日本語レシピデータセットの継続的な構築と複合的な利用

原島純、平松淳、深澤祐援、山口泰弘（クックパッド株式会社）

NLP2022 Workshop on Japanese Evaluation Dataset (JED2022)

背景

インターネットやスマートフォンの普及によりインターネット上のレシピが増加

・日本語だと 70 万レシピ（2010）→ 500 万レシピ（2020）*1

レシピに関する研究やデータセットも増加

・研究：言語理解 [Kiddon+ 15]、文書生成 [Kiddon+ 16]、情報検索 [Salvador+ 17]、質問応答 [Yagcioglu+ 18]、…

・データセット：Recipe1M+ [Marin+ 19]、RISeC [Jiang+ 20]、ARA [Donatelli+ 21]、…

研究にしろデータセットにしろ、メインはやはり英語（特にトップカンファレンス）→ 日本語も負けてられない！

---

*1 クックパッドと楽天レシピに投稿されたレシピの総数（発表者調べ）

目次

クックパッド

インターネット上でレシピの投稿・検索ができる日本最大*1のレシピサービス

・レシピ投稿数：365 万品
・国内月間利用者数：5,600 万人
・プレミアム会員数：183 万人
・展開国・地域数：74 カ国
・対応言語数：32 言語

*1 それぞれ 2021 年 12 月 31 日時点のデータ

レシピ

料理の材料や作り方を記述した文書

多くの場合、以下の要素で構成される

・タイトル
・作者のコメント
・作者の名前
・材料
・作り方
・調理後の写真（場合によっては動画）
・調理中の写真（場合によっては動画）
・…

なすと挽き肉の甘辛炒め

レシピを保存

茄子と挽き肉の相性ばっちり！簡単につくれます。

ryokatsuma

材料 (2人分)

なす	二本
挽き肉	150g
厚揚げ豆腐	半袋
☆ 醤油	大さじ2
☆ みりん	大さじ2
☆ 料理酒	大さじ2
片栗粉	少々

カロリー・塩分を計算

作り方

1

なすを薄く輪切りにします。

2

厚揚げを小さくサイコロ状に切ります。

3

うすく油をひいたフライパンでなすと厚揚げを炒めます。

4

十分に火が通ったら、挽き肉も加えます。

5
挽き肉にも火が通ったら、弱火にして☆を加える。

6
仕上げに水溶き片栗粉を加えてとろみをつけて完成。

Cookpad Dataset

クックパッド株式会社が継続的に構築・公開しているデータセット

・Cookpad Recipe Dataset（2015 年公開）

・Cookpad Image Dataset（2017 年公開）

・Cookpad Comparable Corpus（2017 年公開）

・Cookpad Parsed Corpus（2020 年公開）

Cookpad Recipe Dataset

2014 年 9 末までに投稿された約 172 万レシピのテキスト（**タイトル、作者のコメント、材料、作り方、…**）を収録 [Harashima+ 16]

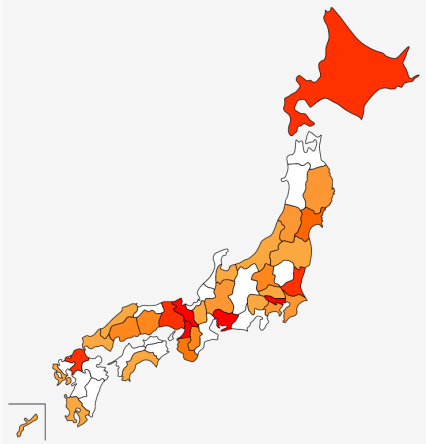
一部のレシピには**カテゴリや献立の情報**もある（逆に言うと、全てのレシピにはない）

