A Large-scale Recipe and Meal Data Collection as Infrastructure for Food Research

Jun Harashima, Michiaki Ariga, Kenta Murata, Masayuki Ioki
Cookpad Inc.
Yebisu Garden Place Tower 12F, 4-20-3 Ebisu, Shibuya-ku, Tokyo, 150-6012, Japan
{jun-harashima, michiaki-ariga, kenta-murata, masayuki-ioki}@cookpad.com

Abstract

Everyday meals are an important part of our daily lives and, currently, there are many Internet sites that help us plan these meals. Allied to the growth in the amount of food data such as recipes available on the Internet is an increase in the number of studies on these data, such as recipe analysis and recipe search. However, there are few publicly available resources for food research; those that do exist do not include a wide range of food data or any meal data (that is, likely combinations of recipes). In this study, we construct a large-scale recipe and meal data collection as the underlying infrastructure to promote food research. Our corpus consists of approximately 1.7 million recipes and 36000 meals in cookpad, one of the largest recipe sites in the world. We made the corpus available to researchers in February 2015 and as of February 2016, 82 research groups at 56 universities have made use of it to enhance their studies.

Keywords: recipe, meal, food research

1. Introduction

Nowadays many people plan meals using the Internet. Because meals need to be planned several times a day, this is a time-consuming and labor-intensive task. However, owing to the many Internet sites available to help users find recipes and plan their meals (e.g., cookpad in Japanese,[1] and Yummly in English[2]), such planning has become much easier.

Alongside the increase in the amount of available food data such as recipes, there has been a corresponding increase in the number of studies on these data such as recipe analysis (Kiddon et al., 2015; Jermsurawong and Habash, 2015; Maeta et al., 2015; Sasada et al., 2015) and recipe search (Yasukawa et al., 2014; Wang et al., 2008). Furthermore, many conferences and workshops have recently been held to promote research on these topics.

Nevertheless, there are few publicly available resources for food research. Additionally, such resources are not widely applicable as they do not include a wide range of food data (a few dozen – hundreds of thousands of recipes) or any meal data (data on likely combinations of recipes). Thus, a novel resource is required to promote food research.

In this study, we constructed a large-scale recipe and meal data collection as the underlying infrastructure for food research. We organized approximately 1.7 million recipes, 36000 meals, and their related data (e.g., reviews) in cookpad and made the collected data available to researchers. In this paper, we report on how we designed our corpus and how this has contributed to enhancing food research.

The remainder of this paper is organized as follows. Section 2. briefly describes relevant previous studies. Sections 3. and 4. discuss, respectively, the recipe and meal data in our corpus. Section 5. summarizes the contributions of this study while our conclusions are presented in Section 6..

2. Related Work

2.1. Large-scale Raw Data Collections

As described in the previous section, few resources exist that include a wide range of food data. To the best of our knowledge, the following corpora are the only two that include a reasonable amount of food data. Rakuten Data[3] is a publicly available resource that includes approximately 440000 recipes written in Japanese, while approximately 100000 recipes in English extracted from Yummly were published in the 11th NTCIR Workshop.[4]

We have constructed a novel corpus that is superior to these resources in terms of the following aspects. The corpus includes many more recipes than either of these existing resources, while also providing meal data.

2.2. Small-scale Annotated Data Collections

Several resources with small amounts of data include manually annotated data. In the flow graph corpus (Mori et al., 2014), 266 Japanese recipes from cookpad are represented as graph structures, composed of vertices (named entities) and arcs (representing relations between two entities). The Kyoto University Smart Kitchen Dataset selected 20 recipes from Mori's corpus and linked cooking processes in each recipe to real human activities in a kitchen (Hashimoto et al., 2014). The Carnegie Mellon University Recipe Database (CMU Recipe Database) consists of 260 English recipes translated into a bespoke machine-readable language (Tasse and Smith, 2008).

