをはじめとしたユーザー投稿型のレシピサービスには多様なレシピが存在する．投稿されるレシピの中には，しばしば特定の材料や調理器具等を使わずに調理できる旨を明示的に記載したものがある．「小麦粉不使用」や「卵アレルギー対応」等特定の食材を使わないことを示すものや，「火を使わずに」「圧力鍋なしで」のように調理法や道具を使わないことを表すもの等，対象となるアイテムや行為，表記方法は様々である．こうした記述は，閲覧者がアレルギーに配慮したレシピを探す際や，より簡単な調理法を探す際に有益な情報となり得る．

一方，このような不使用な食材等に関する記述はレシピ検索等の応用において不都合な場合がある．単純なテキストのマッチングに基づく検索では，「卵」で検索した際に「卵不使用」と記載のあるレシピが検索結果に表示されることがある．しかし材料名で検索を行うユーザーの多くはその材料を使ったレシピを探している可能性が高く，当該材料を使わないレシピが検索結果に載ることは不自然だと考えられる．こうした課題を解決するためには，予めレシピ中から不使用と書かれた材料等の記述を検出しておく，といった対応が必要になる．

そこで本研究では，レシピ中のテキストから系列ラベリングモデルを用いて不使用な材料／調理器具／調理工程を抽出する手法を提案する．また，不使用に関連する記述に着目したいくつかのデータ拡張手法を検討し，実験を通してその効果を調査する．

2　関連研究

系列ラベリングは入力系列の各トークンにラベルを割り当てる手法であり，固有表現認識や品詞タグ付け等のタスクで利用されている．本研究では固有表現認識タスクの代表的な手法である BiLSTM-CRF [1] を元にしたモデルを利用する．

レシピ中のテキストを対象にした系列ラベリングの研究として，材料や調理工程を検出する手法がいくつか提案されている [2] [3]．また，Cookpad Parsed Corpus (CPC) [4] のようにレシピに対して形態素，固有表現，係り受けの情報を付与する取り組みもある．CPC では調理中に除去される材料の検出等，文脈を考慮したより詳細な解析が試みられている．

また，レシピの材料や調理工程に着目した研究として，材料名の正規化 [5]，材料欄に含まれる非材料の検出 [6]，タイトルに基づく材料の予測 [7]，調理工程のグラフ構造化 [8] [9]，材料の上位下位関係の検出 [10] 等がある．これらの研究はレシピ中で使用される材料や調理工程に着目しているが，本研究では不使用な項目に焦点を当てる．

テキストのデータ拡張に関する研究には，単語の挿入・削除・置換等の処理を行う EDA [11] 等があり，文分類タスク等での有効性が確認されている．また，Dai ら [12] は固有表現認識タスクを対象にしたデータ拡張としてメンションの置換やメンション以外のトークンのシャッフル等の手法を提案し，予測性能の向上を示している．これらの研究とは異なり，本研究ではレシピ固有の表現に着目したデータ拡張手法を用いることで性能改善を試みる．

表 1 アノテーションしたレシピの例

	テキスト
タイトル	さつまいものココアクッキー
説明	[裏ごし]$_{\text{PROC}}$ 必要なしのさつまいもを使ったクッキー．[バター]$_{\text{ING}}$，[卵]$_{\text{ING}}$ 不使用のボウルひとつでできます．
材料	さつまいも，ココア，サラダ油，シナモンパウダー，塩，水，砂糖，薄力粉か中力粉

表 2 データセットの統計

	学習	検証	テスト
# recipes	6,000	1,500	2,500
# texts (titles/descriptions)	12,000	3,000	5,000
# texts w/ spans	1,730	467	716
# unused ingredient spans	1,568	451	634
# unused procedure spans	435	106	158
# unused tool spans	211	63	89

3 提案手法

3.1 データセットの作成

不使用に関する記述に着目したデータセットを新たに作成するために，172 万品のレシピデータを含む Cookpad Recipe Dataset [13] から「不使用」等の記述を含むレシピ 7,000 件と無作為に選択したレシピ 3,000 件の計 1 万レシピを抽出した．1 人のアノテータにより，抽出したレシピのタイトル，説明のテキストに対して不使用であることが明記された材料／調理器具／調理工程のスパンのアノテーションを行った．表 1 にアノテーションしたレシピの例を示す．作成したーデータセットは学習用／検証用／テスト用に分割し，学習と評価に利用した．分割したデータセットの統計を表 2 に示す．データの収集やアノテーション方法に関するより詳しい情報は付録 A に記す．

