Overview of the NTCIR-12 IMine-2 Task

Takehiro Yamamoto[1], Yiqun Liu[2], Min Zhang[2], Zhicheng Dou[3], Ke Zhou[4],
Ilya Markov[5], Makoto. P. Kato[1], Hiroaki Ohshima[1], Sumio Fujita[5]
[1]Kyoto University, [2]Tsinghua University, [3]Renmin University of China,
[4]Yahoo! Labs, [5]University of Amsterdam, [6]Yahoo! Japan Corporation
{tyamamot, kato, ohshima}@dl.kuis.kyoto-u.ac.jp; {yiqunliu, z-m}.tsinghua.edu.cn;
{ douzhicheng, zhouke.nlp}@gmail.com, i.markov@uva.nl, sufujita@yahoo-corp.jp

ABSTRACT

In this paper, we provide an overview of the NTCIR-12 IMine-2 task, which is a core task of NTCIR-12 and also a succeeding work of IMine@NTCIR-11, INTENT-2@NTCIR-10, and INTENT@NTCIR-9 tasks. IMine-2 comprises the Query Understanding subtask and the Vertical Incorporating subtask. 23 groups from diverse countries including China, France, India, Portugal, Ireland, and Japan registered to the task. Finally, IMine-2 attracted 9 participating teams; we received 42 runs for the Query Understanding subtask and 15 runs for the Vertical Incorporating subtask. We describe the subtasks, data, evaluation methods, and report the official results for each subtask.

1. INTRODUCTION

The goal of the IMine-2 task, which is a core task of NTCIR-12 and also a succeeding task of the IMine [9], INTENT-2 [6], and INTENT [8] tasks, is to explore and evaluate the technologies of understanding user intents behind the query. Many queries issued by users are short and ambiguous in Web search. Even though two users issue the same query their search intents would be different. Recently, query understanding and search result diversification, which aim at satisfying different user intents behind a Web search query, attracted both IR communities and commercial search engines. The IMine task aims at providing common dataset and evaluation methodology to researchers working for this research area.

In IMine-2, taking over the basic task designs in the IMine-1 task, we focus on vertical intents behind a query as well as its topical intents. Nowadays, many commercial Web search engines merge several types of search results and generate a SERP (search engine results page) in response to a user's query. For example, the results of query "flower" now may contain image results and encyclopedia results as well as usual Web search results. We refer to such "types" of search results as verticals. Many researchers as well as commercial search engines have been focusing on predicting and evaluating appropriate vertical resources for a query [2][3][5].

The IMine-2 task comprises the two subtasks: the Query Understanding subtask and the Vertical Incorporating subtask. The Query Understanding subtask is a successive task of the Subtopic Mining subtask, which was held in the IMine and INTENT tasks. The difference from the past Subtopic Mining subtask is that the participants are asked to identify the relevant verticals for each subtopic. The Vertical Incorporating subtask is also a successive task of the Document Ranking subtask in the past tasks. The difference from the past Document Ranking subtask is that the participants should decide whether the result list should contain vertical results (See Section 2 for the detailed task descriptions).

Table 1 summarizes the differences between IMine-2 and the previous IMine task. Just like the IMine task, we involve dealing with three different languages including English, Chinese and Japanese in the IMine-2 task. One difference other than vertical is that we include more topics than those in the IMine task. A recent study by Sakai [7] suggests that we need to increase the number of topics to guarantee significant differences among runs in terms of D#-nDCG, which was used as the primary metric in the IMine task. To make our test collection more reliable and reusable, we include more topics while reducing the size of pool depth, which is also recommended in [7].

Table 1. Differences between IMine and IMine-2 tasks.

	IMine@NTCIR-11	**IMine-2@NTCIR-12**
# of topics	Chinese: 50 English: 50 Japanese: 50	Chinese: 100 English: 100 Japanese: 100
Query types	Ambiguous Broad Very clear	Ambiguous Faceted Very clear Task-oriented Vertical-oriented
	Subtopic Mining	**Query Understanding**
Language	English Chinese Japanese	English Chinese Japanese
Subtopics	Two-level subtopics	First-level subtopics
Vertical intents	No	Yes
Pool depth	5 (first-level) 10(second-level)	10
	Document Ranking	**Vertical Incorporating**
Language	English Chinese	English Chinese
Pool depth	20	10

23 groups from China, France, India, Portugal, Ireland, and Japan registered to the IMine task. Finally, we received 42 runs from 9 teams for the Query Understanding subtask and 15 runs from two teams for the Vertical Incorporating subtask. Tables 2 and 3 summarize the number of runs and participating teams for each subtask.

The reminder of the paper is organized as follows. Section 2 describes the details of the two subtasks. Section 3 describes the data provided to the participant, including the query topics, document collection, and other resources. Section 4 explains the evaluation strategy and metrics used in the IMine-2 task. Section 5 reports the official results. Finally, Section 6 concludes this paper.

Table 2. Organization of participating groups in IMine-2.

GroupID	Organization
IMC	Beijing Institute of Technology, China
rucir	Renmin University of China, China
HUKB	Hokkaido University, Japan
IRCE	University of Tsukuba, Japan
KDEIM	Toyohashi University of Technology, Japan
THUIR	Tsinghua University, China
YJST	Yahoo Japan Corporation, Japan
HLT01	Université de Caen Normandie, France
NEXTI	Hiroshima City University, Japan

Table 3. Statistics of result submissions.