Some recipes and meals in our corpus have been manually tagged with their categories (see the following sections). Therefore, the corpus is similar to the above resources in that all of these include annotated data. However, our corpus includes a vast amount of recipe and meal data and thus, differs significantly from existing resources in terms of size.

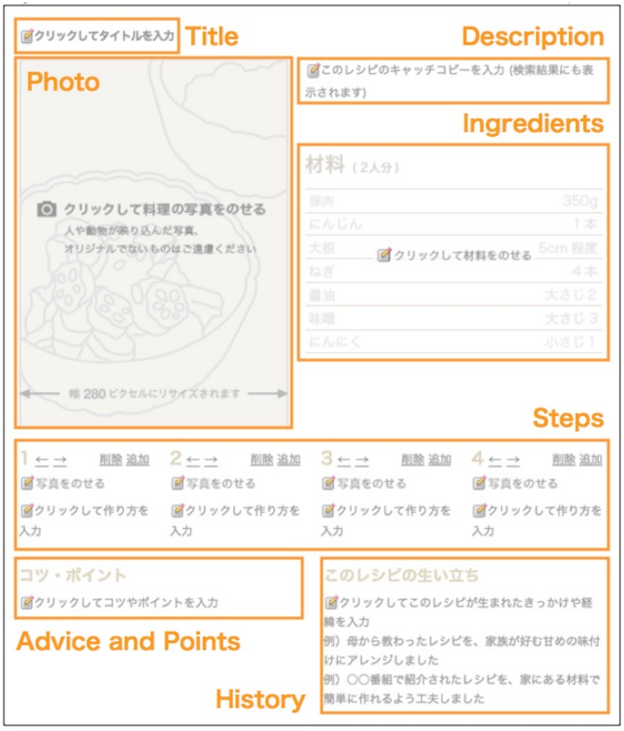
[1] cookpad.com
[2] www.yummly.com
[3] http://rit.rakuten.co.jp/opendata.html
[4] http://research.nii.ac.jp/ntcir/ntcir-11/index.html

2.3. Existing Studies on Food Data

First, we introduce relevant studies focusing on the analysis of recipes. Kiddon et al. (2015) proposed an unsupervised method for extracting and identifying latent connections among actions in a recipe. Based on the expectation-maximization approach, the method defines what actions should be performed on which objects (e.g., ingredients, tools) and in what order. Jermsurawong and Habash (2015) represented a recipe as a dependency tree with the leaves depicting ingredients in the recipe and internal nodes denoting instructions contained therein. They also built a parser that maps a recipe into their proposed tree structure and showed the accuracy of their parser using the CMU Recipe Database. Maeta et al. (2015) proposed a framework for analyzing procedural texts, especially for recipes. Their method tokenizes a recipe, recognizes concepts (e.g., food, tools) like named entity recognition, and then connects these in a graph. Sasada et al. (2015) proposed a named entity recognizer that can be trained from partially annotated data. They focused on the fact that fully annotated data is rarely available in food research and constructed the recognizer to overcome this problem. Nanba et al. (2014) used both manual and automatic means to construct a cooking ontology for recipe analysis. Their resource comprises 474 entry words, 5023 synonyms, 1512 attributes, and 2429 meronymic words. Tachibana et al. (2014) investigated identifying a naming concept from a recipe. As the concept, they extracted differences between elements (e.g., ingredients, tools) of the recipe and typical elements of other recipes for the same dish.

Studies that focus on applications for recipes, such as recipe searches, recipe recommendations, and recipe summarizations, are discussed below. Yasukawa et al. (2014) implemented a search task, called RecipeSearch,[5] to study information access for recipe data in the 11th NTCIR Workshop. To implement the task, they used Rakuten and Yummly data, as described in the previous section, and four research groups participated in the task. Yamakata et al. (2013) focused on extracting a general way of cooking as a summary of cooking procedures. Their approach converts recipes into graph structures and finds common structures in these to form the summary. Forbes and Zhu (2011) proposed a recommendation method for recipes based on matrix factorization. They incorporated content information in a recipe (i.e., ingredients) into their approach and confirmed the usefulness of the method through various experiments. Wang et al. (2008) constructed a similarity search system for Chinese recipes. The system translates recipes into directed graphs and for a given query, displays a recipe together with other similar ones based on the graph structures.