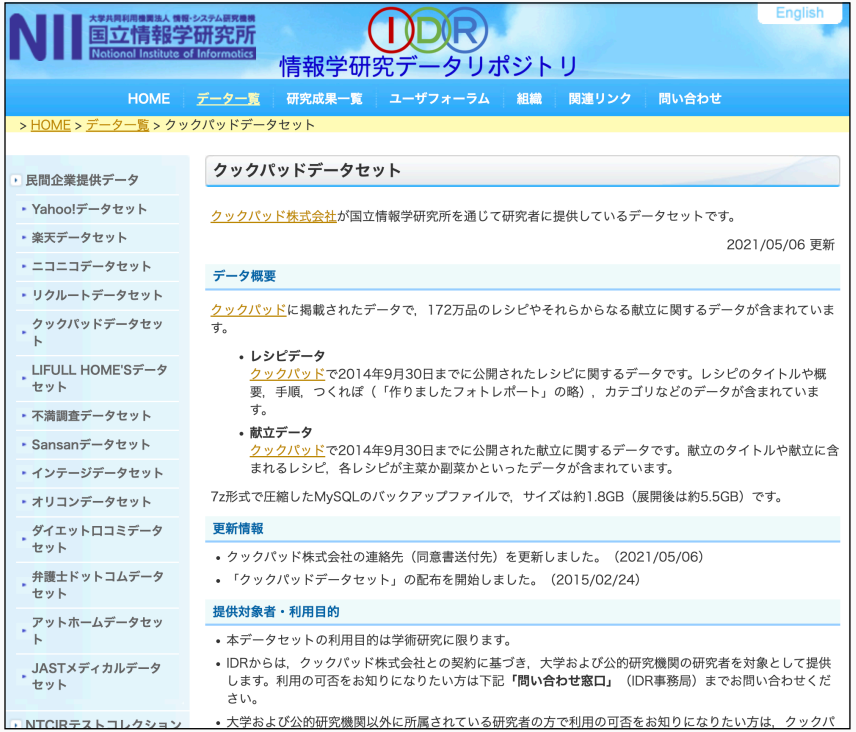
2015 年に公開、レシピ関連のテキストデータセットとしては世界最大

Cookpad Recipe Dataset

後述する他のデータセットと違い、NII*1 経由で公開

2022 年 3 月時点で全国 110 大学 212 研究室*2が利用





*1 https://www.nii.ac.jp/dsc/idr/cookpad/
*2 NLP 以外の研究室も多数

Cookpad Image Dataset

Recipe Dataset と同じ 172 万レシピの画像（**調理後の写真、調理中の写真**）を収録 [Harashima+ 17]

2017 年に公開、レシピ関連の画像データセットとしては世界最大

Cookpad Image Dataset

Table 1: Statistics and features of existing datasets and our dataset.

	year	# of complete images	# of process images	notable features
PFID [4]	2009	4, 545	N/A	fast-food
Rakuten Data [17]	2010	approx. 800, 000	N/A	linkable recipe texts
UEC FOOD 100 [15]	2012	9, 060	N/A	100 categories, bounding boxes
Chen's dataset [5]	2012	5, 000	N/A	50 categories
UEC FOOD 256 [12]	2014	31, 397	N/A	256 categories, bounding boxes
UNICT-FD889 [8]	2014	3, 583	N/A	889 categories
Food-101 [2]	2014	101, 000	N/A	101 categories
Menu-Match [1]	2015	646	N/A	restaurant food, 41 categories, calorie counts
UNIMIB2015 [6]	2015	2, 000	N/A	15 categories
UNIMIB2016 [7]	2016	1, 027	N/A	73 categories
VIREO Food-172 [3]	2016	110, 241	N/A	172 categories, 353 ingredients
Cookpad Image Dataset	2017	1, 642, 450	3, 105, 594	large collection of complete and process images, linkable recipe texts