3.2 データ拡張

作成したデータセットの観察から，不使用に関連する記述には多様性があり，各表現の出現頻度は不均衡であることがわかった．そこで，以下のデータ拡張手法を適用することでこの問題に対処する．

ReplaceSuffix “〜なし” のように，対象となるスパンの後方に続くことでそれらが不使用であることを示す表現について，それらの表現をランダムに入れ替える手法．付録 B.1 に示した表現に該当する箇所を，他の表現でランダムに置き換える．

ReplaceSeparator “A・B を使わない...” のように複数のアイテムを対象に不使用であることを示す表現について，列挙のパターンをランダムに入れ替える手法．アノテーションされた直近 2 つのスパンに挟まれたテキストのうち，付録 B.2 に示した表現に該当するものを他の表現で置き換える．

SuffixToPrefix “〜なし” のようにスパンの後方に続く表現を，“ノン〜” (“ノンオイル”，等) のようにスパンの前方に付く表現に置き換える手法．アノテーションされたスパンの後に付録 B.3 に示す表現が続く場合，“ノン〜” という形式に置き換える．

3.3 同義語による置換

上記のデータ拡張手法では，対象となる表現以外は変化せず，似た表記を持つテキストが多く生成されるため過学習につながる可能性がある．そこで，データのバリエーションを増やすためにデータ拡張によって生成されたテキストを対象としてテキスト中に含まれる単語をランダムに同義語に置き換える処理を合わせて行う．テキストを分かち書きした後，各単語について料理関連語を対象として事前に定義した同義語辞書をもとに他の同義語へとランダムに置き換える．

3.4 不使用に関する記述の抽出

図 1 にスパン抽出に用いる系列ラベリングモデルの構造を示す．BiLSTM-CRF [1] をベースとしたニューラルモデルであり，レシピ中のテキストを入力として，各トークンに対して BIOUL タグを出力することで不使用なアイテムのスパンと，それらのカテゴリ (材料／調理器具／工程) を予測する．

入力のテキストは形態素解析を行なったあと文字レベルでトークン化し，文字系列 $(c_1, \cdots, c_T)$ と品詞系列 $(\tau_1, \cdots, \tau_T)$ に変換する．単語の境界情報を与えるために，品詞系列には図 1 に示すように BIOUL スキーマを用いて文字ごとに品詞情報を割り当てる方法を採用した．エンコーダへの入力は文字／品詞に対応する学習可能な埋め込み表現 $\boldsymbol{e}_{\text{CHAR}}(c_t), \boldsymbol{e}_{\text{POS}}(\tau_t)$ を結合したベクトル $\boldsymbol{x}_t$ の系列とする．

$$\boldsymbol{x}_t = \boldsymbol{e}_{\text{CHAR}}(c_t) \| \boldsymbol{e}_{\text{POS}}(\tau_t) \tag{1}$$

$$\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T = \text{Encoder}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T) \tag{2}$$

エンコーダの出力から，CRF を通してタグ系列 $\boldsymbol{y} = (y_1, \cdots, y_T)$ の確率を計算する．以下の $\theta_{y_{t-1}, y_t}, \boldsymbol{w}_{y_t}$ は CRF の学習パラメータを表す．

$$p(\boldsymbol{y}|\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T) \propto \exp\left(\sum_{t=2}^{T} \theta_{y_{t-1}, y_t} + \sum_{t=1}^{T} \boldsymbol{w}_{y_t}^{\top} \boldsymbol{h}_t\right) \tag{3}$$

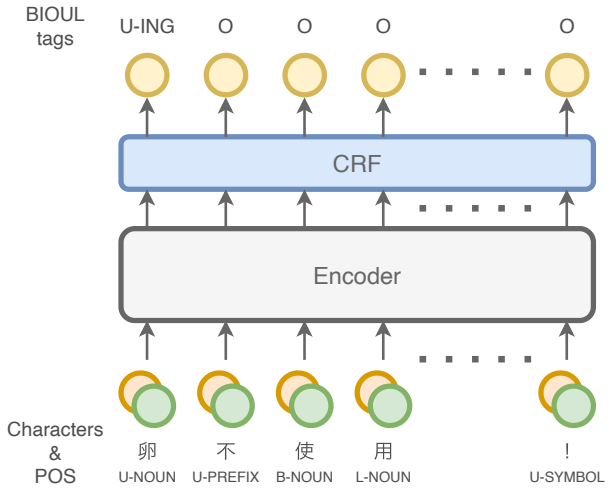

図 1　スパン抽出モデル

学習では以下の負の対数尤度を最小化するようにモデルのパラメータを最適化する．

$$\mathcal{L} = -\sum_{i} \log p(\boldsymbol{y}^{(i)} | \boldsymbol{h}_1^{(i)}, \cdots, \boldsymbol{h}_T^{(i)}) \tag{4}$$