All of the above studies use some food data. However, researchers often need to implement a crawler themselves to obtain the data from recipe sites. We have made our corpus available to researchers with the aim of facilitating the study of food data. Additionally, by creating a common infrastructure for this field, our corpus should assist researchers in discussing their studies with others.



クリックしてタイトルを入力

このレシピのキャッチコピーを入力 (検索結果にも表示されます)

クリックして料理の写真をのせる
人や動物が映り込んだ写真、
オリジナルでないものはご遠慮ください

幅 280 ピクセルにリサイズされます

材料（2人分）

クリックして材料をのせる

豚肉	350g
にんじん	1本
大根	5cm 程度
ねぎ	4本
醤油	大さじ2
味噌	大さじ3
にんにく	小さじ1

1 ← → 削除 追加
写真をのせる
クリックして作り方を入力

2 ← → 削除 追加
写真をのせる
クリックして作り方を入力

3 ← → 削除 追加
写真をのせる
クリックして作り方を入力

4 ← → 削除 追加
写真をのせる
クリックして作り方を入力

コツ・ポイント
クリックしてコツやポイントを入力

このレシピの生い立ち
クリックしてこのレシピが生まれたきっかけや経緯を入力
例）母から教わったレシピを、家族が好む甘めの味付けにアレンジしました
例）〇〇番組で紹介されたレシピを、家にある材料で簡単に作れるよう工夫しました

Figure 1: A recipe template in cookpad.

3. Recipe Data

As at February 2016, the cookpad comprised over 2.2 million recipes written in Japanese, making it one of the largest recipe sites in the world. On this site, users can upload their recipes using the template shown in Figure 1.

We organized the recipes and their related data in cookpad to help researchers use them. First, we collected approximately 1.7 million recipes that had been uploaded to cookpad by September 2014. Figure 2 gives an example of recipes in cookpad with the main information provided by us, demarcated by orange rectangles. In this study, we included text information, but no image information (i.e., a photo for each recipe); this has been left for our future work. Details of the text information fields added are given below.

A. Title

The title, consisting of a maximum of 20 characters, gives a concise summary of the recipe. For example, the title in Figure 2 is 豚のにんにく醤油焼き (pork with garlic soy sauce).

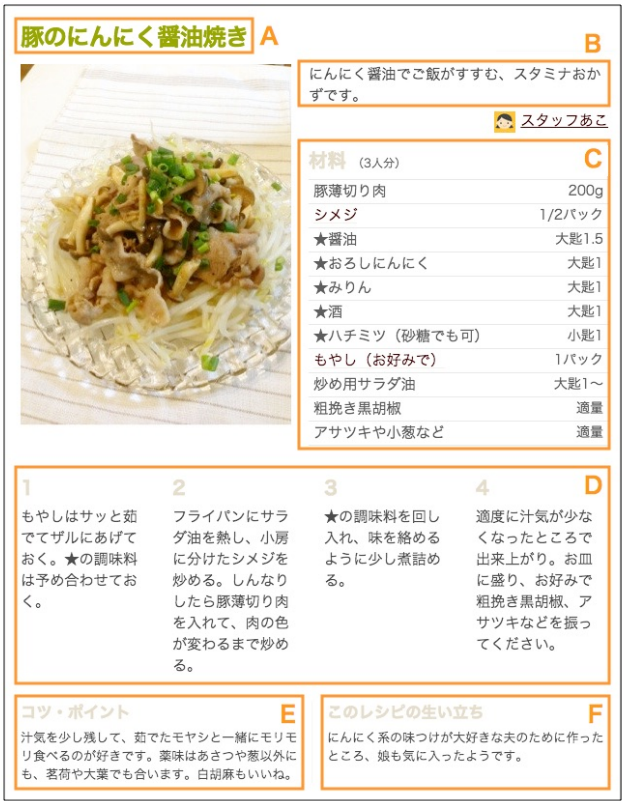
B. Description

Using this data field, authors can provide an eye-catching description of their recipes (e.g., “this goes well with rice” as shown in Figure 2). This text is also shown as a snippet in the recipe search results in cookpad.