調理後の写真数で世界最大

調理中の写真数でも世界最大

Recipe Dataset と紐付け可能

Cookpad Comparable Corpus

16,000 レシピに対する翻訳データ（日→英）を収録

・過去に開発していたサービス（クローズ済み）で使用していたデータ

翻訳プロセス

・1. 日本語ネイティブ 1 名*1*2 が翻訳

・2. 英語ネイティブ 2 名*2 が修正

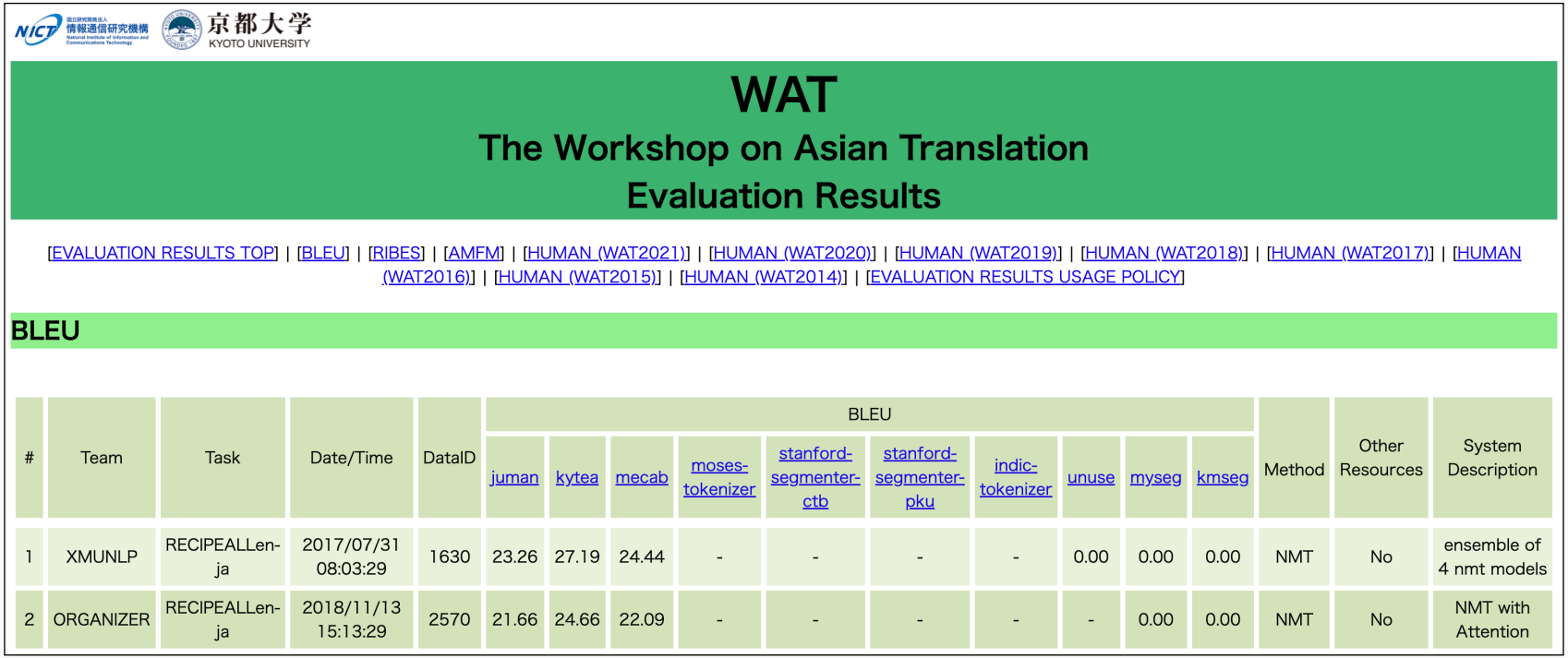
WAT 2017 と 2018*3 の subtask として提供

```
ja: {
  title: 卵豆腐のすまし汁,
  ingredients: [
    卵豆腐,
    …
  ],
  steps: [
    たけのこは上のやわらかい部分だけを薄く切る。,
    …
  ],
},
en: {
  title: Clear Broth with Egg Tofu,
  ingredients: [
    Egg tofu,
    …
  ],
  steps: [
    Take the soft part of the top of the bamboo shoot and thinly slice.,
    …
  ],
}
```