3.5　スパンのフィルタリング処理

本研究で抽出対象とする 3 種類のスパンのうち，特に材料に関するものはレシピの材料欄の情報と組み合わせることで抽出精度の改善が期待できる．そこで，レシピの材料欄に基づいて抽出されたスパンのフィルタリングを行う方法を提案する．

前処理として，抽出された材料のスパンと材料欄の材料名を対象に Harashima ら [5] の手法を利用して材料名の正規化を行う．正規化した材料名同士を比較し，抽出した材料スパンのうち材料欄に含まれるものがあればそれらを取り除く．このフィルタリング処理により誤って抽出された可能性の高いスパンを除去することができるため，モデルの適合率の改善が期待できる．

4　実験と結果

4.1　実験方法

各種データ拡張手法を適用したデータセットを用いてスパン抽出モデルの学習を行い，性能の評価を行った．データ拡張処理では各手法について拡張対象の表現と変換パターンを可能な限り適用した．学習データにデータ拡張を施した場合のデータセット規模の変化を表 3 に示す．より詳細なデータ拡張の設定，結果は付録 B に記す．

表 3　各データ拡張手法適用後のデータセット規模

	# texts	# texts w/ spans
None	12,000	1,730
ReplaceSuffix	21,170	10,900
ReplaceSeparator	12,704	2,434
SuffixToPrefix	12,438	2,168
All	29,499	19,229

各テキストは予め fugashi [14] と IPAdic[1] により形態素解析を行い，その結果得られた品詞情報をモデルの入力とする．また，モデルのエンコーダには 2 層の双方向 GRU (BiGRU) [15] を利用した．

モデルの学習では分割したアノテーションデータのうち学習データを用いてモデルのパラメータの更新を行い，検証データを用いた評価による Early Stopping を実施した．モデルのパラメータの最適化には AdamW [16] を用いた．

モデルの性能評価にはスパンベースの F1 ／ Precision ／ Recall を利用した．学習時には Early Stopping を行い，検証データにおいて最も高い F1 スコアを達成した時点のモデルを採用した．

4.2　ベースライン手法

ベースライン手法として，以下 2 つの抽出方法を採用した．どちらの手法についても，3.5 節のフィルタリング処理と併せて予測を行なった．

Dictionary-based Extractor　Cookpad のサービスで利用されている独自の辞書に基づいてテキスト中から候補となる材料・調理工程・調理器具のスパンを抽出する手法．

Dependency-based Extractor　テキストの依存構造解析を行い，「不使用」等のキーワードと係り受け関係にある名詞を抽出する手法．依存構造解析には GiNZA[2] を利用した．抽出したスパンのラベルは Dictionary-based Extractor でも利用した独自の辞書に基づいて決定した．

5　結果と考察

表 4 に実験結果を示す．2 つのベースライン手法と BiGRU-CRF を比較すると，F1 において BiGRU-CRF モデルが+20%以上高い性能を示した．Dictionary-based Extractor については，Recall は比較的高いが，材料以外はフィルタリング処理されないことや，“〜にもあう” のような表現を伴うレシピ

1）　Cookpad のサービスで利用している独自のレシピドメイン向けのユーザー辞書と併用して解析を行なった．
2）　https://megagonlabs.github.io/ginza/

表 4 テストデータにおけるモデルの予測精度

	F1	Precision	Recall
Dictionary-based Extractor	18.5	10.5	77.3
Dependency-based Extractor	63.9	82.8	52.0
BiGRU-CRF	83.0	81.9	84.2
+ ReplaceSuffix	83.9	85.5	82.3
+ ReplaceSeparator	83.8	88.7	79.5
+ SuffixToPrefix	83.5	85.5	81.6
+ All	**85.1**	**85.8**	**84.3**