C. Ingredients

Ingredients used in the recipe (e.g., pork in Figure 2) are listed in this field together with the quantities thereof. This data field also describes how many servings the recipe provides.

[5] https://sites.google.com/site/ntcir11recipesearch/



豚のにんにく醤油焼き

にんにく醤油でご飯がすすむ、スタミナおかずです。

スタッフあこ

材料 （3人分）

豚薄切り肉	200g
シメジ	1/2パック
★醤油	大匙1.5
★おろしにんにく	大匙1
★みりん	大匙1
★酒	大匙1
★ハチミツ（砂糖でも可）	小匙1
もやし（お好みで）	1パック
炒め用サラダ油	大匙1～
粗挽き黒胡椒	適量
アサツキや小葱など	適量

1 もやしはサッと茹でてザルにあげておく。★の調味料は予め合わせておく。

2 フライパンにサラダ油を熱し、小房に分けたシメジを炒める。しんなりしたら豚薄切り肉を入れて、肉の色が変わるまで炒める。

3 ★の調味料を回し入れ、味を絡めるように少し煮詰める。

4 適度に汁気が少なくなったところで出来上がり。お皿に盛り、お好みで粗挽き黒胡椒、アサツキなどを振ってください。

コツ・ポイント

汁気を少し残して、茹でたモヤシと一緒にモリモリ食べるのが好きです。薬味はあさつや葱以外にも、茗荷や大葉でも合います。白胡麻もいいね。

このレシピの生い立ち

にんにく系の味つけが大好きな夫のために作ったところ、娘も気に入ったようです。

Figure 2: An example of a recipe from cookpad.

D. Steps

This field describes the method of cooking. A recipe usually consists of multiple steps (e.g., four steps in Figure 2) all of which are explained in detail in this data field.

E. Advice and Points to Note

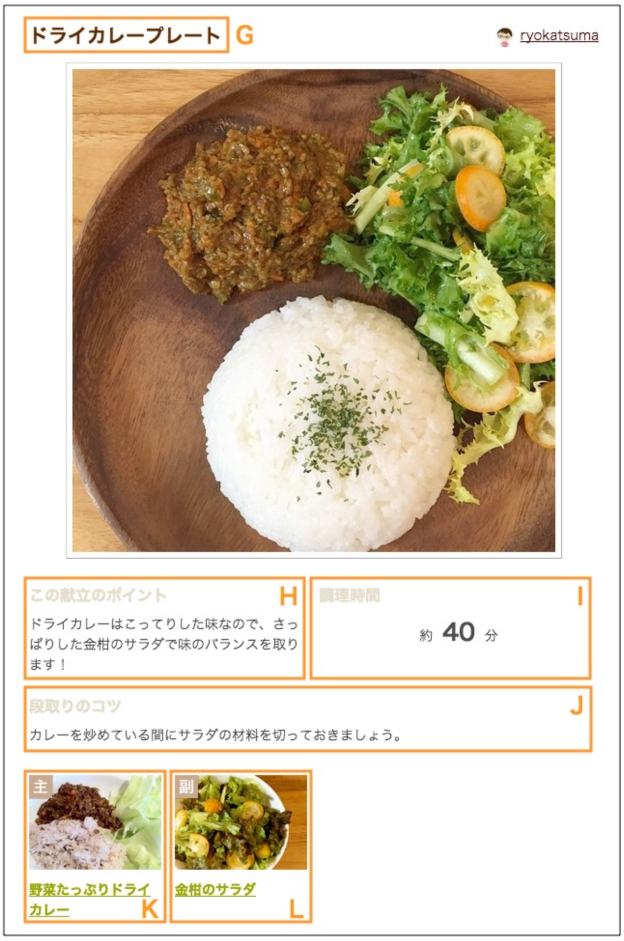
In this field, authors can provide additional hints about their recipe to help readers prepare the dish. An author suggests that Japanese ginger and green perilla can be used instead of chives in the recipe depicted in Figure 2.

