Gunma University, Kiryu University, and RMIT University at the NTCIR-11 Cooking Recipe Search Task

Michiko Yasukawa
Gunma University, Japan
michi@cs.gunma-u.ac.jp

Hiroji Ishii
Kiryu University, Japan
ishii-hi@kiryu-u.ac.jp

Falk Scholer
RMIT University, Australia
falk.scholer@rmit.edu.au

ABSTRACT

We report an empirical study of the NTCIR-11[2] Cooking Recipe Search task[4]. A series of experiments was performed in both Japanese and English based on a collaboration that involved research groups from Gunma University, Kiryu University and RMIT University. We compared baseline, oracle, and test search runs in the task. We also report the findings that we obtained from studies of food synonyms and recipe similarity.

Team Name

GUKURMITIR (Gunma University, Kiryu University and RMIT University collaborative IR research group). Group ID is GUKUR.

Language

English, Japanese.

Subtasks

Recipe Ad-hoc (subtask 1), Recipe Pairing (subtask 2).

Keywords

query expansion, synonyms, evaluation.

1. INTRODUCTION

Searches for cooking recipes work to some extent by using state-of-the-art web search techniques. For example, a quick search can be performed using Google to find a few popular recipes for some dishes. However, it may be unexpectedly difficult to satisfy the information needs of users in realistic situations of cooking and eating activities. A particular problem involves vocabulary mismatches between query words and how people describe the food names in existing recipes[1]. In the present study (Section 3), we evaluated baseline and oracle runs, which were generated using queries and various numbers of words derived from dish or ingredient names in answer recipes. For the Japanese subtasks, manually prepared dictionaries of food synonyms were examined to study synonymous words in recipes (Section 4). Another aspect of the recipe search problem is how to deal with negating words in search queries. Thus, it is necessary to consider how recipes with unwanted ingredients or preparation steps should be excluded. To address this question, we compared baseline runs and search runs, which were generated in the absence of negating query words in the English ad hoc search. We also investigated recipe similarity in the Japanese ad hoc search (Section 5).

Table 1: Evaluation measures calculated by NTCIREVAL

Abbr.	Description
MAP	Mean average precision.
MRR	Mean reciprocal rank.
MSnDCG	normalized discounted cumulative gain. (Microsoft version of nDCG)
nDCG	original nDCG
ERR	Expected Reciprocal Rank
RBP	Rank-biased Precision
NCU	Normalized Cumulative Utility
P@k	precision at k (number of relevant docs in top k divided by k.)
Hit@k	1 if top k contains a relevant doc, and 0 otherwise.

2. METHODOLOGY

As the experimental framework for our evaluation, we selected Indri Search Engine[1] version 5.7, which is a state-of-the-art search system. To perform the English search, we generated our submitted runs using Indri's default settings including the Dirichlet LM ranking function, no-stemming, and no-stopping. To perform the Japanese search, we used the Japanese morphological analyzer MeCab[2] version 0.996 with ipadic-2.7.0. We compared various evaluation values for our submitted runs that are calculated with a basic NTCIR evaluation tool[3], including the three official evaluation values for the task (MAP, MRR, and MSnDCG).

Our submitted runs were generated using formal run queries for the baseline runs and answer recipes for oracle runs and test runs. The answer recipes were provided to guide the search for recipes relevant to the given queries. Examples of the English/Japanese experimental data are provided in the task overview study[4].

3. EXPERIMENTAL RESULTS

We submitted five runs for each of the English/Japanese subtasks 1/2. The total number of runs submitted was 20. The runs were all generated using Indri (default setting) with no-stemming and no-stopping. In the Japanese experiments, MeCab and IPAdic were used for word-breaking in the Japanese texts.

3.1 English ad hoc recipe search (EN1)

Table 7 shows the evaluation values used for the English ad hoc

[1]http://sourceforge.net/projects/lemur/
[2]https://code.google.com/p/mecab/
[3]NTCIREVAL.130507 `http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html`

recipe search (EN1). The system inputs for the submitted runs are explained as follows.

GUKUR-EN1-BASE-01:
The system input comprised queries (all terms).

GUKUR-EN1-BASE-02:
The system input comprised queries (dropping negation terms).

GUKUR-EN1-ORCL-01:
The system input comprised queries and answer examples (recipe title).

GUKUR-EN1-ORCL-02:
The system input comprised queries and answer examples (recipe title and top ingredient lines).

GUKUR-EN1-ORCL-03:
The system input comprised queries and answer examples (recipe title and all ingredient lines).

For all three official measures (MAP, MRR, and MSnDCG@1000), the ORCL-3 run variant performed substantially better than the baseline runs. This is particularly noticeable for MRR, where this Oracle variant achieves a score of 0.939, compared to 0.357 and 0.386 for the two baselines. Adding the full set of all ingredient lines provides a substantial boost to retrieval performance.

3.2 English recipe pairing (EN2)

Table 8 shows the evaluation values used for the English recipe pairing (EN2). The system inputs for the submitted runs are explained as follows.

GUKUR-EN2-ORCL-01:
System input comprised answer examples (recipe title).

GUKUR-EN2-ORCL-02:
System input comprised answer examples (recipe title, either of the 3 attributes). The three attributes are side, salad, or dessert.

GUKUR-EN2-ORCL-03:
System input comprised answer examples (recipe title, all attributes). The attributes include rolls in a menu (side, salad, or dessert), seasons, cuisine styles, etc.

GUKUR-EN2-ORCL-04:
System input comprised answer examples (recipe title, top ingredient lines).

GUKUR-EN2-ORCL-05:
System input comprised answer examples (recipe title, all ingredient lines).

Based on the official measures, Oracle variant 5 which used all ingredient lines showed by far the highest performance, achieving a near-perfect score (0.973 for MAP and MRR, and 0.980 for MSnDCG@1000).

3.3 Japanese ad hoc recipe search (JA1)

Table 9 shows the evaluation values for Japanese ad hoc recipe search (JA1). The system inputs for the submitted runs are explained as follows.

GUKUR-JA1-BASE-01:
System input comprised queries (dish name).

