Cookpad Parsed Corpus: Linguistic Annotations of Japanese Recipes

Jun Harashima and **Makoto Hiramatsu**
Cookpad Inc.
{jun-harashima, himkt}@cookpad.com

Abstract

It has become increasingly common for people to share cooking recipes on the Internet. Along with the increase in the number of shared recipes, there have been corresponding increases in recipe-related studies and datasets. However, there are still few datasets that provide linguistic annotations for the recipe-related studies even though such annotations should form the basis of the studies. This paper introduces a novel recipe-related dataset, named Cookpad Parsed Corpus, which contains linguistic annotations for Japanese recipes. We randomly extracted 500 recipes from the largest recipe-related dataset, the Cookpad Recipe Dataset, and annotated 4,738 sentences in the recipes with morphemes, named entities, and dependency relations. This paper also reports benchmark results on our corpus for Japanese morphological analysis, named entity recognition, and dependency parsing. We show that there is still room for improvement in the analyses of recipes.

1 Introduction

Today, a great number of cooking recipes are available on the Internet. Many people upload their recipes to recipe-sharing services such as Cookpad and Yummly, with Cookpad having over six million recipes and Yummly over two million recipes, to date.

As the number of shared online recipes increases, many recipe-related datasets have been published (Salvador et al., 2017; Yagcioglu et al., 2018; Chandu et al., 2019; Lin et al., 2020). These datasets have successfully contributed their content to a variety of recipe-related studies about recipe understanding, recipe search, recipe generation, and so on.

Nevertheless, there are still few datasets that contain linguistic annotations for cooking recipes. Most of the recipe studies rely on linguistic analyses, like those focusing on other text such as newspaper articles. Since linguistic annotations play an important role in fundamental analyses, they are also deserving of more attention in this field.

In this paper, we introduce our Cookpad Parsed Corpus, which is a novel dataset of Japanese recipes. We extract 500 recipes randomly from the Cookpad Recipe Dataset (Harashima et al., 2016), currently the largest recipe dataset, and annotate 4,738 sentences in the recipes with the most fundamental linguistic information: morphemes, named entities, and dependency relations.

We also report benchmark results of the corpus for morphological analysis (MA), named entity recognition (NER), and dependency parsing (DP) for Japanese, and investigate whether tools or methods which have been commonly used for these analyses perform sufficiently for cooking recipes.

2 Related Works

Table 1 summarizes existing recipe-related datasets and our corpus. As shown in the table, each resource has recipes with their own content, such as graph representations and cooking images of the recipes. This

Both authors equally contributed to this work.

The 14th Linguistic Annotation Workshop, pages 87–92
Barcelona, Spain (Online), December 12, 2020.

Name	Main content (other than recipes)
Carnegie Mellon University Recipe Database (Tasse and Smith, 2008)	Machine-readable language representations
Flow Graph Corpus (Mori et al., 2014)	Graph representations and named entities
SIMMR Recipe Dataset (Jermsurawong and Habash, 2015)	Graph representations
Cookpad Recipe Dataset (Harashima et al., 2016)	Reviews and meals (combinations of recipes)
Cookpad Image Dataset (Harashima et al., 2017)	Food images and cooking images
Recipe1M (Salvador et al., 2017)	Food images
RecipeQA (Yagcioglu et al., 2018)	Question-answer pairs
Storyboarding Data (Chandu et al., 2019)	Cooking images
r-FG BB dataset (Nishimura et al., 2020)	Bounding boxes for cooking images
English Recipe Flow Graph Corpus (Yamakata et al., 2020)	Graph representations and named entities
Microsoft Research Multimodal Aligned Recipe Corpus (Lin et al., 2020)	URLs to YouTube videos
Multi-modal Recipe Structure dataset (Pan et al., 2020)	Graph representations and cooking images
Cookpad Parsed Corpus	**Linguistic annotations**

Table 1: Existing recipe-related datasets and our corpus.

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.

content has successfully promoted a variety of recipe-related studies about recipe understanding, recipe search, recipe generation, and so on.

Our corpus differs from these efforts in that it contains linguistic annotations of cooking recipes. In other words, the existing datasets have not taken account of the information, except for the Flow Graph Corpus (Mori et al., 2014) and English Recipe Flow Graph Corpus (Yamakata et al., 2020), which contain a few limited linguistic annotations such as named entities. By contrast, our corpus contains a variety of linguistic annotations such as morphemes, named entities, and dependency relations.

