レシピ解析の現状と課題: Cookpad Parsed Corpus を例として

平松 淳　原島 純
クックパッド株式会社
{himkt, jun-harashima}@cookpad.com

1 はじめに

インターネット上にレシピを投稿する行為が一般的になり，計算機で処理できる電子化されたレシピが増加している．これにより，レシピデータの解析への注目が高まり，様々なデータセットが整備されている [1, 2, 3]．レシピデータが公開されたことで，データを利用した応用的な研究トピックにも関心が集まり，レシピ検索 [3, 4]，レシピ生成 [5, 6]，レシピ質問応答 [7]，そしてカロリー推定 [8] のような応用的なタスクも提案されている．

レシピ研究が活発になった一方，レシピテキストを解析するための基礎的な言語資源は未だ充分に整備されていない．利用可能な既存の言語資源の多くは新聞記事データなどを基に構築されている [9, 10] が，レシピテキストにはこのようなデータには出現しない用語が数多く存在する．例えば「炒める」「煮込む」のような調理動作に関する表現は新聞記事にあまり出現しない．また，「おにぎらず」のような新語も日々レシピテキストの中で誕生している．このような用語を含むレシピドメインのコーパスを構築し，言語解析器を開発することで，レシピ解析の精度を改善できると考えられる．

この課題に対し，著者ら [11] は Cookpad Parsed Corpus (CPC) を提案した．CPC にはレシピサービスのクックパッド [1] に投稿されたレシピから無作為に抽出した 500 品のデータが含まれており，各レシピのタイトルおよび調理手順に形態素，固有表現，そして係り受けの情報が付与されている．

本稿では， CPC のアノテーション情報について簡単に解説し， CPC を用いたベンチマーク実験の設定・結果を報告する．その後，各実験において明らかになったレシピ解析の現状と課題について考察する．また， 本研究で使用したソースコードおよび Dockerfile を GitHub [2] で公開する．

1） https://cookpad.com
2） https://github.com/cookpad/cpc1.0

2 関連研究

日本語のレシピドメインの言語資源に関する研究として，Mori ら [12] によるフローグラフコーパスが存在する．Mori らはクックパッドに投稿されたレシピから 266 品のレシピを抽出し，調理手順のテキストに対してグラフ形式の情報を付与した r-FG コーパスを提案した．Yamakata ら [13] も英語のレシピで同様のコーパスを提案している．また，笹田ら [14] はクックパッドから無作為に抽出した 436 件のレシピに対し，Mori らの研究で定義された固有表現タグに基づく固有表現情報を付与した．

言語以外のデータセットとして，西村ら [15] は r-FG コーパスを基に，各レシピの画像に対してバウンディングボックスを付与し，r-FG コーパス中の料理用語に対応させた r-FG-BB データセットを提案した．西村らはレシピ画像中のバウンディングボックスが示すエンティティをレシピテキスト中から発見するタスクに取り組み，レシピにおける言語と画像を組み合わせた研究の可能性を示した．

これに対し， CPC は 500 品のレシピに対して形態素，固有表現，係り受けの情報を付与した．CPC は Cookpad Recipe Dataset (CRD) [2] から無作為に抽出した 500 レシピに対してアノテーションを実施している．さらに， CPC は Cookpad Image Dataset (CID) [16] と互換性がある．CID は CRD に含まれるレシピについて，完成写真，および調理手順写真 474 万枚の画像を収録したデータセットである．CPC と CID を組み合わせることで，言語と画像を組み合わせた研究が可能である．

