OKSAT at NTCIR-11 RecipeSearch

- Categorization and Expansion of Search Terms in Topics -

Takashi SATO, Shingo AOKI, Yuta MORISHITA {sato, aoki, morishita}@ss.osaka-kyoiku.ac.jp (Osaka Kyoiku University)

[1] Introduction

- OKSAT submitted five runs for English and Japanese ad hoc recipe search (EN1 and JA1) subtask of NTCIR-11 RecipeSearch.
- For runs of EN1, we tried to categorize search terms of topics.
- We also tried to expand search term for some runs.
- We do not expand search terms of JA1 topics because relatively detail information is obtained from the topic.
- Analyzing experimental results, we observe the effectiveness of our method.

[2] Our Approach

- We searched corpus by the following procedure for English ad hoc recipe search (EN1) and Japanese ad hoc recipe search (JA1), and then we made runs.
  - (1) Extract fields from corpus and made four (EN1) or three (JA1) indices.
  - (2) Prepare search terms from topics to search indices of (1).
  - (3) Score search results of each index (2) using probabilistic model.
  - (4) Merge each scored results into a run.

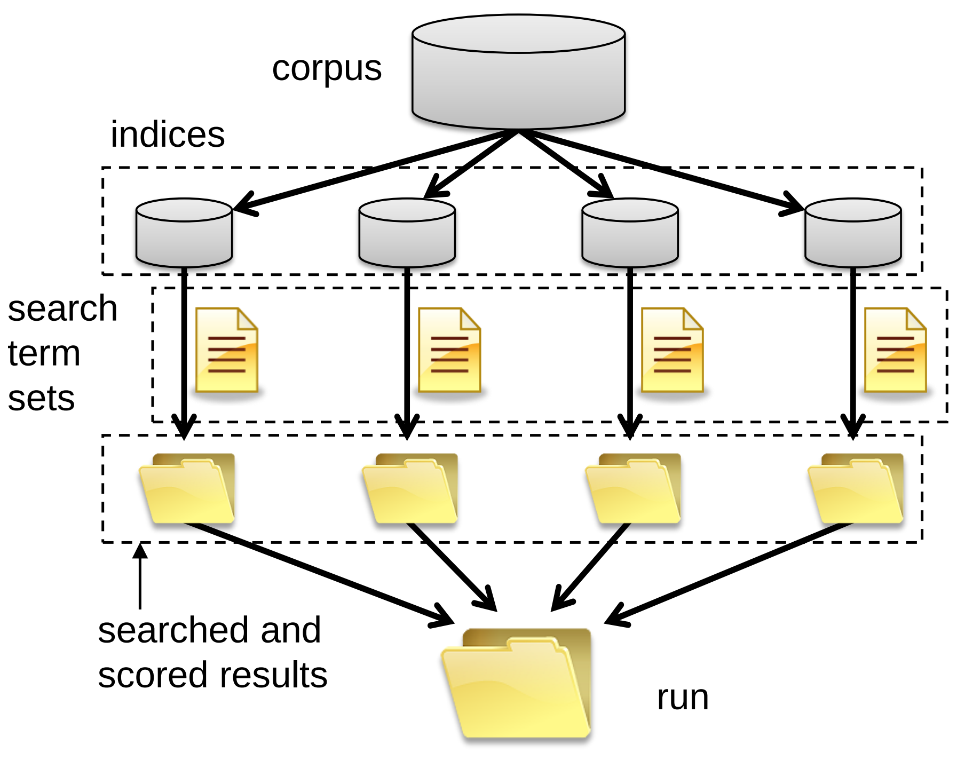

Figure 1. Procedure flow of our approach

[3]EN1 - Indexing

- From title, ingredientLines, preparationSteps and attributes field of English recipe corpus, we made title, ingre, prep and attr index correspondingly.
- We did not use the totalTimeInSeconds field of corpus.
- Indices were gram based, so arbitrary strings search was possible using them.

Table 1. Specifications of computer

CPU	Intel Core i5-4430@3.0GHz 4C/4T
MEM	8GB, DDR3-1600
O S	FreeBSD 8.4, 64bit
HDD	1TB, SATA 6GB/s, 64MB Cache

Table 2. Statistics of EN indices

	title	ingre	prep	attr
data size (MB)	2.77	30.3	64.4	3.19
index size (MB)	9.31	62.8	146	6.91
time (sec.)	1.12	11.5	25.9	.807

[3]EN1 - Categorization of Search Terms

- We made search terms from a topic by the following procedures.
  - (1) Extract words from a topic.
  - (2) Categorize terms into four categories referring our recipe term database.
- The categories are ttl, ing, prp and att intended to search title, ingre, prep and attr index respectively.
- (1) of Figure 2 shows an example.