GroupID	Query Understanding			Vertical Incorporating	
	English	Chinese	Japanese	English	Chinese
IMC		5			
rucir	5	5		5	5
HUKB			5		
IRCE		1	5		
KDEIM	4				
THUIR		5			5
YJST			5		
HULTECH	1				
NEXTI			1		
#Group	3	4	4	1	2
#Run	10	16	16	5	10

2. SUBTASKS

The IMine-2 task comprises the Query Understanding subtask and the Vertical Incorporating subtask. This section first explains the input and output of the two subtasks and then explains several concepts important in the IMine-2 task.

2.1 Query Understanding Subtask

The Query Understanding subtask is defined as follows: given a query, the participant is required to generate a diversified ranked list of not more than 10 subtopics with their relevant vertical intents. In the Query Understanding subtask, a subtopic of a given query is viewed as a search intent that specializes and/or disambiguates the original query. The participants are expected to (1) rank important subtopics higher, (2) cover as many intents of a given query as possible, and (3) predict a relevant vertical for each subtopic.

This subtask corresponds to the Subtopic Mining subtask in the IMine, INTENT-2 and INTENT tasks. The difference from the previous subtask is that participants are also required to identify the relevant vertical for each subtopic. In other words, for a given query, the participants have to identify its important subtopics and which vertical should be presented for the subtopic.

For example, for the query "iPhone 6", a possible result list of the Query Understanding subtask is:

```
[tid]           [subtopic]          [vertical]  [score]
IMINE2-E-000    iPhone 6 apple.com  Web         0.98
IMINE2-E-000    iPhone 6 sales      News        0.90
IMINE2-E-000    iPHone 6 photo      Image       0.88
IMINE2-E-000    iPhone 6 review     Web         0.78
```

where *tid* is a topic ID, *subtopic* is a string that the system generates as a subtopic, *vertical* is an estimated vertical relevant to the subtopic, *score* is an estimated importance of the subtopic. For *vertical*, the system must pick up one vertical out of six verticals defined for each language (See Section 2.4 for the available verticals for each language). For example, for the English Query Understanding subtask, a vertical intent should be "Web", "Image", "News", "QA", "Encyclopedia" or "Shopping". Note that we did not use *score* values for our evaluation and use only the order of subtopics and their vertical intents; the ranks of the subtopics were determined just by their appearance orders in the submission file.

In the Query Understanding subtask, we accepted the following two types of runs:

- **Q-Run**: Runs for the regular Query Understanding subtask; systems are required to identify both subtopics and relevant verticals for given topics.
- **S-Run**: Optional runs designed for those who wants to focus on the subtopic mining; systems are required to identify subtopics, but not vertical intents.

Among 42 runs submitted to the Query Understanding subtask, 31 runs were submitted as Q-Run and 11 runs were submitted as S-Run.

2.2 Vertical Incorporating Subtask

In the Vertical Incorporating subtask, given a query and the document collection, the system is required to return a diversified ranked list of not more than 100 results. The objective of the ranking is to (1) rank documents relevant to important intents higher, (2) rank vertical results (defined as virtual documents) relevant to important intent higher, and (3) cover as many intents as possible.

This subtask corresponds to the Document Ranking (DR) subtask in the IMine, INTENT-2 and INTENT tasks. The difference from the previous subtask is that the participants should decide whether the result list should contain certain types of vertical results. For this purpose, the participants can include *virtual documents* as well as organic documents in their ranking.

A *virtual document* is a special document that represents a search result generated from the vertical. More specifically, for English subtask, the participants could use the following five virtual documents:

- Vertical-Image
- Vertical-News
- Vertical-QA
- Vertical-Encyclopedia
- Vertical-Shopping

For Chinese subtask, the participants could use the following five virtual documents:

- Vertical-Image
- Vertical-News
- Vertical-Download
- Vertical-Encyclopedia
- Vertical-Shopping

A virtual document of a vertical is assumed to be an ideal search result generated by the vertical and always relevant if and only if its vertical is relevant to one of the intents behind the query. By using the document collections and virtual documents, the participants have to decide which virtual documents should be ranked higher while keeping the diversity of the ranking.

For example, a possible result list for the Vertical Incorporating subtask is:

[tid]	*[did]*	*[score]*
IMINE2-E-000	*IMINE-E-000-013.html*	*0.78*
IMINE2-E-000	*Vertical-News*	*0.70*
IMINE2-E-000	*Vertical-Image*	*0.60*
IMINE2-E-000	*IMINE-E-000-113.html*	*0.50*

where *tid* is a topic ID, *did* is either a document ID in the document collection or a virtual document ID, *score* is an estimated importance of the document. Note that we did not use *score* values for evaluation, and used only the order of documents in the evaluation; the ranks of the documents were determined just by their appearance orders in the submission file.

2.3 Subtopics and Intents

In the Understanding subtask, participants were required to return a ranked list of subtopics, not a ranked list of document IDs. We provided the following instruction on the IMine-2 homepage.

A subtopic of a given query is a query that specializes and/or disambiguates the search intent of the original query. If a string returned in response to the query does neither, it is considered irrelevant.

e.g.

original query*: "jaguar" (ambiguous)*

subtopic: "jaguar car brand" (disambiguate)

*incorrect: "*jaguar *jaguar" (does not disambiguate; does not specialize)*

e.g.

original *query: "harry potter" (underspecified)*

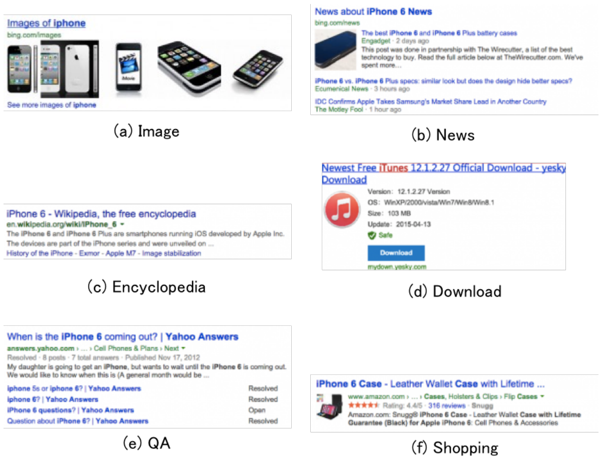
*sub*topic: "harry potter movie" (specialize)

incorrect: "harry potter hp" (does not specialize; does not disambiguate)