GUKUR-JA1-BASE-02:
System input comprised queries (dish name, ingredient names).

GUKUR-JA1-BASE-03:
System input comprised queries (dish name, negation/explanation conditions).

GUKUR-JA1-BASE-04:
System input comprised queries (all).

GUKUR-JA1-TEST-01:
System input comprised answer examples and a hand-made dictionary.

The run GUKUR-JA1-TEST-01 achieved the highest performance on the three official measures. This is due to the inclusion of a hand-made dictionary, which incorporated information such as alternative expressions and abbreviations for food names.

3.4 Japanese recipe pairing (JA2)

Table 10 shows the evaluation values for Japanese recipe pairing (JA2). The system inputs for the submitted runs are explained as follows.

GUKUR-JA2-BASE-01:
System input comprised side dish information in formal run queries (dish name).

GUKUR-JA2-BASE-02:
System input comprised side dish information in formal run queries (ingredient names).

GUKUR-JA2-BASE-03:
System input comprised side dish information in formal run queries (dish name, ingredient names).

GUKUR-JA2-TEST-01:
System input comprised answer examples (dish name, top ingredient names) and a hand-made dictionary.

GUKUR-JA2-TEST-02:
System input comprised answer examples (dish name, all ingredient names) and a hand-made dictionary.

In this subtask, the run GUKUR-JA2-TEST-02 achieved the highest performance across the three official measures. This run incorporated the same handmade dictionary described in the previous section. Note however that the run GUKUR-JA2-TEST-01 also used this dictionary, but performed much less effectively. The additional inclusion of all (versus just the top) ingredient names in the former provided a substantial additional boost to performance.

4. A STUDY OF FOOD SYNONYMS

To investigate how synonymous words are recognized differently by person to person, we performed a user study of synonyms of Japanese food names.

To develop the queries for the Japanese subtasks, hand-made dictionaries of Japanese food names were edited and categorized by word meaning. The dictionaries contained 596 pairs of synonyms for Japanese subtask 1 and 331 pairs of synonyms for Japanese subtask 2. Common synonyms were present in the dictionaries used for Japanese subtask 1 and subtask 2. We obtained 769 unique pairs of synonyms from the dictionaries.

Two assessors then assigned labels to 769 pairs of synonymous words for Japanese food names. One of the assessors (Assessor-X)

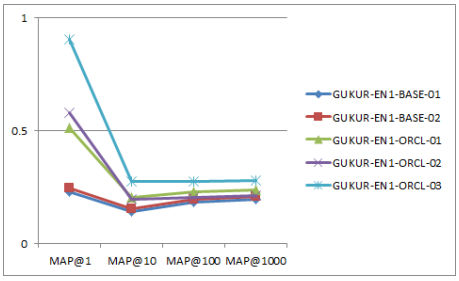

Figure 1: MAP@k for EN1 (Ad hoc)

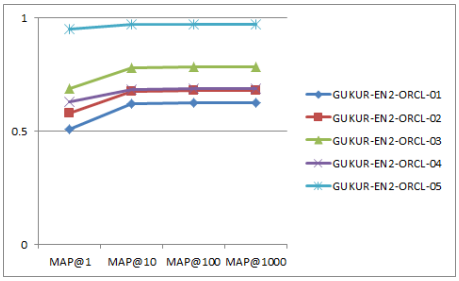

Figure 2: MAP@k for EN2 (Pairing)

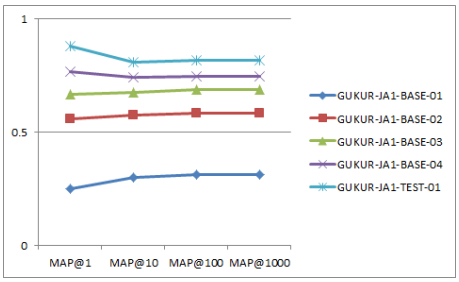

Figure 3: MAP@k for JA1 (Ad hoc)

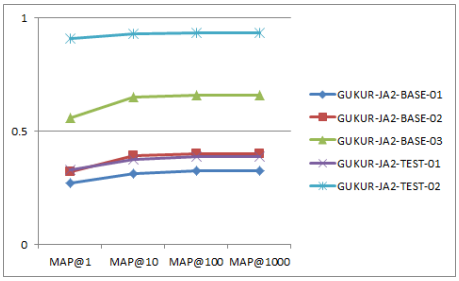

Figure 4: MAP@k for JA2 (Pairing)

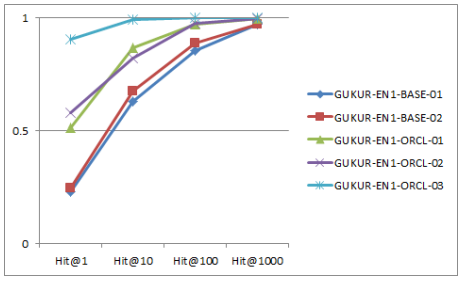

Figure 5: Hit@k for EN1 (Ad hoc)

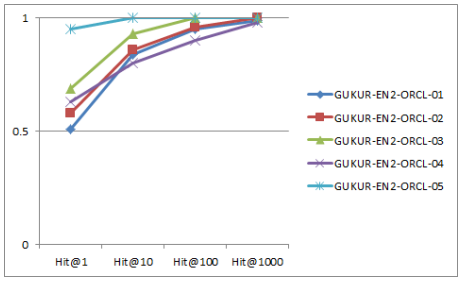

Figure 6: Hit@k for EN2 (Pairing)

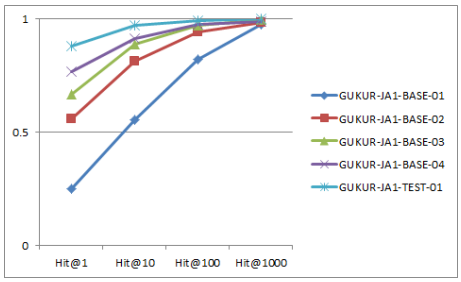

Figure 7: Hit@k for JA1 (Ad hoc)