Table 2 summarizes existing Japanese parsed corpora and our corpus. There are several parsed corpora which have contributed their linguistic annotations to linguistic analyses such as MA, NER, and DP for Japanese. In particular, the Kyoto University Text Corpus (Kawahara et al., 2002) and NAIST Text Corpus (Iida et al., 2007) have been commonly used for such studies.

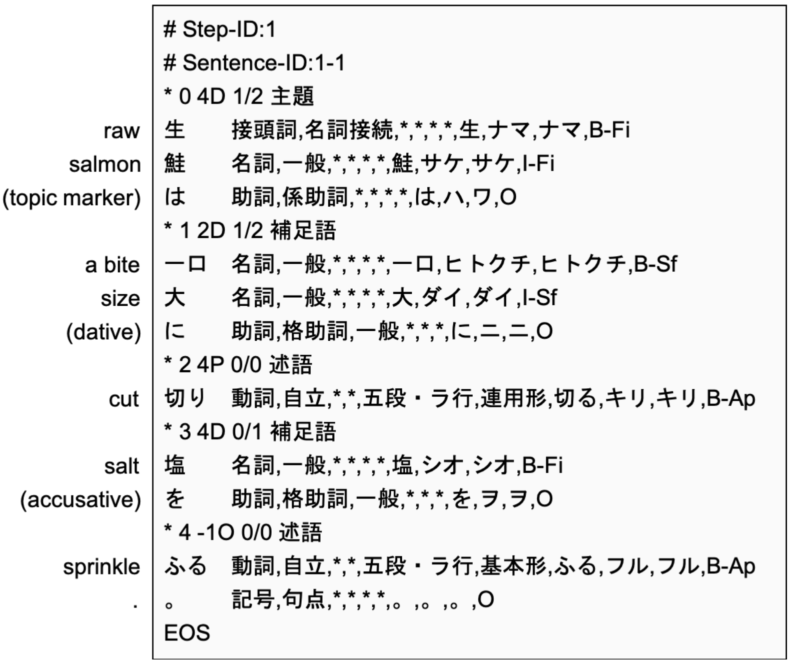
One of the biggest differences between these and our work is the target documents for annotations; that is, the other works focus on newspaper articles, dictionary entries, blogs, web documents, books, and magazines, whereas our corpus is the first to focus on cooking recipes.

3 Cookpad Parsed Corpus

In this study, we constructed a novel recipe-related dataset, named Cookpad Parsed Corpus, which contains linguistic annotations of Japanese recipes. We randomly selected 500 recipes from the largest recipe dataset, Cookpad Recipe Dataset (Harashima et al., 2016), which contains approximately 1.7 million Japanese recipes. We then annotated $4,738$ sentences (hereafter called the target sentences) in the 500 recipes with morphemes, named entities, and dependency relations.

Figure 1 shows linguistic annotations for an example sentence in our corpus. The lines starting with # represent the IDs of the step in the recipe and the sentence in the step, respectively, while EOS represents the end of the sentence. The format of our corpus is based on the Kyoto University's corpora (Kawahara et al., 2002; Hashimoto et al., 2011; Hangyo et al., 2012) and output of CaboCha, which is one of the most popular dependency parsers for Japanese.

We first annotated the target sentences with morphemes. In Figure 1, the lines starting with a word such as 鮭 (salmon) give its morphological information such as part-of-speech (POS), fine-grained POS, base form, reading, pronunciation, and so on. We followed the IPA dictionary (Asahara and Matsumoto, 2003) to decide boundaries and POS for each morpheme because that resource is most commonly used



```
                # Step-ID:1
                # Sentence-ID:1-1
                * 0 4D 1/2 主題
           raw  生     接頭詞,名詞接続,*,*,*,*,生,ナマ,ナマ,B-Fi
        salmon  鮭     名詞,一般,*,*,*,*,鮭,サケ,サケ,I-Fi
(topic marker)  は     助詞,係助詞,*,*,*,*,は,ハ,ワ,O
                * 1 2D 1/2 補足語
        a bite  一口   名詞,一般,*,*,*,*,一口,ヒトクチ,ヒトクチ,B-Sf
          size  大     名詞,一般,*,*,*,*,大,ダイ,ダイ,I-Sf
      (dative)  に     助詞,格助詞,一般,*,*,*,に,ニ,ニ,O
                * 2 4P 0/0 述語
           cut  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
          salt  塩     名詞,一般,*,*,*,*,塩,シオ,シオ,B-Fi
  (accusative)  を     助詞,格助詞,一般,*,*,*,を,ヲ,ヲ,O
                * 4 -1O 0/0 述語
      sprinkle  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
```

Figure 1: Linguistic annotations for an example sentence, 生鮭は一口大に切り塩をふる。 (Cut the raw salmon into bite-size chunks and sprinkle them with salt.), in our corpus.