簡単のため、実際のデータを改変・省略

*1 英語に精通している人を採用
*2 料理に精通している人を採用
*3 http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT{2017,2018}/index.html

Cookpad Comparable Corpus



NICT 国立研究開発法人 情報通信研究機構 National Institute of Information and Communications Technology
京都大学 KYOTO UNIVERSITY

WAT

The Workshop on Asian Translation

Evaluation Results

[EVALUATION RESULTS TOP] | [BLEU] | [RIBES] | [AMFM] | [HUMAN (WAT2021)] | [HUMAN (WAT2020)] | [HUMAN (WAT2019)] | [HUMAN (WAT2018)] | [HUMAN (WAT2017)] | [HUMAN (WAT2016)] | [HUMAN (WAT2015)] | [HUMAN (WAT2014)] | [EVALUATION RESULTS USAGE POLICY]

BLEU

#	Team	Task	Date/Time	DataID	BLEU										Method	Other Resources	System Description
					juman	kytea	mecab	moses-tokenizer	stanford-segmenter-ctb	stanford-segmenter-pku	indic-tokenizer	unuse	myseg	kmseg			
1	XMUNLP	RECIPEALLen-ja	2017/07/31 08:03:29	1630	23.26	27.19	24.44	-	-	-	-	0.00	0.00	0.00	NMT	No	ensemble of 4 nmt models
2	ORGANIZER	RECIPEALLen-ja	2018/11/13 15:13:29	2570	21.66	24.66	22.09	-	-	-	-	-	0.00	0.00	NMT	No	NMT with Attention

ベンチマークの結果や実験用のスクリプトが閲覧・取得可能

Cookpad Parsed Corpus

500 レシピ（タイトルと作り方）に対する形態素解析と構文解析、固有表現認識の正解データを収録 [Harashima&Hiramatsu 20]

・形態素解析：MeCab（ipadic）の結果を人手で修正

・構文解析：CaboCha の結果を人手で修正

・固有表現認識：独自の 17 タグを人手で付与

企業による日本語解析済みコーパスの公開は初？

Name	Target documents
Kyoto University Text Corpus (Kawahara et al., 2002)	Newspaper articles
GDA Corpus (Hashida, 2005)	Newspaper articles and dictionary entries
NAIST Text Corpus (Iida et al., 2007)	Newspaper articles
Kyoto University and NTT Blog Corpus (Hashimoto et al., 2011)	Blogs
Kyoto University Web Document Leads Corpus (Hangyo et al., 2012)	Web documents
Balanced Corpus of Contemporary Written Japanese (Maekawa et al., 2014)	Newspaper articles, books, magazines, etc
Cookpad Parsed Corpus	**Cooking recipes**

Table 2: Existing Japanese parsed corpora and our corpus.

```
# 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
```

Cookpad Parsed Corpus

新聞記事の解析と比べると…

・形態素解析は難しい（未知語が多いため）

・構文解析は易しい（文が短いため）

・固有表現認識は不明（同じタグが付いてないため）

形態素解析器（MeCab）の性能*1

	再学習	適合率	再現率	F 値
単語分割のみ	なし	94,82	95,18	95,00
	あり	95,69	95,84	95,77
単語分割＋品詞タグ付け	なし	88,69	89,02	88,85
	あり	90,91	91,06	90,98

構文解析器（CaboCha）の性能*1

再学習	正解率	
	文節単位	文単位
なし	91,49	70,36
あり	94,20	78,04

固有表現認識器の性能*1

	正解率	適合率	再現率	F 値
[Sasada + 15]	87,48	73,61	81,37	77,30
[Lample+ 16]	90,13	85,95	85,56	85,75