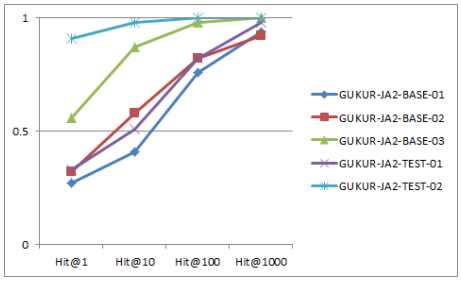

Figure 8: Hit@k for JA2 (Pairing)

was a non-professional who enjoyed cooking at home, whereas the other assessor (Assessor-Y) was a professional who was a qualified dietitian. During the development of the dictionary, the task organizer collected synonymous words from recipes in the Rakuten Recipe corpus and created word pairs based on the meaning of the words in the recipes. During the assessment, the two assessors had no access to the recipes and they judged the word similarity by reviewing word pairs without referring to the word context. The two assessors were instructed to assign a label of '0' if the pair of food names had different meaning, and assign a label of '1' if the pair of food names had the same meaning.

Table 6 shows some examples of the synonym pairs and their inter-rater agreements/disagreements. In Table 6, the symbol 'F' indicates that the word pair was labeled as '0' and the symbol 'T' indicates that the word pair was labeled as '1'. Of the 769 synonym pairs, Assessor-X disagreed with the task organizer in 13 cases, but Assessor-Y had 77 disagreements with the task organizer. Table 2 shows the numbers of inter-rater agreements/disagreements for the 769 pairs of synonymous words.

Table 2: Number of inter-rater agreements/disagreements of the 769 pairs of synonymous words for Japanese food names. The symbol 'T' indicates that the pair of food names is similar. The symbol 'F' indicates that the pair of food names is not similar.

Task Organizer	Assessor-X	Assessor-Y	# pairs
T	T	T	688
T	T	F	68
T	F	T	4
T	F	F	9
		total	769

According to the inter-rater agreements shown in Table 2, the proportions of agreements (T-T-T) and disagreements (T-T-F, T-F-T, or T-F-F) are 89% and 11%, respectively. Krippendorff's alpha is calculated as alpha = 0.0637 (Subjects = 769, Raters = 3). As shown in Table 6, inter-rater agreements are more likely if the synonyms are related to the spelling or word order in phrasal expressions. They are detected easily by knowledgeable speakers of the Japanese language. Thus, it would be useful if these synonymous words could be automatically recognized as common synonyms in recipe search systems.

If the synonymous words require semantic interpretation according to the word context in recipes, inter-rater disagreements are more likely. While all of the word pairs had been made by the Task Organizer (one of the authors of this report), double-check of each pair using the sources (i.e., recipes in the Rakuten Recipe corpus) was necessary to verify the synonymous word pairs that resulted in disagreements (T-T-F, T-F-T, or T-F-F). The nine disagreements where both Assessor-X and Assessor-Y assigned the label '0' and the Task Organizer assigned the label '1' can be classified into the following two types. One type is a spelling error of a short word, such as 'paprika' and 'pabrika.' The other type is a food name that contains an irreversible cooking process, such as 'garlic' and 'sliced garlic.'

We consider that the 68 disagreements (T-T-F) and the four disagreements (T-F-T) were caused by the subjective views of the Assessor-X and Assessor-Y. Assessor-X who enjoyed cooking as a hobby was more tolerant of different food name expressions. Assessor-Y who specialized in nutrition was more rigorous about the correspondence between food concepts and food names. Among the 81 cases where one or both of the assessors disagree with the task organizer (T-T-F, T-F-T, or T-F-F), the proportions of agreements by Assessor-X and Assessor-Y (T-F-F), and disagreements by Assessor-X and Assessor-Y (T-T-F, T-F-T) were 11% and 89%, respectively. The pairwise Cohen's kappa score for the inter-rater agreements between Assessor-X and Assessor-Y for the 81 pairs of food names is Kappa = -0.103 (Subjects = 81, Raters = 2, z = -4.69, p-value < 0.0001).

The results of this study demonstrate that independent food names in recipes can be recognized differently from person to person. Our next question is whether a recipe can be recognized as similar or different from person to person because multiple food names are present in a single recipe. Thus, we present our investigation of recipe similarity in the next section.

5. A STUDY OF RECIPE SIMILARITY

To investigate recipe similarity, we developed evaluation data for similar recipes to the answer recipes for the Japanese ad hoc search.

The evaluation data was obtained based on a manual relevance assessment by five students and two of the authors of this report. Candidate similar recipes were collected using the search engine, Indri (default settings), with no-stemming, no-stopping, and Boolean (AND) search. The system input comprised the recipe title and the top four ingredients in each answer recipe. The number of the ingredients was chosen by a trial-and-error method. If the number of the ingredients was large, greatly similar recipe search was performed. In this case, recipe pairs were not obtained in most cases because the only search result was the answer recipe itself. If the number of the ingredients was small, similar recipe search was poorly performed because many pairs of recipes were obtained, but the pairs were not similar in most cases.

Because of the limited number of human assessors and the time constraints on the task schedule, we choose four for the number of ingredients, and set the run depth at four to create a pool of similar recipe pairs. As a result, we obtained 236 pairs for recipe similarity assessment with 111 ad hoc queries.

By learning from the latest open-source relevance assessment system, Relevation![3], we developed an intuitive assessment system using Apache, PHP, and PostgreSQL on a Linux server. In our assessment system, the assessors could compare two similar recipes side by side and input relevance labels one after another. The relevance labels used by our assessment system were three types: "similar," "not similar," and "not judged."

The assessors were instructed to: (1) assume that you are a user of a similar recipe search system and that you are given a pair of recipes as a search result, (2) select the radio button 'similar' if you think that the query recipe (on the left) and the searched recipe (on the right) are similar; and (3) select the radio button 'not similar' if you do not think that the two recipes are similar.

Each of the similar pairs was judged by two assessors, and 236 multilevel judgments (L0, L1, L2) were collected for the 111 queries[4]. The multilevel judgments collected in the user study were as follows.