表 5 ラベルごとの予測精度

	F1	Precision	Recall
材料	86.8	87.3	86.4
調理器具	86.0	89.2	83.1
調理工程	77.3	78.1	76.6

と直接関係のない食品のスパンも抽出されたことで， Precision は他の手法に比べて特に低い結果となった．Dependency-based Extractor では BiGRU-CRF と比較して高い Precision を達成したが，予め定義したパターン以外は抽出できないことや依存構造解析の誤り等の影響で，特に Recall が低くなる結果となった．

データ拡張の効果　各データ拡張手法に着目すると，すべての手法でデータ拡張を用いない場合と比べて F1 スコアの上昇が見られた．データ拡張を個々に適用した場合，いずれの手法においても Precision の上昇が確認できたが，Recall については下がる傾向が見られた．一方，全てのデータ拡張手法を適用した場合には全ての指標について改善が見られ，Precision では+3.9%の改善，Recall においても +0.1% と若干の上昇となった．データ拡張で考慮された記述パターンを含むテキストについて検出精度が向上することで Precision の上昇につながったと考えられる．一方，false-negative な誤りの多くは “〜の代わりに” や “〜を避けたい” のように今回のデータ拡張の設定で考慮されていないものであり，Recall の更なる改善にはこうした表現に対応する学習データの追加やデータ拡張の設定が必要になると考えられる．

ラベルごとの予測精度　表 5 に BiGRU-CRF + All モデルのラベルごとの予測精度を示す．F1 による比較では材料の抽出精度が最も高く，次いで調理器具，調理工程 という結果になった．特に調理工程については，調理器具よりも学習データに含まれるスパンの数が多いにも関わらず，他のラベルと比較していずれの指標においても低い性能を示した．材料や調理器具のラベルが付いたスパンは，その多くは一般名詞であることから予測が比較的容易であると考えられる．しかし，調理工程については学習データに含まれる事例が比較的少ない一方で「一晩寝かせ」のように複数の単語からなるもの等表現のバリエーションが多いため，他の 2 つのラベルに比べて予測が困難であると考えられる．

表 6 同義語による置換とスパンのフィルタリングの効果

	F1	Precision	Recall
BiGRU-CRF + All	**85.1**	**85.8**	84.3
w/o synonym replacement	83.7	85.5	82.1
w/o span filtering	84.1	83.5	**84.8**

同義語による置換の効果　同義語による置換処理を施していないデータセットで BiGRU-CRF + All モデルを学習した結果を表 6 に示す．置換処理を行わない場合，すべての指標において性能の低下が見られたが，特に Recall において -2.2% の顕著な性能低下が見られた．同義語置換なしの設定では共通した語句を持つテキストが多く生成されるため，モデルが限られたパターンのみに適合したことで Recall が減少したと考えられる．

スパンのフィルタリング処理の影響　スパンのフィルタリング処理を適用しない場合の性能比較を同様に表 6 に示す．フィルタリング処理を適用した場合と比較して，F1 で -1.0%，Precision では -2.3% の性能低下が見られた．特に Precision の変化から，提案したフィルタリング処理を用いることで BiGRU-CRF モデルが誤って検出したスパンを排除できることがわかった．一方，Recall においてはフィルタリング処理しない場合の方が +0.5% ほど高い精度となった．これはフィルタリング処理によって本来抽出するべきスパンが誤って排除されたことによるものと考えられる．フィルタリング処理の導入により， Recall においては若干の性能低下が見られたが，Precision においては期待した改善が見られ，F1 の向上に寄与した．

6　おわりに

レシピ中のテキストに含まれる不使用な材料／調理器具／調理工程に関する記述を抽出するために，系列ラベリングを用いた抽出手法と，レシピ特有の表現に着目したデータ拡張手法を提案した．実験の結果，F1 スコアで 85.1% の抽出精度を達成し，提案したデータ拡張手法の有効性が示された．今後は抽出精度の改善やレシピ検索等の応用に向けて，系列ラベリング以外の抽出手法の検討や検索タスクによる評価等を行いたい．

参考文献

[1] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)**, 2016.

[2] Makoto Hiramatsu, Kei Wakabayashi, and Jun Harashima. Named Entity Recognition by Character-based Word Classification using a Domain Specific Dictionary. In **Proceedings of the 20th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2019)**, 2019.

[3] Tetsuro Sasada, Shinsuke Mori, Tatsuya Kawahara, and Yoko Yamakata. Named Entity Recognizer Trainable from Partially Annotated Data. In **Conference of the Pacific Association for Computational Linguistics (PACLING 2015)**, 2015.