F. History

This field explains how and why the recipe was created. For example, the history field in Figure 2 states that the recipe was devised for the author's husband, who likes garlic.

Although Figure 2 only shows the main information for the recipe, our corpus also includes the following data for each recipe: the unique ID of the recipe, the unique ID of the author, and the date when the recipe was uploaded.

In addition, we provide researchers with related data for each recipe. In cookpad, more than 146000 recipes have been classified into approximately 1100 categories by users (e.g., meat dishes, seafood dishes, vegetable dishes). Moreover, there are almost 10 million reviews of recipes (called "tsukurepo") in cookpad, which are also included in our corpus.



ドライカレープレート

ryokatsuma

この献立のポイント

ドライカレーはこってりした味なので、さっぱりした金柑のサラダで味のバランスを取ります！

調理時間

約 40 分

段取りのコツ

カレーを炒めている間にサラダの材料を切っておきましょう。

野菜たっぷりドライカレー

金柑のサラダ

Figure 3: An example of a meal from cookpad.

4. Meal Data

Unlike the existing resources, our corpus includes not only recipes but also meals (that is, likely combination of recipes). In addition to the recipes, we collected approximately 36000 meals in cookpad that had been uploaded by September 2014. Similar to recipes, meals are uploaded using the template provided by cookpad. Figure 3 shows an example of a meal, with the main information for the meal provided by us, explained below.

G. Title

The title, consisting of a maximum of 20 characters, gives a concise summary of the meal. For example, the title in Figure 3 is ドライカレープレート (curried pilaf).

H. Noteworthy Points

Similar to the advice field in recipes, this field explains various aspects of the respective meal to help readers prepare it. In Figure 3, the author states that a salad was added as a side dish to balance the flavors.

I. Cooking Time

This field states how long it took the author to prepare the meal (e.g., 40 min in Figure 3). When an author submits his/her meal, the total preparation time in minutes must be selected from the given options.

	language	# of recipes	# of meals
Our data	Japanese	approx. 1715000	approx. 36000
Rakuten data	Japanese	approx. 440000	N/A
Yummly data	English	approx. 100000	N/A
(Mori et al., 2014)	Japanese	266	N/A
(Hashimoto et al., 2014)	Japanese	20	N/A
(Tasse and Smith, 2008)	English	260	N/A

Table 1: Statistics of our data and existing data.

J. Advice
A meal usually consists of multiple dishes, some of which may be difficult to prepare simultaneously. This field provides various hints about the meal to help readers prepare multiple dishes efficiently.

K. Main Dishes
Main dishes for the meal, of which there can be many, are listed in this field. The recipe data for the dishes are also included in our corpus (see the previous section).

L. Side Dishes
Side dishes for the meal are listed in this field. As in the case of main dishes, there can be multiple side dishes, the recipe data for which are also included in our corpus.

Our corpus also includes the following data for each meal: the unique ID of the meal, the unique ID of the author, and the date when the meal was uploaded.
Moreover, we provide researchers with related data for each meal. In cookpad, meals are classified into categories (e.g., Japanese style, Western style) and some meals are voted for by users. Such user-created data are also included in our corpus.

5. Data Release

We have made our corpus available in cooperation with the National Institute of Informatics (NII),[6] a Japanese Research Institute with the goal of advancing informatics research. The NII aggregates information on various kinds of datasets for informatics research. Our corpus, which is available from a site administered by the Institute,[7] is distributed as a MySQL dump, consisting of 12 tables: six for recipe data and six for meal data. Any researcher in public institutions such as universities can obtain access to the corpus if he/she wishes to use it for research purposes.
In Table 1, we summarize the statistics of our corpus and others described in Section 2.. From the table, it is clear that our corpus is the biggest in the world.
Between its release in February 2015 and February 2016, our corpus has already been acquired by 82 research groups at 56 universities. With the availability of our corpus, there is no need for individual researchers to implement crawlers themselves, as described in Section 2.. Additionally, the corpus makes it easier for researchers to compare their studies with those of others. As described above, our corpus forms the basic infrastructure for food research and we strongly believe that the corpus will enhance research in this field.