The submitted subtopics are clustered into several clusters so as to form a set of intents, which represents the possible search intents for a query. (See Section 4.3)

2.4 Verticals

Nowadays, many commercial Web search engines merge several types of search results and generate a SERP (search engine results page) in response to a user's query. For example, the results of query "flower" now may contain image results and encyclopedia results as well as usual Web search results. We refer to such "types" of search results as verticals. For example, "image", "news" can be a vertical. Figure 1 shows the typical representation of each vertical in a SERP.



Images of iphone
bing.com/images
See more images of iphone

(a) Image

News about iPhone 6 News
bing.com/news
The best iPhone 6 and iPhone 6 Plus battery cases
Engadget · 2 days ago
This post was done in partnership with The Wirecutter, a list of the best technology to buy. Read the full article below at TheWirecutter.com. We've spent more…
iPhone 6 vs. iPhone 6 Plus specs: similar look but does the design hide better specs?
Ecumenical News · 3 hours ago
IDC Confirms Apple Takes Samsung's Market Share Lead in Another Country
The Motley Fool · 1 hour ago

(b) News

iPhone 6 - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/IPhone_6
The iPhone 6 and iPhone 6 Plus are smartphones running iOS developed by Apple Inc. The devices are part of the iPhone series and were unveiled on ...
History of the iPhone - Exmor - Apple M7 - Image stabilization

(c) Encyclopedia

Newest Free iTunes 12.1.2.27 Official Download - yesky Download
Version: 12.1.2.27 Version
OS: WinXP/2000/vista/Win7/Win8/Win8.1
Size: 103 MB
Update: 2015-04-13
Safe
Download
mydown.yesky.com

(d) Download

When is the iPhone 6 coming out? | Yahoo Answers
answers.yahoo.com › ... › Cell Phones & Plans › Next
Resolved · 8 posts · 7 total answers · Published Nov 17, 2012
My daughter is going to get an iPhone, but wants to wait until the iPhone 6 is coming out. We would like to know when this is (A general month would be ...
iphone 5s or iphone 6? | Yahoo Answers Resolved
iphone 6? | Yahoo Answers Resolved
iPhone 6 questions? | Yahoo Answers Open
Question about iPhone 6? | Yahoo Answers Resolved

(e) QA

iPhone 6 Case - Leather Wallet Case with Lifetime ...
www.amazon.com › ... › Cases, Holsters & Clips › Flip Cases
★★★★★ Rating: 4.4/5 · 316 reviews · Snugg
Amazon.com: Snugg® iPhone 6 Case - Leather Wallet Case with Lifetime Guarantee (Black) for Apple iPhone 6: Cell Phones & Accessories

(f) Shopping

Figure 1. Typical representation of each vertical in SERP.

In IMine-2, we selected six verticals for each of Japanese, Chinese, and English topics so that we could pick up the popular verticals for different countries. More specifically, we considered the following verticals:

- **English Query Understanding and Vertical Incorporating subtasks:**
  - Web
  - Image
  - News
  - QA
  - Encyclopedia
  - Shopping
- **Chinese Query Understanding and Vertical Incorporating subtasks:**
  - Web
  - Image
  - News
  - Download
  - Encyclopedia
  - Shopping
- **Japanese Query Understanding subtask:**
  - Web
  - Image
  - News
  - QA
  - Encyclopedia
  - Shopping

Relevant verticals depend on the intents behind a query. For a user who searches for "iPhone 6 photo," for example, the image vertical might be much more relevant than usual Web search results. A *vertical intent* is defined as a preference on verticals for a given intent. In Query Understanding subtask, the participants were required to identify relevant vertical intent for each subtopic.

3. DATA

This sections describes the query topics, document collection and other resources provided to the IMine-2 participants.

3.1 Topics

The same query topics were adopted in both Query Understanding subtask and the Vertical Incorporating subtask for all languages.

100 queries were prepared for each of the languages. Similar to the IMine and INTENT tasks, the topics are sampled from the median-frequency queries collected from AOL, Sogou and Bing search engine logs. Five types of queries, namely, ambiguous, faceted, very clear, task-oriented, and vertical-oriented, were included in the query topic set so that we could investigate the performances of different algorithms with diverse queries. The details of the five query types are as follows:

- **Ambiguous**: The concepts/objects behind the query are ambiguous (e.g., “jaguar” -> car, animal, *etc*).
- **Faceted**: The information needs behind the query include many facets or aspects (e.g., “harry potter” -> movie, book, Wikipedia, *etc*).
- **Very clear**: The information need behind the query is very clear so that usually a single relevant document can satisfy his information needs.(e.g., “apple.com homepage”)
- **Task-oriented**: The search intent behind the query relates the searcher’s goal (e.g., “lose weight” -> exercise, healthy food, medicine, *etc*).
- **Vertical-oriented**: The search intent behind the query strongly indicates a specific vertical (e.g., “iPhone photo” -> Image vertical).

The differences from the IMine task is we included the task-oriented and vertical-oriented queries as our topics. As for the task-oriented queries, we decided to include them since many researchers recently have been studying on understanding of tasks behind a user’s query [10] as in the TREC2015 Tasks track[1] and IMine TaskMine subtask. As for the vertical-oriented query, we included them so that we could guarantee that several queries highly indicate the specific verticals rather than usual Web search results.