3 Cookpad Parsed Corpus

CPC では図 1 のようなフォーマットでアノテーションが実施されている． 1 つの調理手順 (Step) は複数の文を含む可能性があるため， Step の中に Sentence が存在する．本章では CPC の形態素，固有表現，そして係り受け情報について簡単に述べる．

```
# Step-ID:1
# Sentence-ID:1-1
* 0 4D 1/2 主題
生　　接頭詞, 名詞接続,*,*,*,*, 生, ナマ, ナマ, B-Fi
鮭　　名詞, 一般,*,*,*,*, 鮭, サケ, サケ, I-Fi
は　　助詞, 係助詞,*,*,*,*, は, ハ, ワ, O
* 1 2D 1/2 補足語
一口　名詞, 一般,*,*,*,*, 一口, ヒトクチ, ヒトクチ, B-Sf
大　　名詞, 一般,*,*,*,*, 大, ダイ, ダイ, I-Sf
に　　助詞, 格助詞, 一般,*,*,*, に, ニ, ニ, O
* 2 4P 0/0 述語
切り　動詞, 自立,*,*, 五段・ラ行, 連用形, 切る, キリ, キリ, B-Ap
* 3 4D 0/1 補足語
塩　　名詞, 一般,*,*,*,*, 塩, シオ, シオ, B-Fi
を　　助詞, 格助詞, 一般,*,*,*, を, ヲ, ヲ, O
* 4 -1O 0/0 述語
ふる　動詞, 自立,*,*, 五段・ラ行, 基本形, ふる, フル, フル, B-Ap
。　記号, 句点,*,*,*,*,。,。,。, O
EOS
```

図 1: CPC のアノテーション例

表 1: 固有表現タグの一覧

大分類	固有表現タグ	説明	頻度
F (食材)	Fi	食材	5,768
	Fe	除外される食材	381
	Fd	料理名	534
	Fa	料理の属性	623
T (道具)	Tg	調理器具	2,148
	To	道具	174
	Ta	道具の属性	29
A (動作)	Ap	調理者の動作	7,959
	Af	食材の動作	973
	At	道具の動作・変化	95
S (状態)	Sf	食材の状態	1,210
	St	道具の状態	344
	Sap	動作の状態	792
X (その他)	X	上記以外	389

3.1 形態素

CPC のテキストには IPA 品詞体型をベースにした形態素情報が付与されている．形態素のアノテーション作業は文の分割，自動解析，そして解析結果の修正の 3 ステップで実施した．まず人手で文境界の判定を実施した．その後，各文を MeCab [17] および mecab-ipadic を用いて解析した．最後に解析結果を人手で確認し，誤った形態素情報を修正した．

3.2 固有表現

レシピテキストには様々な料理ドメインの用語が出現する．食材，調理器具，調理動作 (例. 切る, 煮る, 炒める)，分量・時間 (例. 200g, 1 カップ, 1 時間, 冷めるまで) など，多様な用語が存在する．

CPC では笹田ら [14] の用語タグを参考に表 1 に示すような用語タグを定義した．CPC の用語タグは笹田らが定義したものより細分化されている．例えば，笹田らのコーパスでは食材を表すタグは “F” であるが，CPC では “Fi”，“Fe” などのタグの細分化されたタグを定義している．これは，レシピ中に出現する食材の中には調理中に除去される部分（例. 魚の「わた」）が存在し，それらは調理中に使われる材料とは区別すべきであると考えたためである．

固有表現のアノテーションは IOB2 形式で実施されている．図 1 のように，各形態素に対して “B-Fi” “I-Fi” “O” のようなタグが付与されている．

3.3 係り受け

CPC のテキストには文節同士の係り受け情報も付与されている．図 1 の例において，各文節の先頭には “*” から始まる行が存在し，文節の ID (0 オリジン), 係り先文節 ID + 係り受けラベル, 主辞/機能語フィールド, 文節ラベルの情報が記述されている．係り受けのアノテーションは自動解析と結果の修正からなる 2 ステップで実施した．はじめにテキストを CaboCha [18] で解析し，その後誤った文節の区切りおよび係り受けを人手で修正した．

4 実験

本章では CPC のベンチマークタスクの実験設定，およびそれぞれの結果について報告する．実験の際は CPC を訓練・検証・評価のために 3 つに分割しており，各データに含まれる文，形態素，固有表現，そして文節の数は表 2 に示す通りである．

表 2: 実験データの統計量

	訓練	検証	評価
レシピ数	400	50	50
文数	3,740	529	469
形態素数	48,826	7,280	6,040
固有表現数	17,645	2,562	2,153
文節数	20,934	3,016	2,551