*1 実験用のスクリプトは https://github.com/cookpad/cpc1.0 で公開

たべみる（ついでに紹介）

クックパッドの検索データを蓄積、法人向けに
展開している分析ツール

2016 年に公開

・データセットとして公開しているわけではなく
　アカウントを無償で提供（研究者のみ）

目次

日本語レシピデータセットの継続的な構築

日本語レシピデータセットの複合的な利用

まとめと今後の展望

複合的な利用？

各データセットは個別に利用可能（当たり前）

一方、複合的に利用することで初めて取り組めるタスクや手法も

個別の利用

Recipe Dataset

・文書推薦（主菜推薦・副菜推薦）
・文書生成（タイトル・作り方生成）
・キーワード推薦（材料推薦）
・…

Image Dataset

・超解像
・…

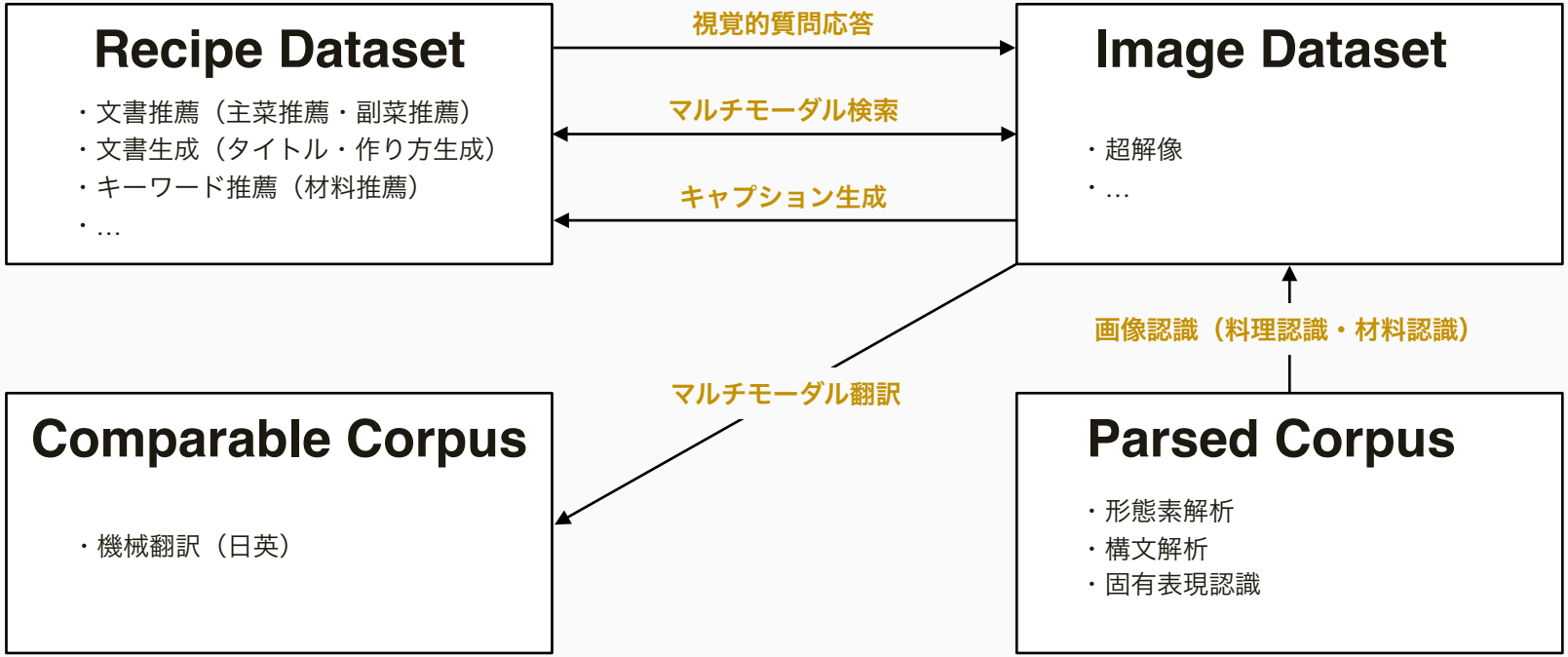
Comparable Corpus

・機械翻訳（日英）

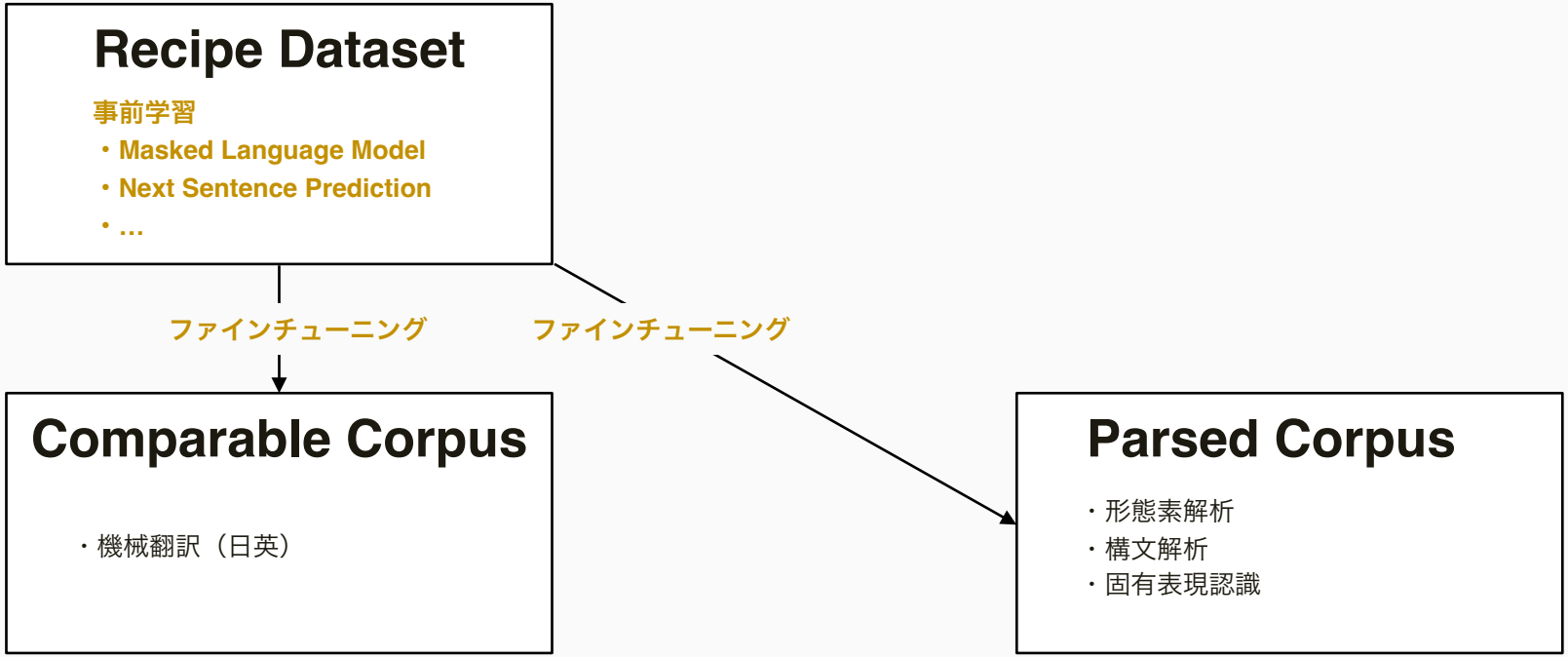