Several topics are also shared among different languages for possible future cross-language research purposes. Table エラー! ブックマークが定義されていません。 summarizes the statistics of the query topics in the IMine-2 task. Tables in Appendix A shows the complete list of the query topics used in the IMine-2 task. As for the Query Understanding subtask, queries with very clear intents were not evaluated because they are not expected to contain subtopics.

Table 4. Statistics of IMine-2 query topics.

Language	Query types				
	Ambiguous	Faceted	Very Clear	Task-oriented	Vertical-oriented
English	24	24	3	24	25
Chinese	9	19	9	20	43
Japanese	25	25	0	25	25

3.2 Document Collection

Unlike the past IMine and INTENT tasks, we provided the document collection designed for the IMine-2 task. The document collection, which we call the IMine-2 Web corpus, contains the top 500 ranked documents that were returned by the Bing Web search API[2] in response to each query. This crawling was conducted from July 1st to August 17th 2015. As we failed to access some of the documents, the number of crawled documents per query is fewer than 500.

The participants were asked to use the IMine-2 Web corpus for generating a ranked-list for the Vertical Incorporating subtask.

3.3 Other Resources

The following data was provided to the participants so that the participants can predict/mine intents for a given query. Also, we encouraged the participants to use other external resources for their runs on both the Query Understanding and Vertical Incorporating subtasks.

- **Web Search Related Query Data from Yahoo! JAPAN** (for Japanese subtask): This dataset is generated from the query log of Yahoo! Japan Search from July 2009 to June 2013[3].
- **SogouQ search user behavior data** (for Chinese subtask): The collection contains queries and click-through data collected and sampled in November, 2008 (consistent with SogouT). A new version of SogouQ is also available now which is a sample of data collected in 2012. Further information regarding the data can be found on the page http://www.sogou.com/labs/dl/q.html.
- **Query suggestions/completions of several commercial search engines** (for Chinese, English, Japanese subtasks): A list of query suggestions/completions collected from popular commercial search engines such as Google, Yahoo!, Bing, Baidu are provided as possible subtopic candidates.

4. EVALUATION METRICS

This section first explains the evaluation metrics used for the Query Understanding and Vertical Incorporating subtasks. It then explains how we construct the ground truth data.

4.1 Query Understanding Subtask

In the QU subtask, the quality of the participants’ runs are evaluated based on both the *diversity of intents* and the *accuracy of vertical intent prediction*.

The diversity of intents is measured by D#-measure [4], which was proposed by Sakai *et al.*, and also used in the IMine and INTENT tasks. The purpose of D#-measure is to intuitively evaluate a ranked-list in terms of both its diversity and relevance. Let I be the set of known intents for given query q. For each $i \in I$, let $P(i|q)$ denote its *intent probability* and let $g_i(r)$ be the gain value of the subtopic at rank *r* with respect to intent *i*, which we defined as 1 if the subtopic belongs to intent *i* and 0 otherwise. The global gain for this *r*-th ranked subtopic is defined as:

$$\mathrm{GG}(r) = \sum_i P(i|q) g_i(r)$$

The “globally ideal” ranked list of subtopics is obtained by sorting all relevant subtopic by the global gain. Let $GG^*(r)$ denote the global gain in this ideal list. D-nDCG at cufoff *l* is defined as:

$$\text{D-nDCG}@l = \frac{\sum_{r=1}^{l} GG(r)/\log{(r+1)}}{\sum_{r=1}^{l} GG^*(r)/\log{(r+1)}} .$$

Let $I'(\subseteq I)$ be the set of intents covered by a ranked list. Then the recall of intents I-rec is defined as:

[1] http://www.cs.ucl.ac.uk/tasks-track-2015/

[2] https://datamarket.azure.com/dataset/bing/search

[3] http://research.nii.ac.jp/ntcir/news-20150717-ja.html

$$\text{I-rec} = \frac{|\mathrm{I}'|}{|\mathrm{I}|} \ .$$

While D-nDCG measures an overall relevance in terms of all the possible intents, I-rec measures the number of intents covered by the ranked list. D#-nDCG@l is computed as a linear combination of D-nDCG@l and I-rec:

$$\text{D\#-nDCG} = \gamma\text{I-rec} + (1-\gamma)\text{D-nDCG} \ \ ,$$

where we let $\gamma = 0.5$ throughout the paper, as in the past IMine and INTENT tasks.

As for the accuracy of vertical intent prediction, we employed the simple metric since it is the first trial in the NTCIR tasks to incorporate the accuracy of vertical intent prediction. Let V be the set of available verticals. For each $v \in V$, let $P(v|i)$ denote the importance of vertical v with respect to intent i. The accuracy of the vertical intent prediction of the r-th ranked subtopic is defined as:

$$\text{Accuracy}(r) = \frac{P(v_r|i_r)}{\max_{v \in V} P(v|i_r)} \ ,$$

where v_r denotes the predicted vertical of the r-th ranked subtopic and $i_r$denotes the intent to which the r-th ranked subtopic belongs. Note that Accuracy(r) becomes 0 if the r-th ranked subtopic is irrelevant.

Having the above equation, V-score@l, which measures the accuracy of vertical intent prediction for a ranked list of subtopics at cutoff l, is computed as:

$$\text{V-score@}l = \frac{1}{l}\sum_{r=1}^{l} \text{Accuracy}(r) \ .$$

Finally, we linearly combine D#-nDCG and V-score. The definition of QU-score, which is used as the main evaluation metric for the Query Understanding subtask, is as follows:

$$\text{QU-score} = \lambda\text{D\#-nDCG@}l + (1-\lambda)\text{V-score@}l$$

where we use $l = 10$ and $\lambda = 0.5$ throughout the paper.