L2 two assessors judged the two recipes as similar.

L1 one assessor judged the two recipes as similar.

L0 zero assessor judged the two recipes as similar.

[4]We differentiated a "similar" recipe from the "same" recipe, and multi-level relevance for the JA1 evaluation comprised L0 (not relevant), L1 (somewhat similar to the answer recipe), L2 (highly similar to the answer recipe), and L3 (the answer recipe itself).

The percentages of L2, L1, and L0 levels assigned to the 236 pairs are 44%, 26%, and 30%, respectively. The proportion of inter-rater agreements (T-T or F-F) and disagreements (T-F or F-T) are 74% and 26%, respectively. The details of the numbers of inter-rater agreements/disagreements for similar recipes are shown in Table 3.

Table 3: Number of inter-rater agreements/disagreements for the 236 pairs of recipes. The symbol 'T' indicates that the pair of recipes is similar. The symbol 'F' indicates that the pair of recipes is not similar.

S-1	S-2	S-3	S-4	S-5	Auth-1	Auth-2	# pairs
T	T						51
F	F						19
T	F						12
F	T						4
T				T			28
F				F			17
T				F			17
F				T			0
		T	T				24
		F	F				35
		T	F				15
		F	T				12
					T	F	1
					F	F	1
						total	236

Table 4: Number and percentage of similar/not-similar pairs of recipes. The symbol 'T' indicates that the pair of recipes is similar. The symbol 'F' indicates that the pair of recipes is not similar.

	S-1	S-2	S-3	S-4	S-5	Auth-1	Auth-2
# T	108	55	39	36	28	1	0
# F	40	31	47	50	34	1	2
# total	148	86	86	86	62	2	2
pct. T	73%	64%	45%	41%	45%	50%	0%
pct. F	27%	36%	55%	59%	55%	50%	100%

Among the five students, two students (Student-4 and Student-5) had advanced knowledge of cooking, and they were relatively more efficient in the assessment process. Student-1 and Student2 were students who did not cook often and they assessed the recipe similarity based on their overall impressions when reading the recipes. They spent more time making assessments than Student-4 and Student-5 did. Student-3 was a student who had the same level of advanced cooking knowledge as Student-4 and Student-5. However, Student-3 was an international student and needed more time to read the Japanese recipes.

After the similarity assessment, Student-4 and Student-5 completed a quick survey about all of the assessment data, which showed that they employed much clearer criteria for assessing similarity compared with the other students. Their criteria are explained as follows.

Student-4: If the two recipes only differed in terms of a few ingredients, they were assessed as similar. Two recipes that could produce the same food were assessed as similar even if the toppings or sauces tasted differently. If the ingredient lines or preparation steps differed greatly (e.g., the text length of recipes differed substantially), the two recipes were assessed as not similar.

Student-5: If the completed dishes prepared using the two recipes had the same taste/flavor/texture and the same role (main dish or side dish) in a multi-course menu, the two recipes were assessed as similar. If the major ingredients were typically common and the differences in the ingredient lines were subtle, such as two dishes with different levels of richness, saltiness, or sweetness, the two recipes used to produce the two dishes were assessed as not being similar.

Both Student-4 and Student-5 recognized that their recipe similarity assessments were different from those of the other students. Student-5 thought that Student-5's criteria were relatively closer to those of Student-3 rather than those of Student-4. In terms of cooking activity, Student-3 and Student-5 were more interested in preparing everyday meals, whereas Student-4 was more interested in preparing sweets and desserts.

For the inter-rater agreements in Table 3, Krippendorff's alpha is calculated as alpha = 0.475 (Subjects = 236, Raters = 7). The pairwise Cohen's kappa scores are presented in Table 5. Note that assessors Auth-1 and Auth-2 only judged 2 items. They agreed on one, and disagreed on the other; this is what would be expected by chance, and hence Kappa is 0 in this case.

Table 5: The pairwise Cohen's kappa scores for the inter-rater agreements of the 236 pairs of recipes.

Assessors	Kappa	*p*-value
S-1 and S-2	0.572	p<0.0001
S-1 and S-5	0.475	p<0.0001
S-3 and S-4	0.362	p=0.0008
Auth-1 and Auth-2	0	NaN

The results described above indicate that the recognition of similarity of cooking recipes can differ from person to person. The frequency and preferences of cooking of assessors may be important factors that affect recipe similarity assessments, but further investigation are required to verify whether this is the case.

Based on the recipe similarity assessments given by the students, one of the authors of this report prepared a small run pool from the submitted runs and found another 24 similar recipes for 17 example answer recipes in the Japanese ad hoc (JA1). As a result, the total number of relevant recipes and similar recipes for JA1 was 760, and these recipes were used for the official evaluation of the task. For the Japanese recipe pairing (JA2), four similar recipes for three example answers were found from the submitted runs. The total number of relevant recipes and similar recipes for JA2 is 104 and these recipes were used for the official evaluation of the task. There might have been other unjudged relevant recipes for JA1 and JA2, but we had insufficient time to perform an additional assessment.

6. CONCLUSIONS

We compared the search effectiveness of our submitted runs using various evaluation values. Query expansion using synonymous words was effective for the recipe search task. Building and maintaining a synonymous food name list by human assessors is feasible, but it is too costly. At present, it is considerably difficult for

Table 6: Examples of inter-rater agreement/disagreement for synonymous food names.