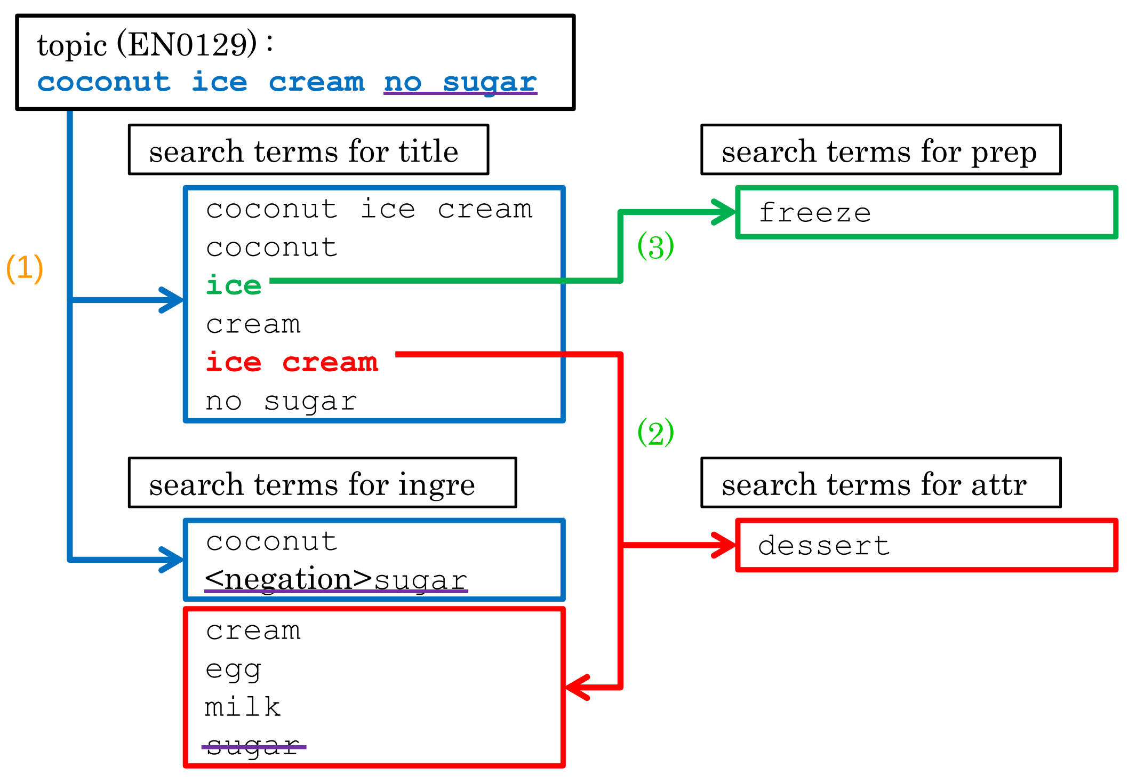

Figure 2. Categorization and expansion of search terms

[3]EN1 - Expansion of Search Terms

- We expanded search terms using words from example answer recipes and/or from the Internet search (Google, Wikipedia, Weblio, etc.).
- (2), (3) of Figure 2 shows an example.

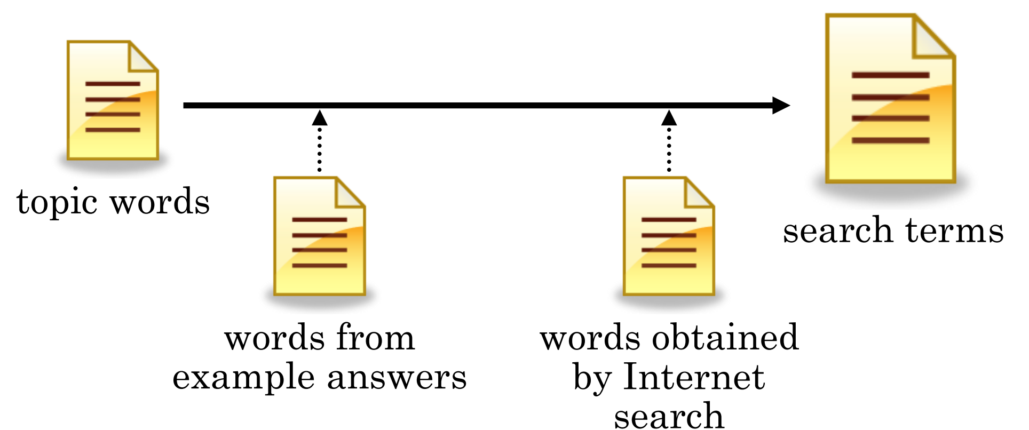

Figure 3. Expansion of search terms

[3]EN1 - Expansion of Search Terms – Cnt'd

Table 3. Part of word expansion list

type	source	expanded words
by grammer	strawberry	strawberries
ttl → ing	bread	flour, baking powder
ing → ing	fruit	apple, lemon, …
ttl → att	cake, …	dessert

[3]EN1 - Searching, Scoring and Merging

- We search four indices (title, ingre, prep, attr) by four search term sets (ttl, ing, prp, att).
- Scoring each of search results using probabilistic model, we got four ranked document list namely title-ttl, ingre-ing, prep-prp and attr-att.
- We multiplied the ranked results by weight of 0.4, 0.4, 0.1, 0.1 in the order, and then we merged them into one list for a run.