Parsed Corpus

・形態素解析
・構文解析
・固有表現認識

複合的な利用

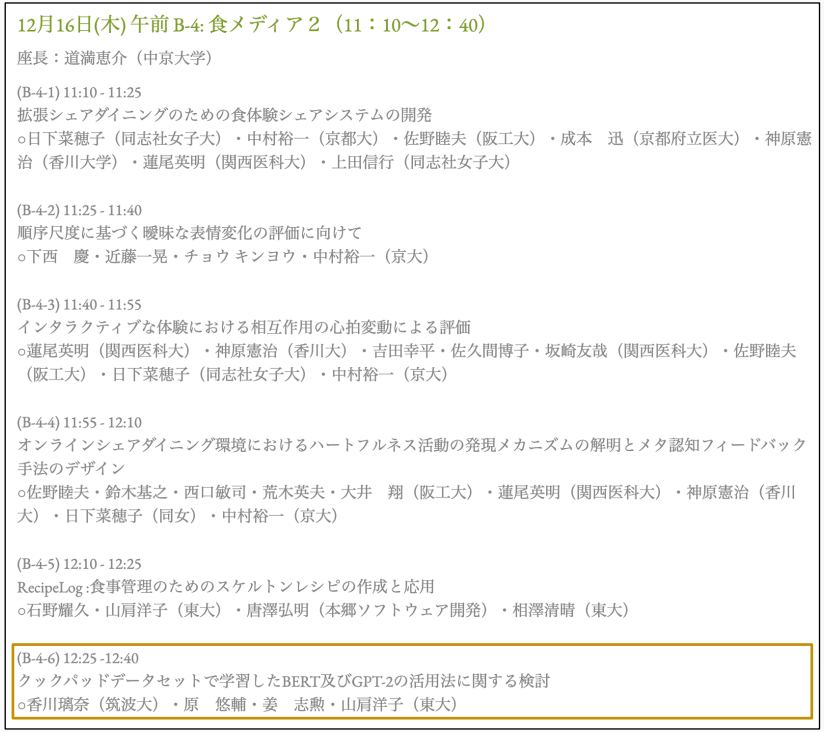

Recipe Dataset
・文書推薦（主菜推薦・副菜推薦）
・文書生成（タイトル・作り方生成）
・キーワード推薦（材料推薦）
・…
視覚的質問応答
マルチモーダル検索
キャプション生成
Image Dataset
・超解像
・…
画像認識（料理認識・材料認識）
マルチモーダル翻訳
Comparable Corpus
・機械翻訳（日英）
Parsed Corpus
・形態素解析
・構文解析
・固有表現認識

複合的な利用（手法の観点）



事前学習モデルの構築

既に取り組みはじめてくださっている方も



12月16日(木) 午前 B-4: 食メディア２（11：10〜12：40）

座長：道満恵介（中京大学）

(B-4-1) 11:10 - 11:25
拡張シェアダイニングのための食体験シェアシステムの開発
○日下菜穂子（同志社女子大）・中村裕一（京都大）・佐野睦夫（阪工大）・成本　迅（京都府立医大）・神原憲治（香川大学）・蓮尾英明（関西医科大）・上田信行（同志社女子大）

(B-4-2) 11:25 - 11:40
順序尺度に基づく曖昧な表情変化の評価に向けて
○下西　慶・近藤一晃・チョウ キンヨウ・中村裕一（京大）

(B-4-3) 11:40 - 11:55
インタラクティブな体験における相互作用の心拍変動による評価
○蓮尾英明（関西医科大）・神原憲治（香川大）・吉田幸平・佐久間博子・坂崎友哉（関西医科大）・佐野睦夫（阪工大）・日下菜穂子（同志社女子大）・中村裕一（京大）

(B-4-4) 11:55 - 12:10
オンラインシェアダイニング環境におけるハートフルネス活動の発現メカニズムの解明とメタ認知フィードバック手法のデザイン
○佐野睦夫・鈴木基之・西口敏司・荒木英夫・大井　翔（阪工大）・蓮尾英明（関西医科大）・神原憲治（香川大）・日下菜穂子（同女）・中村裕一（京大）

(B-4-5) 12:10 - 12:25
RecipeLog :食事管理のためのスケルトンレシピの作成と応用
○石野耀久・山肩洋子（東大）・唐澤弘明（本郷ソフトウェア開発）・相澤清晴（東大）