Food name (A)	Food name (B)	Task Organizer	Assessor-X	Assessor-Y
いちご/ジャム	苺ジャム	T	T	T
いちじく/乾燥	ドライフィグ	T	T	T
うなぎ/かば焼	うなぎ蒲焼	T	T	T
かいわれだいこん	貝割れ	T	T	T
かに風味かまぼこ	カニカマ	T	T	T
きな粉	黄粉	T	T	T
しょうが/おろし	生姜/チューブ入り	T	T	T
ひき肉/合い挽き	牛豚合挽肉	T	T	T
アーモンド粉	アーモンドプードル	T	T	T
うどん/玉	うどん玉	T	T	F
お好み焼きソース	お好みソース	T	T	F
だし類/コンソメ/キューブ	キューブコンソメ	T	T	F
チーズ/とろけるタイプ	溶けるチーズ	T	T	F
プレミックス粉/お好み焼き用	おこのみ粉	T	T	F
一味唐辛子	一味	T	T	F
かつお節	おかか	T	F	T
からし	マスタード	T	F	T
ごはん	雑穀米/ごはん	T	F	T
月桂樹	ローレル	T	F	T
パプリカ	パプリカ	T	F	F
かたくり粉	水溶き片粉	T	F	F
にんにく	ニンニク/スライス	T	F	F
ねぎ	刻み葱	T	F	F
塩	アジシオ	T	F	F
鶏卵	目玉焼き	T	F	F

computers to provide intuitive synonymous food names. As a result of our user study, we found that the similarity assessments of recipes and food name synonyms could differ from person to person. In future work, we will investigate how to predict a side dish suggestion for a given main dish.

7. ACKNOWLEDGMENTS

This work was supported partly by JSPS KAKENHI Grant Number 26330363. For the English subtasks, we used Yummly Recipe Data v1 provided by Yummly. For the Japanese subtasks, we used the Rakuten Data provided by Rakuten, Inc. We express our gratitude to the data providers. We are thankful to the student assessors at Gunma University and the faculty member at Kiryu University for their cooperation in our user study.

8. REFERENCES

[1] Sally Jo Cunningham and David Bainbridge. An analysis of cooking queries: implications for supporting leisure cooking. In *Proceedings of iConference*, pages 112–123, 2013.

[2] Hideo Joho and Kazuaki Kishida. Overview of NTCIR-11. In *Proceedings of NTCIR-11*, 2014.

[3] Bevan Koopman and Guido Zuccon. Relevation!: an open source system for information retrieval relevance assessment. In *Proceedings of SIGIR2014*, pages 1243–1244, 2014.

[4] Michiko Yasukawa, Fernando Diaz, Gregory Druck, and Nobu Tsukada. Overview of the NTCIR-11 cooking recipe search task. In *Proceedings of NTCIR-11*, 2014.

Table 7: Detailed results for EN1 (Ad hoc). The evaluation values in the table were calculated by using a basic NTCIR evaluation tool called 'NTCIREVAL.' See the README file of NTCIREVAL.130507 for more information on each value.

Run ID	MAP	MRR	ERR	RBP
GUKUR-EN1-BASE-01	0.1949	0.3566	0.1672	0.0591
GUKUR-EN1-BASE-02	0.2080	0.3859	0.1812	0.0636
GUKUR-EN1-ORCL-01	0.2381	0.6352	0.2639	0.0588
GUKUR-EN1-ORCL-02	0.2126	0.6691	0.2628	0.0449
GUKUR-EN1-ORCL-03	0.2815	0.9389	0.3681	0.0531

Run ID	O-measure	P-measure	Q-measure	P-plus
GUKUR-EN1-BASE-01	0.3643	0.3643	0.2314	0.3643
GUKUR-EN1-BASE-02	0.3931	0.3931	0.2453	0.3931
GUKUR-EN1-ORCL-01	0.6399	0.6399	0.2681	0.6399
GUKUR-EN1-ORCL-02	0.6728	0.6728	0.2335	0.6728
GUKUR-EN1-ORCL-03	0.9399	0.9399	0.3024	0.9399

Run ID	NCUgu,BR	NCUgu,P	NCUrb,BR	NCUrb,P
GUKUR-EN1-BASE-01	0.2314	0.1949	0.2418	0.2084
GUKUR-EN1-BASE-02	0.2453	0.2080	0.2571	0.2231
GUKUR-EN1-ORCL-01	0.2681	0.2381	0.2870	0.2593
GUKUR-EN1-ORCL-02	0.2335	0.2126	0.2528	0.2330
GUKUR-EN1-ORCL-03	0.3024	0.2815	0.3317	0.3118

Run ID	MAP@1	MAP@10	MAP@100	MAP@1000
GUKUR-EN1-BASE-01	0.2280	0.1419	0.1850	0.1949
GUKUR-EN1-BASE-02	0.2480	0.1537	0.1976	0.2080
GUKUR-EN1-ORCL-01	0.5140	0.2043	0.2292	0.2381
GUKUR-EN1-ORCL-02	0.5820	0.1957	0.2054	0.2126
GUKUR-EN1-ORCL-03	0.9040	0.2750	0.2736	0.2815

Run ID	Hit@1	Hit@10	Hit@100	Hit@1000
GUKUR-EN1-BASE-01	0.2280	0.6300	0.8540	0.9740
GUKUR-EN1-BASE-02	0.2480	0.6760	0.8880	0.9740
GUKUR-EN1-ORCL-01	0.5140	0.8660	0.9740	0.9980
GUKUR-EN1-ORCL-02	0.5820	0.8240	0.9760	0.9960
GUKUR-EN1-ORCL-03	0.9040	0.9920	1.0000	1.0000

Run ID	MSnDCG@1	MSnDCG@10	MSnDCG@100	MSnDCG@1000
GUKUR-EN1-BASE-01	0.2280	0.2280	0.3595	0.4381
GUKUR-EN1-BASE-02	0.2480	0.2476	0.3804	0.4571
GUKUR-EN1-ORCL-01	0.5140	0.3250	0.4139	0.4893
GUKUR-EN1-ORCL-02	0.5820	0.3063	0.3674	0.4362
GUKUR-EN1-ORCL-03	0.9040	0.4136	0.4574	0.5347

Run ID	P@1	P@10	P@100	P@1000
GUKUR-EN1-BASE-01	0.2280	0.1720	0.0609	0.0100
GUKUR-EN1-BASE-02	0.2480	0.1890	0.0643	0.0103
GUKUR-EN1-ORCL-01	0.5140	0.1870	0.0526	0.0090
GUKUR-EN1-ORCL-02	0.5820	0.1476	0.0383	0.0071
GUKUR-EN1-ORCL-03	0.9040	0.1810	0.0418	0.0081