4.1 形態素解析

形態素解析の実験には MeCab [17] を利用した．辞書として mecab-ipadic および mecab-ipadic をベースに CPC でパラメータを再学習した mecab-ipadic-cpc

表 3: 形態素解析の実験結果

タスク	システム辞書	精度	再現率	F 値
単語分割	mecab-ipadic	94.82	95.18	95.00
	mecab-ipadic-cpc	95.69	95.84	95.77
素性完全一致	mecab-ipadic	88.69	89.02	88.85
	mecab-ipadic-cpc	90.91	91.06	90.98

の 2 つを用意し，性能を比較した．

実験結果を表 3 に示す．評価指標は単語分割の精度・再現率・F 値，および単語の素性の完全一致の精度・再現率・F 値とした．ここで，単語の素性とは単語の読み，品詞，活用形などの情報を指す．単語分割・素性完全一致両者の指標において，mecab-ipadic-cpc が mecab-ipadic を上回っている．この結果から， CPC の有用性が確認できた．

一方で，Kudo ら [17] は新聞記事を基にして構築されたコーパスにおいて同様の実験を実施し，素性完全一致の評価で 96 ポイント以上の精度を報告している．このことから，レシピテキストの形態素解析にはまだ改善の余地があることが示唆される．

mecab-ipadic-cpc の解析結果について調査したところ，カタカナの単語のひらがな表記や略語を含む誤りが存在することがわかった．「チーズクリーム サラミ あすぱら パスタ」というレシピタイトルを例に挙げる．mecab-ipadic-cpc はこれを「チーズ / クリーム / サラミ / あす / ぱら / パスタ」と分割した．「あすぱら」は本来 1 単語になるべきだが，「あす (明日)」が辞書に含まれていたため，このように 2 つの形態素に分離してしまったと考えられる．

略語を含む誤りの例として「新じゃがは２～３等分に切る」というレシピテキストの解析結果を示す．mecab-ipadic-cpc はこのテキストを「新 / じゃが / は / ２ / ～ / ３ / 等分 / に / 切る」と正しく分割した．しかしながら，素性をみると「じゃが」の品詞として「接続詞」が予測されていた．これは解析器が「じゃが (じゃがいも)」を「～じゃが」のような逆接の接続詞として認識した結果である．CPC を用いた再学習の有用性は示された一方，ユーザ投稿型レシピサービス特有の難しさも明らかになった．

4.2 固有表現抽出

CPC の固有表現タグに基づいて訓練された既存の固有表現抽出モデルは存在しない．このため，既存の 2 種類の手法を用いてモデルを訓練し，参考値としてそれぞれのモデルの性能を報告する．

表 4: 固有表現抽出の実験結果

モデル	正解率	精度	再現率	F 値
PWNER	87.48	73.61	81.37	77.30
BiLSTM-CRF	90.13	85.95	85.56	85.75

表 5: 係り受け解析の実験結果

タスク	モデル	正解率
level0	cabocha-pretrain	91.49
	cabocha-cpc	94.20
level1	cabocha-pretrain	89.58
	cabocha-cpc	92.89
level2	cabocha-pretrain	86.90
	cabocha-cpc	91.14
sentence	cabocha-pretrain	70.36
	cabocha-cpc	78.04

1 つ目の手法は PWNER [19] である．PWNER は点推定と動的計画法に基づく手法であり．笹田ら [14] によるレシピ用語の抽出の研究で利用されている．本研究における実験では， パラメータはすべてデフォルトのものを利用した．

もう 1 つの手法は BiLSTM-CRF [20] である．BiLSTM-CRF の実装は pyner [21] を利用した．BiLSTM-CRF には文字レベルの特徴量と単語レベルの特徴量が存在する．文字レベルの特徴量については 25 次元の文字分散表現を 50 次元の Bi-LSTM に入力して抽出し，単語レベルの特徴量については文字レベルの特徴量と 100 次元の単語分散表現を結合した 150 次元の特徴ベクトルを 200 次元の Bi-LSTM に入力して抽出している．モデルの訓練には SGD を利用し，学習率は 0.01，バッチサイズは 10 に設定した．また，勾配のクリッピングを実施しており，しきい値は 5.0 とした．