4.2 Vertical Incorporating Subtask

As for the Vertical Incorporating subtask, D#-nDCG@l is also used to measure whether the system can generate a diversified ranked list. The difference from the usual D#-measure is we consider the importance of a vertical to compute a gain value of a document. Let $g_i(d)$ be the gain value of document d with respect to intent i, $g_i(d)$ is defined as:

$$g_i(d) = \sum_{v \epsilon V} \delta_v(d) P(v|i) \text{rel}_i(d) \ ,$$

where $\delta_v(d)$ is an indicator that if the type of the vertical of document d is v, $\delta_v(d)$ is 1; otherwise 0. Note that the vertical type of non-virtual documents (i.e. ones from the IMine-2 Web corpus) is regarded as "Web". $\text{rel}_i(d)$ is the relevance of document d with respect to intent i. The range of $\text{rel}_i(d)$ is { 0 (irrelevant), 1 (relevant), 2 (highly relevant) }. Note that, as for the virtual documents, their relevances are assumed to be highly relevant. The D#-nDCG@l for the Vertical Incorporating subtask can be computed by replacing gain value $g_i(r)$ in the Query Understanding subtask with $g_i(d)$.

4.3 Ground Truth Construction

This subsection describes the assessment procedures to construct the ground truth data. All the assessments were completed by the assessors hired in Kyoto University. For both Japanese and Chinese subtasks, the assessments were completed with the native speakers. For the English subtask, the assessors who have sufficient English skills were hired for completing the assessments.

4.3.1 Query Understanding Subtask

For the Query Understanding subtask, the queries except for the very clear ones were annotated by the assessors. The annotation process for the Query Understanding subtask is completed in the following steps:

- **Result pooling**: The submitted runs were first pooled for the later annotation process. The result pool of the English subtask contained 2,503 subtopics. The result pool of the Chinese subtask contained 6,119 subtopics. The result pool of the Japanese subtask contained 6,422 subtopics.

- **Clustering subtopics into intents**: For each topic, the assessors were asked to cluster them into several clusters. These clusters are regard as intents for a query. This clustering assessments were done by the clustering interface as shown in Figure 2.

- **Importance voting:** Having clustered subtopics (i.e., *intents*), we asked five assessors to individually judge whether each intent is important or not with respect to the topic. After the annotation, we selected the TEN most important ones and obtained their intent probabilities $P(i|q)$ by normalizing the number of votes for each intent by the total number of votes of the TEN most important intents. Note that the intents that were not included in the TEN most important ones were regarded as irrelevant when computing the evaluation metrics.

- **Vertical importance voting**: For each of the TEN most important intent, we asked five assessors to judge whether each vertical is important. The assessors were asked to judge their importance with a 3-grade score; 0 (irrelevant), 1 (relevant) and 2 (highly relevant). We finally obtained $P(v|i)$, the importance of vertical v with respect to intent i, by normalizing the scores.

4.3.2 Vertical Incorporating Subtask

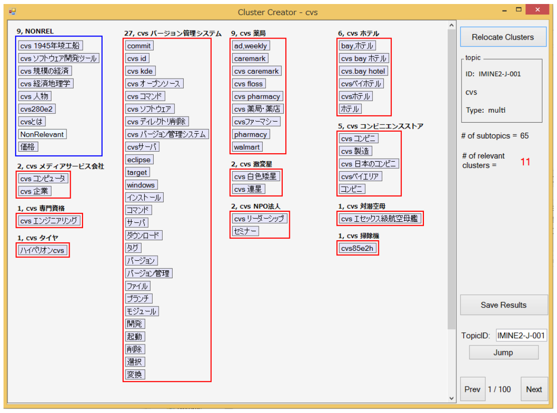
For the Vertical Incorporating subtask, all the queries including the very clear ones were assessed by the assessors to obtain the document relevance. Note that, in the Vertical Incorporating subtask, we only use the top FIVE intents to assess their per-intent document relevance while we use the top TEN intents in the Query Understanding subtask. One reason why we use the top five intents is the results of the IMine task suggested that the five intents were enough to evaluate the diversified results. Another reason is to reduce our assessment cost.

Document relevance assessments were completed via the developed Web interface shown in Figure 3. The annotation process for the Vertical Incorporating subtask is completed in the following steps:

- **Result pooling**: The submitted runs were first pooled for the later annotation process. In IMine-2, the pool depth size was set to 10. The result pool of the English subtask contained 5,564 documents. The result pool of the Chinese subtask contained 6,788 documents.

- **Per-topic relevance judgment:** For each document-query pair, the assessors were asked to judge whether the document is relevant with respect to the query with a 4-grade score (2: highly relevant, 1: relevant, 0: irrelevant, -1: spam).
- **Per-intent relevant judgment:** For each document-intent pair for the queries except for very clear ones, the assessors were asked to judge whether the document is relevant to the intent with a 3-grade score (2: highly relevant, 1: relevant, 0: irrelevant).