Tag	Description
Fi	Food (ingredient)
Fe	Food (part to be eliminated)
Fd	Food (dish)
Fa	Food (attribute)
Tg	Tool (general)
Ta	Tool (attribute)
To	Tool (other)
Nd	Number (duration)
Nq	Number (quantity)
No	Number (other)
Af	Action (food)
At	Action (tool)
Ap	Action (person)
Sf	State (food)
St	State (tool)
Sap	State (person's action)
X	Unclassified named entity

Table 3: Our named entity tags.

Type	Description
主題	Topic
補足語	Complement
連体修飾語	Adnominal modifier
連用修飾語	Predicative modifier
述語	Predicate
独立語	Independent
その他	Other

Table 4: Our bunsetsu types.

for Japanese MA. We also determined boundaries and POS for unknown words so that they fit the policies of the dictionary as much as possible. Consequently, 62,146 morphemes were annotated in the target sentences.

We then annotated the 62,146 morphemes with named entity tags. Table 3 shows our defined 17 tags, which are based on the 8 tags in the Flow Graph Corpus. In our corpus, the tags are located at the end of morphological information, as seen in Figure 1. Note that we used the common IOB2 format in NER to represent the inside (I), outside (O), and the beginning (B) of a named entity. For example, we can see from the figure that 生 (raw) is the beginning of the ingredient 生鮭 (raw salmon) because the morpheme is annotated with B-Fi. In this annotation, 22,359 entities were finally obtained from the target sentences.

We annotated a further 26,501 bunsetsus in the target sentences with dependency relations. A bunsetsu is a conventional unit of Japanese that consists of one or more content words (e.g., noun) and zero or more function words (e.g., particle). The example sentence in Figure 1 consists of five bunsetsus: 生鮭は, 一口大に, 切り, 塩を, and ふる。, and the lines starting with * give bunsetsu information. For example, the 0 in the first line starting with * denotes the index of the bunsetsu 生鮭は. The 4 in 4D denotes the index of the bunsetsu ふる。, which is the head of 生鮭は, while the D denotes the type of dependency relations, such as D (normal dependency), P (coordination dependency), and A (appositive dependency). Note that the index is set to -1 for the last bunsetsu in each sentence. This is because a dependency basically goes from left to right in Japanese, and thus the last bunsetsu has no head bunsetsu. Finally, we defined the 7 bunsetsu types in Table 4 and annotated all the bunsetsus with the types to clarify their roles in a sentence.

Our corpus enables researchers to use the above linguistic annotations for their studies. It was designed to link with two existing datasets: Cookpad Recipe Dataset and Cookpad Image Dataset (Harashima et al., 2017), which provide a variety of content such as reviews, meals, and images of the same 1.7 million recipes. As described, we extracted our 500 recipes from these recipes. This enables researchers to use not only the linguistic annotations of the 500 recipes in our corpus but also access the variety of content of the recipes in the two datasets for their studies.

	Precision	Recall	F1
MeCab	88.91	88.95	88.93
MeCab w/ DA	91.12	91.04	91.08

(a) MA.

	Accuracy
CaboCha	92.21
CaboCha w/ DA	94.68

(c) DP.

	Accuracy	Precision	Recall	F1
Sasada et al. (2015)	88.30	74.65	82.77	78.50
Lample et al. (2016)	91.41	88.17	87.18	87.67

(b) NER.

Table 5: Benchmark results on our corpus.

4 Experiments

Finally, we present benchmark results of our corpus for fundamental linguistic analyses in Japanese. In our experiments, we randomly divided the corpus into 400 recipes ($3,783$ sentences) for a training set, 50 recipes (472 sentences) for a validation set, and 50 recipes (483 sentences) for a test set. Then, we trained, tuned, and tested popular tools or methods for Japanese MA, NER, and DP using these recipes.

Table 5(a) shows the results for MA. We measured the performance of MeCab, the de facto standard morphological analyzer for Japanese, with and without performing domain adaptation (DA) of the tool for cooking recipes using our training set. The precision, recall, and F1 in the table were calculated based on the correct morphemes which the analyzer could recognize in our test set. From the table, we can see that all the metrics were approximately 91% even if we performed DA. This indicates that MA for informal cooking recipes is still an unsolved problem, compared to that for formal newspaper articles on which the tool has already achieved metrics of over 98% in Kudo et al. (2004).