実験結果を表 4 に示す．評価指標は単語レベルの正解率，フレーズレベルの精度・再現率・F 値である．フレーズレベルの指標については各固有表現タグの指標の平均をとる．BiLSTM-CRF が PWNER よりよい性能を発揮していることが確認できる．

BiLSTM-CRF がどのような入力に対して誤った予測をしているのか調査したところ，特に未知語を含む固有表現の抽出に失敗していることがわかった．例えば，「トマト / の / 水煮 / 缶 / を / 加える / と / 、 / 気分 / が / 変わり / ます」という文の「気分」に対して，BiLSTM-CRF は「Fa (食材の状態)」のカテゴリを予測していた．「気分」はレシピ作者の感情で

表 6: アノテーション誤りの例 (“//” は文節の区切りを表す)

原文	CPC のアノテーション	誤り修正後	修正点
フライパンに長ネギ を並べてワインコンソメを入れて お好みの硬さまで煮る。	フライパンに // 長ネギ を // 並べて // **ワインコンソメ**を // 入れて // お好みの // 硬さまで // 煮る。 //	フライパンに // 長ネギ を // 並べて // **ワイン // コンソメ**を // 入れて // お好みの // 硬さまで // 煮る。 //	**文節境界**の修正 「**ワインコンソメ**」 => 「**ワイン / コンソメ**」
じゃがいももまた、 ラップにくるんで、 電子レンジ (600w設定)で焼く 4〜5分ほど加熱調理をしておく。	じゃがいももまた、 // ラップに // くるんで、 // 電子レンジ (600w設定)で // **焼く [EOS]** 4〜5分ほど // 加熱調理を // しておく。	じゃがいももまた、 // ラップに // くるんで、 // 電子レンジ (600w設定)で // **焼く(約) 4〜5分ほど** // 加熱調理を // しておく。	**文境界**の修正 「電子レンジで**焼く [EOS]**」 => 「電子レンジで**焼く (約)** …」

あり，該当する固有表現タグは存在しないため，本来は「O」と予測するべきであった．「気分」は CPC にほとんど出現しない単語だったため未知語であり，BiLSTM-CRF には「(食材の) 色が変わる」のような表現と区別できなかったためにこのような予測誤りをしたと考えられる．BiLSTM-CRF が予測誤りをしたフレーズは 301 件あり，このうち 109 件は 1 つ以上の未知語を含むフレーズであった．

4.3 係り受け解析

係り受け解析の実験には CaboCha [18] を利用した．CaboCha 付属のモデル，および付属モデルのパラメータを CPC で再学習したモデルを比較した．[3] 再学習にはデフォルトのパラメータを利用した．

実験結果を表 5 に示す．評価指標は全文節の係り受け正解率 (level0)，最後の文節以外の正解率 (level1)，最後および最後から 2 番目以外の文節の正解率 (level2)，文全体での正解率 (sentence) である．末尾の文節，およびその前の文節は係り先の推定が容易であるため，それらを除外した level1 および level2 の正解率は level0 と比較すると低いことが確認できる．すべての評価方法において， CPC で再学習したモデルの正解率が学習済みモデルの正解率を上回っており， CPC の有用性が確認できた．なお，解析誤りを調査したが形態素解析や固有表現抽出のようなレシピ特有の誤りは見つからなかった．

5 分野固有のアノテーションミス

CPC のアノテーション結果を見直したところ，少ないながら表 6 のようなアノテーション誤りを発見した．誤りは大きく分けると文節境界に関わるものと文境界に関わるものの 2 種類が存在した．

1 つ目の例は「長ネギを並べてワインコンソメを入れて」という文を「長ネギを / 並べて / ワインコンソメを / 入れて」のような文節に分割している．しかしながら，レシピの材料欄を確認したところ「ワイン」「コンソメ」は独立した材料であり，実際には「長ネギを / 並べて / ワイン / コンソメを / 入れて」のような文節に分割しなければならなかった．