With the above procedure, we obtained the document relevance both to queries and their intents. For very clear queries, the original nDCG score is calculated as the evaluation result.



topic

ID: IMINE2-J-001

cvs

Type: multi

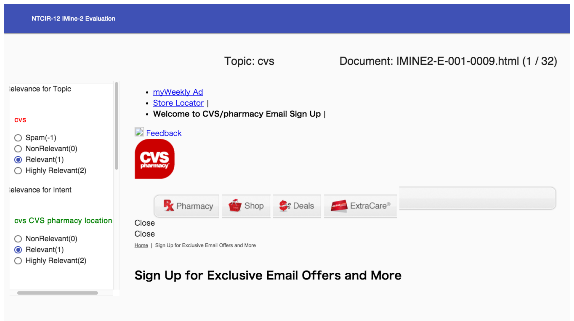
# of subtopics = 65

# of relevant clusters = 11

TopicID: IMINE2-J-001

1 / 100

Figure 2. Clustering tool developed in the INTENT-2 task. By using the tool, assessors can (1) judge whether the subtopics are non-relevant or not, (2) cluster relevant subtopics into clusters, and (3) assign intent label to cluster.



Topic: cvs

Document: IMINE2-E-001-0009.html (1 / 32)

ıelevance for Topic

cvs

○ Spam(-1)
○ NonRelevant(0)
● Relevant(1)
○ Highly Relevant(2)

ıelevance for Intent

cvs CVS pharmacy location:

○ NonRelevant(0)
● Relevant(1)
○ Highly Relevant(2)

- myWeekly Ad
- Store Locator |
- Welcome to CVS/pharmacy Email Sign Up |

Close
Close
Home | Sign Up for Exclusive Email Offers and More

Sign Up for Exclusive Email Offers and More

Figure 3. Developed Web interface for document relevance annotation.

5. EVALUATION

In this section, we present the official evaluation results of the IMine-2 task. We first report the results of the Query Understanding subtask. We then report the results of the Vertical Incorporating subtask. We used ntcireval[4] developed by Sakai to compute I-rec@10, D-nDCG@10, and D#-nDCG@10. The two-sided randomized Tukey's HSD test at the significant level $\alpha = 0.05$ was applied to the results to find significantly different run pairs. We also used discpower [11] developed by Sakai to conduct the statistical tests.

5.1 Japanese Query Understanding subtask

Figures 4, 5, and 6 show the mean I-rec@10, D-nDCG@10, and D#-nDCG@10 performances of the Japanese Query Understanding subtask runs. The significantly different run pairs are also reported in Appendix B. As described in Section 4.1, these metrics measure a subtopic quality returned by the algorithms. It can be observed that (a) NEXTI-Q-J-1Q is the top performer for all the metrics, and (b) HUKB-Q-J-4Q is the second best performer in terms of the intent recall (i.e. I-rec@10), while YJST-Q-J-1Q achieves the second best performance in terms of the subtopic relevance (i.e. D-nDCG@10). Although NEXTI-Q-J-1Q achieves the best performance in terms of D#-nDCG, we found no significant differences among NEXTI-Q-J-1Q, HUKM-Q-J-4Q, and YJST-Q-J-1Q. Figure 6 shows the I-rec/D-nDCG graph. From the figure, we can see that there is the strong correlation between I-rec@10 and D-nDCG@10.

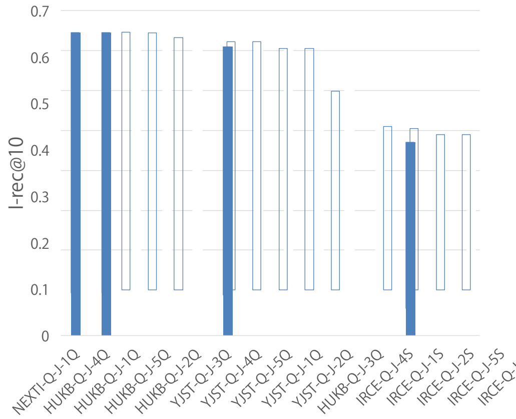


Figure 4. I-rec@10 for unclear topics in Japanese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

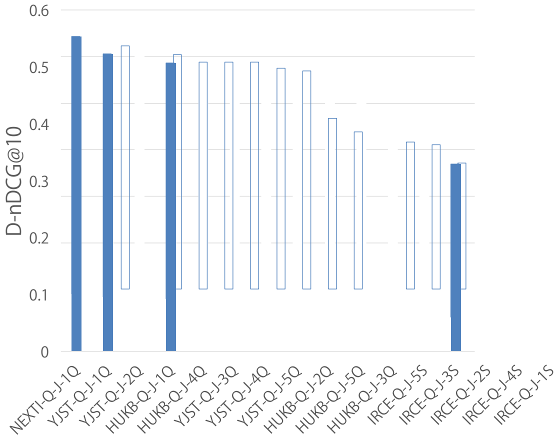


Figure 5. D-nDCG@10 for unclear topics in Japanese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

[4] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html

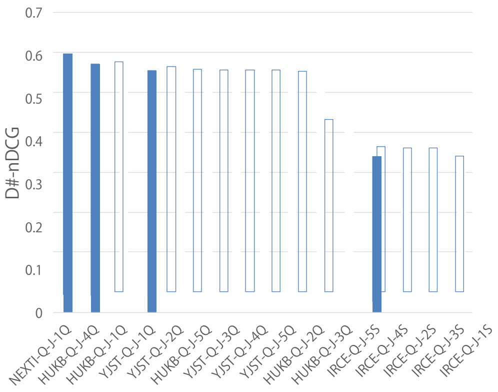


Figure 6. D#-nDCG@10 for unclear topics in Japanese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

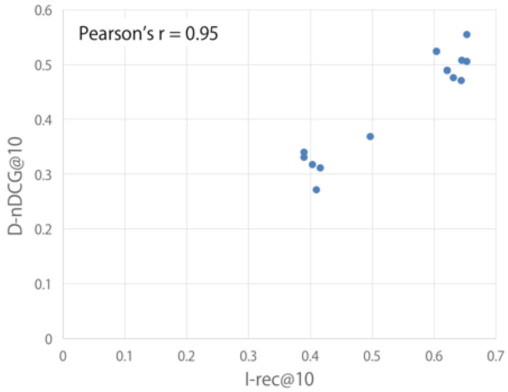


Figure 7. I-rec@10/D-nDCG@10 graph for Japanese Query Understanding.