[4] Jun Harashima and Makoto Hiramatsu. Cookpad parsed corpus: Linguistic annotations of Japanese recipes. In **Proceedings of the 14th Linguistic Annotation Workshop (LAW 2020)**, 2020.

[5] Jun Harashima and Yoshiaki Yamada. Two-step validation in character-based ingredient normalization. In **Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management (CEA 2018)**, 2018.

[6] Yasuhiro Yamaguchi, Shintaro Inuzuka, Makoto Hiramatsu, and Jun Harashima. Non-ingredient Detection in User-generated Recipes using the Sequence Tagging Approach. In **Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)**, 2020.

[7] 深澤祐援, 西川荘介, 原島純. マルチラベル分類による材料推薦モデル. 言語処理学会第 27 回年次大会発表論文集 (NLP 2021), 2021.

[8] Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. Flow Graph Corpus from Recipe Texts. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)**, 2014.

[9] Lucia Donatelli, Theresa Schmidt, Debanjali Biswas, Arne Köhn, Fangzhou Zhai, and Alexander Koller. Aligning actions across recipe graphs. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)**, 2021.

[10] Young-joo Chung. Finding food entity relationships using user-generated data in recipe service. In **Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM 2012)**, 2012.

[11] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)**, 2019.

[12] Xiang Dai and Heike Adel. An analysis of simple data augmentation for named entity recognition. In **Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)**, 2020.

[13] Jun Harashima, Michiaki Ariga, Kenta Murata, and Masayuki Ioki. A large-scale recipe and meal data collection as infrastructure for food research. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)**, 2016.

[14] Paul McCann. fugashi, a tool for tokenizing Japanese in python. In **Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS 2020)**, 2020.

[15] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In **Proceedings of Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST 2014)**, 2014.

[16] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In **Proceedings of International Conference on Learning Representations (ICLR 2019)**, 2019.

A　データセットの詳細

A.1　レシピの収集方法

不使用に関する記述を含むレシピを効率的に収集するために，Cookpad Recipe Dataset [13] に収録されたレシピを対象に，タイトル・説明文に以下の表現を含むものをランダムに 7,000 件収集した:

使わ／つかわ／使い／使って／つかって／使用／不要／なし／無し／いら／要ら／せず／しない／しなく／アレルギー

また，データの偏りを考慮して上記で選ばれなかったレシピの中からランダムに 3,000 件のレシピを取得し，データセットに加えた.

A.2　アノテーションの詳細

収集したレシピのタイトル (title), 説明 (description) を対象に，当該テキストの記述のみからそのレシピ全体で不使用であることがわかるもののみを対象にスパンのアノテーションを付与した．また「〇〇なしでも OK」のように，アイテムを使わないことがオプションである旨が記載されたものに関してはアノテーションを行わないこととした.

B　データ拡張手法の詳細

各種データ拡張手法について，詳細な設定を以下に示す．また，表 7 に各手法で対象となる学習データ中のテキスト／スパン数の統計情報を示す.

表 7　データ拡張の対象となるテキスト／スパン数

	w/ prefix	w/ suffix	w/ separator
# texts	60	915	187
# spans	69	946	242

B.1　ReplaceSuffix

スパンの後方に続く表現が以下のいずれかに当てはまるとき，他の表現による置き換えを行う: なし／いらず／使わない／使ワナイ／を使わない／不要／不使用／要らず／なし／無し／ゼロ

B.2　ReplaceSeparator

スパンに挟まれた表現が以下のいずれかに当てはまるとき，他の表現による置き換えを行う: ・(中点) ／&(アンパサンド) ／，(読点) ／や（並立助詞）／(空白)

B.3　SuffixToPrefix

スパンの後方に続く表現が以下のいずれかに当てはまるとき，それらを消してスパンの前方に “ノン” を追加する処理を行う: なし／ナシ／無し／不使用

B.4　データ拡張の例

全てのデータ拡張手法と同義語による置換を行ない得られたテキストを表 8 に示す.

表 8　データ拡張により生成された文の例

(a) 例文 1

原文	卵・乳製品不使用の簡単もっちりスコーン♪
生成例 1	卵乳製品いらずの超簡単モッチリスコーン♪
生成例 2	卵＆乳製品不要のかんたんもっちりすこーん♪

(b) 例文 2

原文	おからでイタリアン☆ピザ風☆小麦粉不使用
生成例 1	卵ノ花でイタリア料理☆ピザ風☆ノン小麦粉
生成例 2	オカラでイタリア☆ピザ風☆小麦粉を使わない