2 つ目の例は文境界に関する誤りである．「電子レンジ (600w 設定) で焼く [EOS] 4〜5 分ほど加熱処理をしておく」と 2 つの文に分割されているが，「焼く」は「約」の誤りであり，正しくは「電子レンジ (600w 設定) で焼く (約) 4〜5 分ほど加熱処理をしておく」となるべきであった．調理手順で実際に食材を加熱しており「焼く」と誤字があっても意味が成立してしまっていた．また，後続の文も単体で成立しており，発見しにくい誤りであった．

現在公開されている CPC (version 1.0) ではこれらのアノテーション誤りはそのまま残されているが，今後のバージョンアップ時に修正する予定である．

6 結論

本研究では CPC を用いて実施したレシピ解析について，実験の詳細を述べ，現状と課題を報告した．CPC を用いた解析器の再学習によって各タスクの性能が改善し， CPC の有用性が確認された．

また，各タスクの実験結果のエラー分析を実施し，形態素解析，固有表現抽出でレシピ固有の解析誤りを発見した．さらに，料理の知識，およびレシピ特有のひらがな表記や略語に関する知識が必要となるアノテーションが難しい事例を発見した．今後もこれらの事例を活用し， CPC の拡充およびレシピ解析の改善に取り組む予定である．

3) Harashima ら [11] の実験では再学習を実施せずに直接 CPC でモデルを訓練した．これに対し，本実験では形態素解析の実験条件に合わせて再学習を実施した．

参考文献

[1]Dan Tasse and Noah A. Smith. SOUR CREAM: Toward Semantic Processing of Recipes. Technical report, Carnegie Mellon University, 2008.

[2]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 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2455–2459, 2016.

[3]Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning Cross-modal Embeddings for Cooking Recipes and Food Images. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pp. 3020–3028, 2017.

[4]Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings. In *Proceedings of the 41th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018)*, pp. 35–44, 2018.

[5]Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. Globally Coherent Text Generation with Neural Checklist Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pp. 329–339, 2016.

[6]Amaia Salvador, Michal Drozdzal, Xavier Giro i Nieto, and Adriana Romero. Inverse Cooking: Recipe Generation from Food Images. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, 2019.

[7]Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 1358–1368, 2018.

[8]Jun Harashima, Makoto Hiramatsu, and Satoshi Sanjo. Calorie Estimation in a Real-World Recipe Service. In *Proceedings of the 32nd Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-20)*, pp. 13306–13313, 2020.

[9]Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hashida. Construction of a Japanese Relevance-tagged Corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 2008–2013, 2002.

[10]Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop (LAW 2007)*, pp. 132–139, 2007.

[11]Jun Harashima and Makoto Hiramatsu. Cookpad Parsed Corpus: Linguistic Annotations of Japanese Recipes. In *Proceedings of the 14th Linguistic Annotation Workshop (LAW 2020)*, pp. 87–92, 2020.

[12]Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. Flow Graph Corpus from Recipe Texts. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2370–2377, 2014.

[13]Yoko Yamakata, Shinsuke Mori, and John Carroll. English recipe flow graph corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 5187–5194, 2020.

[14]Tetsuro Sasada, Shinsuke Mori, Tatsuya Kawahara, and Yoko Yamakata. Named Entity Recognizer Trainable from Partially Annotated Data. In *Proceedings of the 14th International Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, pp. 148–160, 2015.

[15]Taichi Nishimura, Suzushi Tomori, Hayato Hashimoto, Atsushi Hashimoto, Yoko Yamakata, Jun Harashima, Yoshitaka Ushiku, and Shinsuke Mori. Visual grounding annotation of recipe flow graph. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 4275–4284, 2020.

[16]Jun Harashima, Yuichiro Someya, and Yohei Kikuta. Cookpad Image Dataset: An Image Collection as Infrastructure for Food Research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pp. 1229–1232, 2017.

[17]Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pp. 230–237, 2004.

[18]Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002.

[19]PWNER. http://www.lsta.media.kyoto-u.ac.jp/resource/tool/PWNER/. Accessed: 2021-01-09.

[20]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)*, pp. 260–270, 2016.

[21]pyner. https::/github.com/himkt/pyner. Accessed: 2021-01-09.