Run ID	Q@1	Q@10	Q@100	Q@1000
GUKUR-EN1-BASE-01	0.2280	0.1470	0.2137	0.2314
GUKUR-EN1-BASE-02	0.2480	0.1586	0.2268	0.2453
GUKUR-EN1-ORCL-01	0.5140	0.2085	0.2520	0.2681
GUKUR-EN1-ORCL-02	0.5820	0.1981	0.2205	0.2335
GUKUR-EN1-ORCL-03	0.9040	0.2773	0.2879	0.3024

Run ID	nDCG@1	nDCG@10	nDCG@100	nDCG@1000
GUKUR-EN1-BASE-01	0.2280	0.2251	0.3402	0.4095
GUKUR-EN1-BASE-02	0.2480	0.2447	0.3608	0.4286
GUKUR-EN1-ORCL-01	0.5140	0.3139	0.3914	0.4578
GUKUR-EN1-ORCL-02	0.5820	0.2875	0.3406	0.4010
GUKUR-EN1-ORCL-03	0.9040	0.3867	0.4246	0.4927

Run ID	nERR@1	nERR@10	nERR@100	nERR@1000
GUKUR-EN1-BASE-01	0.2280	0.2870	0.3111	0.3124
GUKUR-EN1-BASE-02	0.2480	0.3133	0.3374	0.3385
GUKUR-EN1-ORCL-01	0.5140	0.4825	0.4999	0.5010
GUKUR-EN1-ORCL-02	0.5820	0.4865	0.5022	0.5035
GUKUR-EN1-ORCL-03	0.9040	0.6913	0.7041	0.7053

Table 8: Detailed results for EN2 (Pairing). The evaluation values in the table were calculated by using a basic NTCIR evaluation tool called 'NTCIREVAL.' See the README file of NTCIREVAL.130507 for more information on each value.

Run ID	MAP	MRR	ERR	RBP
GUKUR-EN2-ORCL-01	0.6255	0.6252	0.2118	0.0217
GUKUR-EN2-ORCL-02	0.6795	0.6792	0.2299	0.0222
GUKUR-EN2-ORCL-03	0.7824	0.7820	0.2642	0.0240
GUKUR-EN2-ORCL-04	0.6888	0.6883	0.2319	0.0206
GUKUR-EN2-ORCL-05	0.9725	0.9725	0.3286	0.0259

Run ID	O-measure	P-measure	Q-measure	P-plus
GUKUR-EN2-ORCL-01	0.6795	0.6795	0.6800	0.6795
GUKUR-EN2-ORCL-02	0.7252	0.7252	0.7257	0.7252
GUKUR-EN2-ORCL-03	0.8214	0.8214	0.8221	0.8214
GUKUR-EN2-ORCL-04	0.7175	0.7175	0.7182	0.7175
GUKUR-EN2-ORCL-05	0.9807	0.9807	0.9807	0.9807

Run ID	NCUgu,BR	NCUgu,P	NCUrb,BR	NCUrb,P
GUKUR-EN2-ORCL-01	0.6800	0.6255	0.6800	0.6255
GUKUR-EN2-ORCL-02	0.7257	0.6795	0.7257	0.6795
GUKUR-EN2-ORCL-03	0.8221	0.7824	0.8221	0.7824
GUKUR-EN2-ORCL-04	0.7182	0.6888	0.7182	0.6888
GUKUR-EN2-ORCL-05	0.9807	0.9725	0.9807	0.9725

Run ID	MAP@1	MAP@10	MAP@100	MAP@1000
GUKUR-EN2-ORCL-01	0.5100	0.6201	0.6253	0.6255
GUKUR-EN2-ORCL-02	0.5800	0.6751	0.6793	0.6795
GUKUR-EN2-ORCL-03	0.6900	0.7794	0.7824	0.7824
GUKUR-EN2-ORCL-04	0.6300	0.6850	0.6885	0.6888
GUKUR-EN2-ORCL-05	0.9500	0.9725	0.9725	0.9725

Run ID	Hit@1	Hit@10	Hit@100	Hit@1000
GUKUR-EN2-ORCL-01	0.5100	0.8400	0.9500	0.9900
GUKUR-EN2-ORCL-02	0.5800	0.8600	0.9600	1.0000
GUKUR-EN2-ORCL-03	0.6900	0.9300	1.0000	1.0000
GUKUR-EN2-ORCL-04	0.6300	0.8000	0.9000	0.9800
GUKUR-EN2-ORCL-05	0.9500	1.0000	1.0000	1.0000

Run ID	MSnDCG@1	MSnDCG@10	MSnDCG@100	MSnDCG@1000
GUKUR-EN2-ORCL-01	0.5100	0.6737	0.6980	0.7031
GUKUR-EN2-ORCL-02	0.5800	0.7202	0.7416	0.7466
GUKUR-EN2-ORCL-03	0.6900	0.8168	0.8315	0.8315
GUKUR-EN2-ORCL-04	0.6300	0.7132	0.7332	0.7429
GUKUR-EN2-ORCL-05	0.9500	0.9795	0.9795	0.9795

Run ID	P@1	P@10	P@100	P@1000
GUKUR-EN2-ORCL-01	0.5100	0.0870	0.0099	0.0010
GUKUR-EN2-ORCL-02	0.5800	0.0890	0.0100	0.0010
GUKUR-EN2-ORCL-03	0.6900	0.0970	0.0104	0.0010
GUKUR-EN2-ORCL-04	0.6300	0.0830	0.0093	0.0010
GUKUR-EN2-ORCL-05	0.9500	0.1040	0.0104	0.0010

Run ID	Q@1	Q@10	Q@100	Q@1000
GUKUR-EN2-ORCL-01	0.5100	0.6699	0.6797	0.6800
GUKUR-EN2-ORCL-02	0.5800	0.7174	0.7254	0.7257
GUKUR-EN2-ORCL-03	0.6900	0.8164	0.8221	0.8221
GUKUR-EN2-ORCL-04	0.6300	0.7109	0.7176	0.7182
GUKUR-EN2-ORCL-05	0.9500	0.9807	0.9807	0.9807