(B-4-6) 12:25 -12:40
クックパッドデータセットで学習したBERT及びGPT-2の活用法に関する検討
○香川璃奈（筑波大）・原　悠輔・姜　志勲・山肩洋子（東大）

さらなる併用も？

・楽天データセット

・フローグラフコーパス [Mori+ 14]

・料理オントロジー [Nanba+ 14]

・基本料理知識ベース [清丸+ 18]

・r-FG-BB データセット [Nishimura+ 20]

・…

いずれもレシピや料理に関する
日本語のデータセット

目次

まとめ

日本語レシピデータセットの継続的な構築

・Cookpad Recipe Dataset（2015 年公開）

・Cookpad Image Dataset（2017 年公開）

・Cookpad Comparable Corpus（2017 年公開）

・Cookpad Parsed Corpus（2020 年公開）

日本語レシピデータセットの複合的な利用

・タスク：視覚的質問応答、マルチモーダル検索、キャプション生成、…

・手法：事前学習＋ファインチューニング

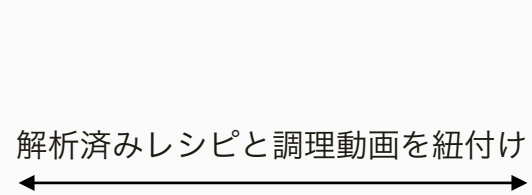
今後の展望

Cookpad Video Dataset with OMRON SINIC X 鋭意開発中！

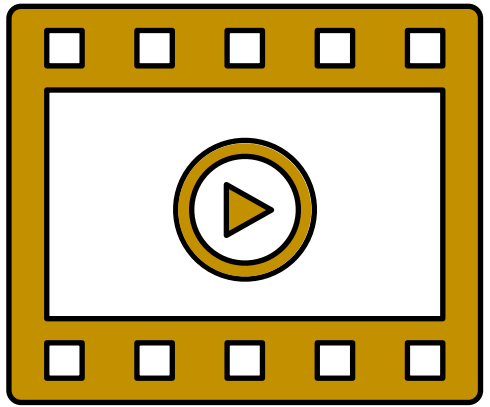
Parsed Corpus

```
# 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
　　⋮
```

解析済みレシピと調理動画を紐付け



Video Dataset



参考文献

- [Donatelli+ 21] Aligning Actions Across Recipe Graphs
- [Harashima+ 16] A Large-Scale Recipe and Meal Data Collection as Infrastructure for Food Research
- [Harashima+ 17] Cookpad Image Dataset: An Image Collection as Infrastructure for Food Research
- [Harashima&Hiramatsu 20] Cookpad Parsed Corpus: Linguistic Annotations of Japanese Recipes
- [Jiang+ 20] Recipe Instruction Semantics Corpus (RISeC): Resolving Semantic Structure and Zero Anaphora in Recipes
- [Kiddon+ 15] Mise en Place: Unsupervised Interpretation of Instructional Recipes
- [Kiddon+ 16] Globally Coherent Text Generation with Neural Checklist Models
- [Lample+ 16] Neural Architectures for Named Entity Recognition
- [Marin+ 19] Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images
- [Mori+ 14] Flow Graph Corpus from Recipe Texts
- [Nanba+ 14] Construction of a Cooking Ontology from Cooking Recipes and Patents
- [Nishimura+ 20] Visual Grounding Annotation of Recipe Flow Graph
- [Salvador+ 17] Learning Cross-modal Embeddings for Cooking Recipes and Food Images
- [Sasada+ 15] Named Entity Recognizer Trainable from Partially Annotated Data
- [Yagcioglu+ 18] RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes
- [香川+ 21] クックパッドデータセットで学習した BERT 及び GPT-2 の活用法
- [清丸+ 18] 料理レシピとクラウドソーシングに基づく基本料理知識ベースの構築