6. Conclusion

In this study, we constructed a novel language resource to promote food research. Our corpus includes more than 1.7 million Japanese recipes taken from cookpad. Unlike previous resources, it also includes data for approximately 36000 meals. Since making the corpus available for research purposes, researchers at 56 universities have used it in their studies.
In future, as described in Section 3., we plan to provide image information to promote image recognition research. Additionally, we plan to provide a comparable recipe corpus, consisting of Japanese recipes with their English translations, to contribute to machine translation research.

Acknowledgments

We gratefully acknowledge the support of the National Institute of Informatics.

7. Bibliographical References

Forbes, P. and Zhu, M. (2011). Content-boosted Matrix Factorization for Recommender Systems: Experiments with Recipe Recommendation. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys 2011)*, pages 261–264.

Hashimoto, A., Sasada, T., Yamakata, Y., Mori, S., and Minoh, M. (2014). KUSK Dataset: Toward a Direct Understanding of Recipe Text and Human Cooking Activity. In *Proceedings of the Workshop on Smart Technology for Cooking Eating Activities (CEA 2014)*, pages 583–588.

Jermsurawong, J. and Habash, N. (2015). Predicting the Structure of Cooking Recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 781–786.

Kiddon, C., Ponnuraj, G. T., Zettlemoyer, L., and Choi, Y. (2015). Mise en Place: Unsupervised Interpretation of Instructional Recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 982–992.

Maeta, H., Sasada, T., and Mori, S. (2015). A Framework for Procedural Text Understanding. In *Proceedings of the 14th International Conference on Parsing Technologies (IWPT 2015)*, pages 50–60.

Mori, S., Maeta, H., Yamakata, Y., and Sasada, T. (2014). Flow Graph Corpus from Recipe Texts. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2370–2377.

[6] http://www.nii.ac.jp/
[7] http://www.nii.ac.jp/dsc/idr/cookpad/cookpad.html

Nanba, H., Doi, Y., Tsujita, M., Takezawa, T., and Sumiya, K. (2014). Construction of a Cooking Ontology from Cooking Recipes and Patents. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication (UbiComp 2014 Adjunct)*, pages 507–516.

Sasada, T., Mori, S., Kawahara, T., and Yamakata, Y. (2015). Named Entity Recognizer Trainable from Partially Annotated Data. In *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, pages 10–17.

Tachibana, A., Wakamiya, S., Nanba, H., and Sumiya, K. (2014). Extraction of Naming Concepts based on Modifiers in Recipe Titles. In *Proceedings of the International MultiConference of Engineers and Computer Scientists 2014 (IMECS 2014)*, pages 507–512.

Tasse, D. and Smith, N. A. (2008). SOUR CREAM: Toward Semantic Processing of Recipes. Technical report, Carnegie Mellon University.

Wang, L., Li, Q., Li, N., Dong, G., and Yang, Y. (2008). Substructure Similarity Measurement in Chinese Recipes. In *Proceedings of the 17th International World Wide Web Conference (WWW 2008)*, pages 979–988.

Yamakata, Y., Imahori, S., Sugiyama, Y., Mori, S., and Tanaka, K. (2013). Feature Extraction and Summarization of Recipes using Flow Graph. In *Proceedings of the 5th International Conference on Social Informatics (SocInfo 2013)*, pages 241–254.

Yasukawa, M., Diaz, F., Druck, G., and Tsukada, N. (2014). Overview of the NTCIR-11 Cooking Recipe Search Task. In *Proceedings of the 11th NTCIR Conference (NTCIR-11)*, pages 483–496.