Run ID	nDCG@1	nDCG@10	nDCG@100	nDCG@1000
GUKUR-EN2-ORCL-01	0.5100	0.7327	0.7568	0.7619
GUKUR-EN2-ORCL-02	0.5800	0.7726	0.7936	0.7986
GUKUR-EN2-ORCL-03	0.6900	0.8693	0.8843	0.8843
GUKUR-EN2-ORCL-04	0.6300	0.7390	0.7592	0.7689
GUKUR-EN2-ORCL-05	0.9500	0.9950	0.9950	0.9950

Run ID	nERR@1	nERR@10	nERR@100	nERR@1000
GUKUR-EN2-ORCL-01	0.5100	0.6201	0.6251	0.6253
GUKUR-EN2-ORCL-02	0.5800	0.6751	0.6792	0.6794
GUKUR-EN2-ORCL-03	0.6900	0.7792	0.7822	0.7822
GUKUR-EN2-ORCL-04	0.6300	0.6848	0.6882	0.6886
GUKUR-EN2-ORCL-05	0.9500	0.9725	0.9725	0.9725

Table 9: Detailed results for JA1 (Ad hoc). The evaluation values in the table were calculated by using a basic NTCIR evaluation tool called 'NTCIREVAL.' See the README file of NTCIREVAL.130507 for more information on each value.

Run ID	MAP	MRR	ERR	RBP
GUKUR-JA1-BASE-01	0.3146	0.3517	0.1814	0.0339
GUKUR-JA1-BASE-02	0.5846	0.6490	0.3307	0.0460
GUKUR-JA1-BASE-03	0.6871	0.7465	0.3821	0.0521
GUKUR-JA1-BASE-04	0.7489	0.8207	0.4189	0.0528
GUKUR-JA1-TEST-01	0.8168	0.9138	0.4669	0.0568

Run ID	O-measure	P-measure	Q-measure	P-plus
GUKUR-JA1-BASE-01	0.3938	0.3938	0.3607	0.3938
GUKUR-JA1-BASE-02	0.6837	0.6837	0.6240	0.6837
GUKUR-JA1-BASE-03	0.7739	0.7739	0.7206	0.7739
GUKUR-JA1-BASE-04	0.8381	0.8381	0.7722	0.8381
GUKUR-JA1-TEST-01	0.9240	0.9240	0.8353	0.9240

Run ID	NCUgu,BR	NCUgu,P	NCUrb,BR	NCUrb,P
GUKUR-JA1-BASE-01	0.3607	0.3146	0.3618	0.3157
GUKUR-JA1-BASE-02	0.6240	0.5846	0.6259	0.5865
GUKUR-JA1-BASE-03	0.7206	0.6871	0.7222	0.6889
GUKUR-JA1-BASE-04	0.7722	0.7489	0.7743	0.7511
GUKUR-JA1-TEST-01	0.8353	0.8168	0.8380	0.8197

Run ID	MAP@1	MAP@10	MAP@100	MAP@1000
GUKUR-JA1-BASE-01	0.2520	0.2996	0.3132	0.3146
GUKUR-JA1-BASE-02	0.5580	0.5755	0.5838	0.5846
GUKUR-JA1-BASE-03	0.6680	0.6771	0.6865	0.6871
GUKUR-JA1-BASE-04	0.7680	0.7413	0.7484	0.7489
GUKUR-JA1-TEST-01	0.8800	0.8088	0.8164	0.8168

Run ID	Hit@1	Hit@10	Hit@100	Hit@1000
GUKUR-JA1-BASE-01	0.2520	0.5560	0.8200	0.9760
GUKUR-JA1-BASE-02	0.5580	0.8140	0.9420	0.9860
GUKUR-JA1-BASE-03	0.6680	0.8900	0.9740	0.9980
GUKUR-JA1-BASE-04	0.7680	0.9140	0.9780	0.9880
GUKUR-JA1-TEST-01	0.8800	0.9740	0.9940	1.0000

Run ID	MSnDCG@1	MSnDCG@10	MSnDCG@100	MSnDCG@1000
GUKUR-JA1-BASE-01	0.2520	0.3551	0.4195	0.4476
GUKUR-JA1-BASE-02	0.5580	0.6307	0.6678	0.6811
GUKUR-JA1-BASE-03	0.6680	0.7255	0.7587	0.7688
GUKUR-JA1-BASE-04	0.7680	0.7817	0.8078	0.8157
GUKUR-JA1-TEST-01	0.8800	0.8477	0.8723	0.8780

Run ID	P@1	P@10	P@100	P@1000
GUKUR-JA1-BASE-01	0.2520	0.0628	0.0111	0.0015
GUKUR-JA1-BASE-02	0.5580	0.0904	0.0124	0.0014
GUKUR-JA1-BASE-03	0.6680	0.1012	0.0132	0.0015
GUKUR-JA1-BASE-04	0.7680	0.1038	0.0131	0.0015
GUKUR-JA1-TEST-01	0.8800	0.1104	0.0137	0.0015

Run ID	Q@1	Q@10	Q@100	Q@1000
GUKUR-JA1-BASE-01	0.2520	0.3333	0.3581	0.3607
GUKUR-JA1-BASE-02	0.5580	0.6076	0.6224	0.6240
GUKUR-JA1-BASE-03	0.6680	0.7029	0.7193	0.7206
GUKUR-JA1-BASE-04	0.7680	0.7589	0.7712	0.7722
GUKUR-JA1-TEST-01	0.8800	0.8213	0.8345	0.8353

Run ID	nDCG@1	nDCG@10	nDCG@100	nDCG@1000
GUKUR-JA1-BASE-01	0.2520	0.3815	0.4420	0.4682
GUKUR-JA1-BASE-02	0.5580	0.6598	0.6933	0.7049
GUKUR-JA1-BASE-03	0.6680	0.7463	0.7757	0.7844
GUKUR-JA1-BASE-04	0.7680	0.7925	0.8150	0.8216
GUKUR-JA1-TEST-01	0.8800	0.8459	0.8668	0.8716

Run ID	nERR@1	nERR@10	nERR@100	nERR@1000
GUKUR-JA1-BASE-01	0.2520	0.3270	0.3380	0.3389
GUKUR-JA1-BASE-02	0.5580	0.6203	0.6264	0.6268
GUKUR-JA1-BASE-03	0.6680	0.7207	0.7260	0.7263
GUKUR-JA1-BASE-04	0.7680	0.7924	0.7961	0.7963
GUKUR-JA1-TEST-01	0.8800	0.8771	0.8805	0.8806

Table 10: Detailed results for JA2 (Pairing). The evaluation values in the table were calculated by using a basic NTCIR evaluation tool called 'NTCIREVAL.' See the README file of NTCIREVAL.130507 for more information on each value.