Figures 8 and 9 show the mean V-score@10 and QU-score performances of the Japanese Query Understanding subtask runs. Note that only the Q-Run runs are evaluated. The significantly different run pairs are also reported in Appendix B. From the figures, we can see that NEXTI-Q-J-1Q again achieves the best performance in terms of V-score and QU-score. Further, the differences between NEXTI-Q-J-1Q and the other runs are significantly different in terms of both V-score and QU-score. Figure 10 shows the V-score/D#-nDCG graph. From the figure we can see that the correlation between V-score and D#-nDCG is smaller than that between I-rec and D-nDCG. The result indicates that, to achieve a high V-score performance, we need to take an approach different from that achieving a high D#-nDCG performance. Figure 11 shows the per-topic Max/Average QU-score performances of the Japanese Query Understanding subtask runs.

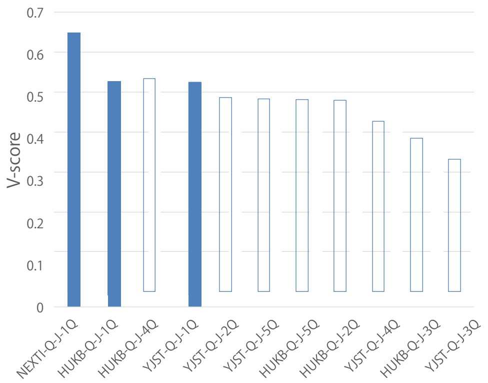


Figure 8. V-score@10 for unclear topics in Japanese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

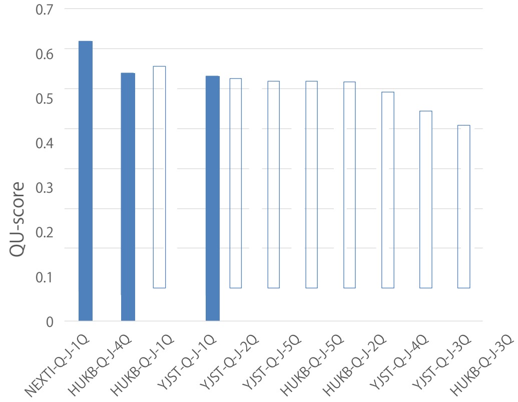


Figure 9. QU-score (official measure) for unclear topics in Japanese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

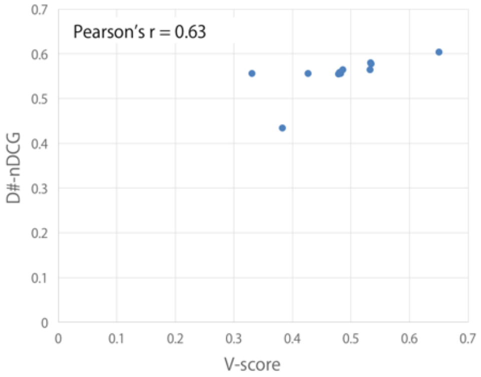


Figure 10. V-score@10/D#-nDCG@10 graph for Japanese Query Understanding.

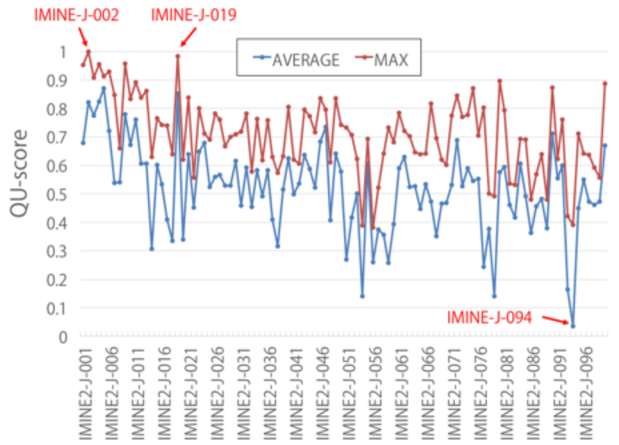


Figure 11. Per-topic QU-score performances for Japanese Query Understanding.

5.2 English Query Understanding subtask

Figures 12, 13, and 14 show the mean I-rec@10, D-nDCG@10, and D#-nDCG@10 performances of the English Query Understanding subtask runs. The significantly different run pairs are also reported in Appendix B. Figure 15 shows the corresponding I-rec/D-nDCG graph. From the figures, we found that (a) rucir-Q-E-4Q achieves the best performance in terms of I-rec@10, (b) HULTECH-Q-E-1Q is the top performer in terms of i.e. D-nDCG; and (c) KDEIM-Q-E-1S is the overall winner in terms of D#-nDCG@10. However, the differences between these three runs are not statistically significant.

Figure 16 and 17 show the mean V-score and QU-score performances of the English Query Understanding subtask runs. The significantly different run pairs are also reported in Appendix B. From the figures, we can see that rucir-Q-E-5Q, which is the third performer in terms of D#-nDCG, achieves the best performance in terms of V-score and QU-score. Further, the differences between rucir-Q-E-5Q and the other runs are statistically significant in terms of both V-score and QU-score. From the figure 18, unlike the Japanese and Chinese Query Understanding subtask, we found that the correlation between V-score and D#-nDCG is quite low. Finally, Figure 19 shows the per-topic QU-score for English Query Understanding subtask.