Table 4. Time and MAP of submitted EN1 runs

	time (min.)	MAP
OKSAT-EN1-TEST-01	5	0.6790
OKSAT-EN1-TEST-02	8	0.6999
OKSAT-EN1-TEST-03	9	0.7287
OKSAT-EN1-TEST-04	12	0.7499

[3]EN1 - Strength of Search Terms

- In order to enable Boolean type search, our system has the means of document filtering by the term strength defined below.
  - (1) Essential: should have the term
  - (2) Negation: should not have the term
  - (3) Essential + Parallel: at least one of grouped terms should appear in a document
  - (4) Negation + Not Negation: same as Negation if Not Negation terms appear in a document.

[3]EN1 - Statistics of Topic Words

- While processing topics, we observed some characteristics of topic words.
  - (1) Most words relate to title (dish name) and ingredientLines.
  - (2) There are words relate to cooking method (bake, fry, ...), cooking tool (casserole, slow cooker, ...), and manufacturing company.
  - (3) There are words relate to attribute such as season, region (country), time of the day, etc.
  - (4) Few words relate to cooking show, well-known cook, etc.
  - (5) 159 topics out of 500 topics in all have negation expression (... free, ... less, no ..., without ...). Most of them relate to ingrediantLines, however, a expression such as 'no bake' relates to preparetionSteps.

[3]EN1 - Submitted Runs

- We added words from example answer recipes and/or from the Internet search to words from topic categorized.
- We made the following four runs by combinations of these search term sets.
  - OKSAT-EN1-TEST-01: words from topic only
  - OKSAT-EN1-TEST-02: topic + example answer
  - OKSAT-EN1-TEST-03: topic + internet search
  - OKSAT-EN1-TEST-04: topic + example answer +internet search

[3]EN1 - Topic by Topic Analysis

- We show some easy and difficult topics for us.
- (1) Topics in which titles (dishes) and/or ingredients, and/or cook tools are listed are easy. For example the following topics are such type.
  - EN0308: crock pot chicken mushrooms potatoes
  - EN0318: fish sticks without eggs
  - EN0322: baked potato with bacon and cheddar

  We search titles (ingredients, cook tools) by title (ingre, prep) index with strength Essential or Negation.
- (2) Topics in which include low fat, low calorie, etc. are difficult because we don't know these criterion.
  - EN0074: acorn squash low calorie soup
  - EN0118: diabetic low fat low cholesterol
  - EN0218: soba noodle salad low fat
- (3) Topics which have few clues are difficult also.
  - EN0275: asian
  - EN0350: overnight breakfast

  In those cases, we tried to search attr index.

[4]JA1 - Indexing

- From recipe title and dish name fields in recipe_all file, we made title index.
- From material name field of recipe_material file, we made mat index.
- From tag 1, tag 2, tag 3 and tag 4 fields in recipe_all file, we made tag index.

Table 5. Statistics of JA Indices

	title	mat	tag
data size (MB)	19.4	28.4	8.93
index size (MB)	31.9	44.4	12.6
time (sec.)	3.39	5.64	1.57

[4]JA1 - Relations Between Topic Field and Index

- We made the following three search term sets from JA1 topic file.
  - (1) ttl from dishName and negation field
  - (2) mt from foodName
  - (3) tg from negation
- The negation field in the topic was used twice.
- Topic of JA1 consisted of plural fields unlike a case of EN1, we searched indices by corresponding search term sets above.
- We did not expand search terms in JA1 because relatively detail information was obtained from JA1 topic.

[4]JA1 - Searching, Scoring and Merging

- We search three indices (title, mat, tag) by three search term sets (ttl, mt, tg).
- Scoring each of search results using probabilistic model, we got three ranked document list namely title-ttl, mat-mt and tag-tg.
- We multiplied the ranked results by weight of 0.4, 0.4, 0.2 in the order, and then we merged them into one list for a run.

[4]JA1 - Submitted Runs

- As JA1 has no expanded search term sets we prepared different from EN1, we submitted only one run, namely OKSAT-JA1-TEST-01.
- This MAP is obtained using NTCIREVAL and it is the same as the official results for JA1.

Table 6. Time and MAP of submitted JA1 run

	time (min.)	MAP
OKSAT-JA1-TEST-01	19	0.6849

[5] EN1 vs. JA1

- It is difficult to understand questioner's intension because topics of JA1 have plural fields.
- For example, foods listed in food name field in topics should be included or the same in recipes.
- More ad hoc query similar to EN1 may help to compare language by language difference.
- As extensions of JA1 topic, topics which intended to refer 'Standard Tables of Food Composition' might be interesting.

[5] CONCLUSIONS

- OKSAT submitted five runs for English and Japanese ad hoc recipe search (EN1 and JA1) subtask of NTCIR-11 Cooking Recipe Search (RecipeSearch).
- For EN1, while processing of topics, we made a categorization database from topic word and an expansion list for search terms.
- The expansion list was created manually about half of topics, and we tried to use this list to other topics by our expansion program.
- And then we tried to automate categorization and expansion of search terms using them.