Run ID	MAP	MRR	ERR	RBP
GUKUR-JA2-BASE-01	0.3272	0.3273	0.1637	0.0237
GUKUR-JA2-BASE-02	0.3992	0.4054	0.2029	0.0295
GUKUR-JA2-BASE-03	0.6577	0.6598	0.3303	0.0430
GUKUR-JA2-TEST-01	0.3890	0.3891	0.1946	0.0277
GUKUR-JA2-TEST-02	0.9326	0.9401	0.4705	0.0493

Run ID	O-measure	P-measure	Q-measure	P-plus
GUKUR-JA2-BASE-01	0.3601	0.3601	0.3600	0.3601
GUKUR-JA2-BASE-02	0.4488	0.4488	0.4429	0.4488
GUKUR-JA2-BASE-03	0.7066	0.7066	0.7043	0.7066
GUKUR-JA2-TEST-01	0.4257	0.4257	0.4255	0.4257
GUKUR-JA2-TEST-02	0.9498	0.9498	0.9427	0.9498

Run ID	NCUgu,BR	NCUgu,P	NCUrb,BR	NCUrb,P
GUKUR-JA2-BASE-01	0.3600	0.3272	0.3600	0.3272
GUKUR-JA2-BASE-02	0.4429	0.3992	0.4430	0.3993
GUKUR-JA2-BASE-03	0.7043	0.6577	0.7044	0.6578
GUKUR-JA2-TEST-01	0.4255	0.3890	0.4255	0.3890
GUKUR-JA2-TEST-02	0.9427	0.9326	0.9429	0.9328

Run ID	MAP@1	MAP@10	MAP@100	MAP@1000
GUKUR-JA2-BASE-01	0.2700	0.3135	0.3264	0.3272
GUKUR-JA2-BASE-02	0.3200	0.3909	0.3987	0.3992
GUKUR-JA2-BASE-03	0.5600	0.6524	0.6577	0.6577
GUKUR-JA2-TEST-01	0.3300	0.3768	0.3881	0.3890
GUKUR-JA2-TEST-02	0.9100	0.9315	0.9326	0.9326

Run ID	Hit@1	Hit@10	Hit@100	Hit@1000
GUKUR-JA2-BASE-01	0.2700	0.4100	0.7600	0.9400
GUKUR-JA2-BASE-02	0.3200	0.5800	0.8200	0.9200
GUKUR-JA2-BASE-03	0.5600	0.8700	0.9800	1.0000
GUKUR-JA2-TEST-01	0.3300	0.5100	0.8200	0.9800
GUKUR-JA2-TEST-02	0.9100	0.9800	1.0000	1.0000

Run ID	MSnDCG@1	MSnDCG@10	MSnDCG@100	MSnDCG@1000
GUKUR-JA2-BASE-01	0.2700	0.3366	0.4071	0.4308
GUKUR-JA2-BASE-02	0.3200	0.4354	0.4824	0.4961
GUKUR-JA2-BASE-03	0.5600	0.7022	0.7285	0.7308
GUKUR-JA2-TEST-01	0.3300	0.4084	0.4695	0.4917
GUKUR-JA2-TEST-02	0.9100	0.9427	0.9495	0.9495

Run ID	P@1	P@10	P@100	P@1000
GUKUR-JA2-BASE-01	0.2700	0.0410	0.0077	0.0010
GUKUR-JA2-BASE-02	0.3200	0.0580	0.0084	0.0010
GUKUR-JA2-BASE-03	0.5600	0.0870	0.0102	0.0010
GUKUR-JA2-TEST-01	0.3300	0.0510	0.0082	0.0010
GUKUR-JA2-TEST-02	0.9100	0.0990	0.0104	0.0010

Run ID	Q@1	Q@10	Q@100	Q@1000
GUKUR-JA2-BASE-01	0.2700	0.3338	0.3583	0.3600
GUKUR-JA2-BASE-02	0.3200	0.4270	0.4419	0.4429
GUKUR-JA2-BASE-03	0.5600	0.6945	0.7043	0.7043
GUKUR-JA2-TEST-01	0.3300	0.4024	0.4238	0.4255
GUKUR-JA2-ORCL-02	0.9100	0.9405	0.9427	0.9427

Run ID	nDCG@1	nDCG@10	nDCG@100	nDCG@1000
GUKUR-JA2-BASE-01	0.2700	0.3588	0.4297	0.4529
GUKUR-JA2-BASE-02	0.3200	0.4748	0.5215	0.5350
GUKUR-JA2-BASE-03	0.5600	0.7495	0.7753	0.7775
GUKUR-JA2-TEST-01	0.3300	0.4263	0.4879	0.5094
GUKUR-JA2-TEST-02	0.9100	0.9580	0.9643	0.9643

Run ID	nERR@1	nERR@10	nERR@100	nERR@1000
GUKUR-JA2-BASE-01	0.2700	0.3135	0.3265	0.3273
GUKUR-JA2-BASE-02	0.3200	0.3950	0.4026	0.4030
GUKUR-JA2-BASE-03	0.5600	0.6545	0.6590	0.6590
GUKUR-JA2-TEST-01	0.3300	0.3768	0.3883	0.3890
GUKUR-JA2-TEST-02	0.9100	0.9366	0.9373	0.9373