The results for NER are given in Table 5(b). As there is no existing named entity recognizer that uses our named entity tags, two recognizers proposed in Sasada et al. (2015) and Lample et al. (2016), which are popular in recipe and general domains, respectively, were trained and evaluated using our training and test sets. The metrics in the table were calculated in the same way in the CoNLL-2003 shared task (Sang and Meulder, 2003). From the table, it is clear that there is still room for improvement in NER for cooking recipes. Most of the errors in our experiment were caused by domain-specific unknown words. Thus, a domain-specific lexicon such as cooking ontology (Nanba et al., 2014) may play an important role in reducing such errors.

Table 5(c) shows the results for DP. For this experiment, we used CaboCha, also mentioned in the previous section. The accuracy in the table was the percentage of the correct dependencies which the parser could recognize in our test set. Interestingly, the accuracy of 92-94$\%$ of the parser on our corpus was higher than the 90% for the Kyoto University Text Corpus, reported in Kudo and Matsumoto (2002). In other words, DP for informal cooking recipes might be slightly easier than for formal newspaper articles. This is probably because sentences in cooking recipes are relatively short, compared to those in newspaper articles. Having said that, we found that over 20% of the sentences in our test set had at least one parsing error even when we performed DA with the parser. This suggests that DP for cooking recipes also remains an unsolved problem.

5 Conclusion

This paper introduced the Cookpad Parsed Corpus, which contains linguistic annotations of Japanese recipes. The corpus was composed of 500 recipes which were extracted from the Cookpad Recipe Dataset. A total of $4,738$ sentences in the recipes were annotated with morphemes, named entities, and dependency relations. Benchmark results on our corpus for Japanese MA, NER, and DP were reported, and showed that there is still room for improvement in these analyses of cooking recipes. We believe the linguistic annotations will form a basic infrastructure not only for the improvement of the fundamental analyses, but also for the variety of recipe-related studies based on the analyses. The dataset can be obtained by sending an e-mail request to the authors. In future work, we plan to enrich our corpus with further linguistic annotations such as predicate-argument structures and co-references.

References

Masayuki Asahara and Yuji Matsumoto. 2003. ipadic version 2.7.0 User's Manual.

Khyathi Raghavi Chandu, Eric Nyberg, and Alan W Black. 2019. Storyboarding of recipes: Grounded contextual generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 6040–6046.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a Diverse Document Leads Corpus Annotated with Semantic Relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC 2012)*, pages 535–544.

Jun Harashima, Michiaki Ariga, Kenta Murata, and Masayuki Ioki. 2016. 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)*, pages 2455–2459.

Jun Harashima, Yuichiro Someya, and Yohei Kikuta. 2017. 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)*, pages 1229–1232.

Koichi Hashida. 2005. Global Document Annotation (GDA) Manual.

Chikara Hashimoto, Sadao Kurohashi, Daisuke Kawahara, Keiji Shinzato, and Masaaki Nagata. 2011. Construction of a Blog Corpus with Syntactic, Anaphoric, and Sentiment Annotations (in Japanese). *Natural Language Processing*, 18(2):175–201.

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

Jermsak Jermsurawong and Nizar Habash. 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.

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

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

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

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. 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)*, pages 260–270.

Angela S. Lin, Sudha Rao, Asli Celikyilmaz, Elnaz Nouri, Chris Brockett, Debadeepta Dey, and Bill Dolan. 2020. A recipe for creating multimodal aligned datasets for sequential tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4871–4884.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

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

Hidetsugu Nanba, Yoko Doi, Miho Tsujita, Toshiyuki Takezawa, and Kazutoshi Sumiya. 2014. Construction of a Cooking Ontology from Cooking Recipes and Patents. In *Proceedings of the 6th Workshop on Multimedia for Cooking and Eating Activities (CEA 2014)*, pages 507–516.

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

Liangming Pan, Jingjing Chen, Jianlong Wu, Shaoteng Liu, Chong-Wah Ngo, Min-Yen Kan, Yugang Jiang, and Tat-Seng Chua. 2020. Multi-modal cooking workflow construction for food recipes. In *Proceedings of the 27th ACM International Conference on Multimedia (ACM MM 2020)*.

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. 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)*, pages 3020–3028.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL 2003)*, pages 142–147.

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

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

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. 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)*, pages 1358–1368.

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