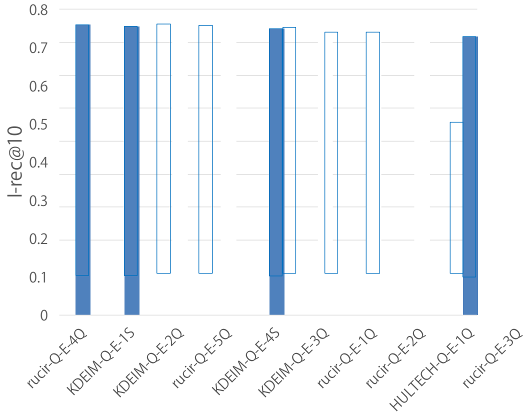


Figure 12. I-rec@10 for 97 unclear topics in English Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

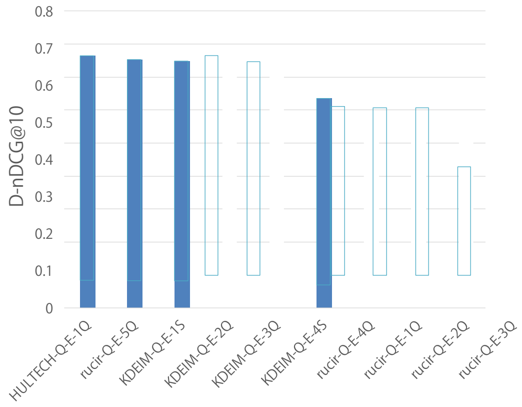


Figure 13. D-nDCG@10 for 97 unclear topics in English Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

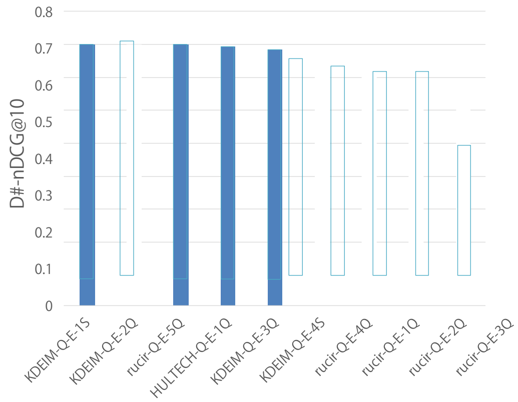


Figure 14. D#-nDCG@10 for 97 unclear topics in English Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

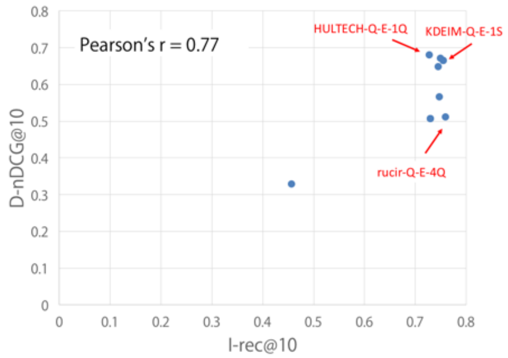


Figure 15. I-rec@10/D-nDCG@10 graph for English Query Understanding.

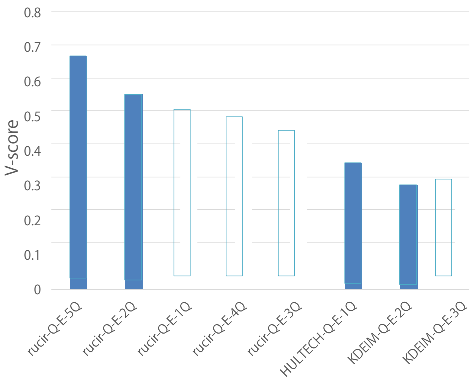


Figure 16. V-score@10 for 97 unclear topics in English Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

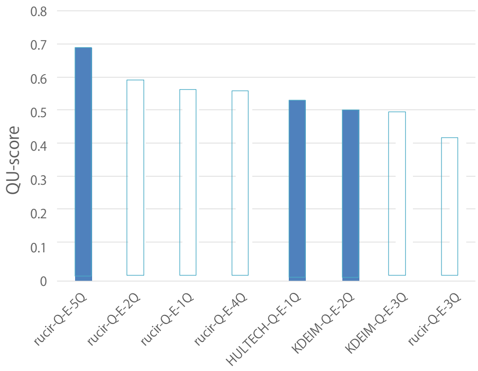


Figure 17. QU-score (official measure) for 97 unclear topics in English Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

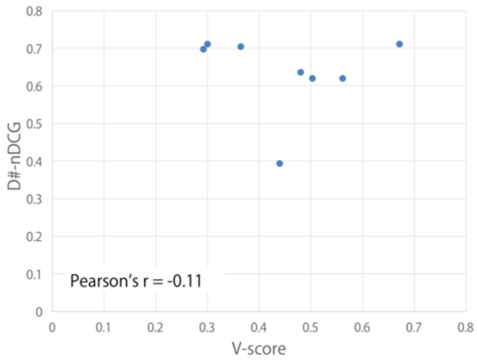


Figure 18. V-score@10/D#-nDCG@10 graph for English Query Understanding.

5.3 Chinese Query Understanding subtask

Figures 20, 21, and 22 show the mean I-rec@10, D-nDCG@10, and D#-nDCG@10 performances of the Chinese Query Understanding subtask runs. The significantly different run pairs are also reported in Appendix B. Figure 23 shows the corresponding I-rec/D-nDCG graph. From the results, we found that (a) thuir-Q-C-3Q is the top performer in terms of I-rec@10, and (b) rucir-Q-C-5Q achieves the best in terms of D-nDCG@10 and D#-nDCG@10. However, there is no significant difference between thuir-Q-C-3Q and rucir-Q-C-5Q in terms of D#-nDCG.

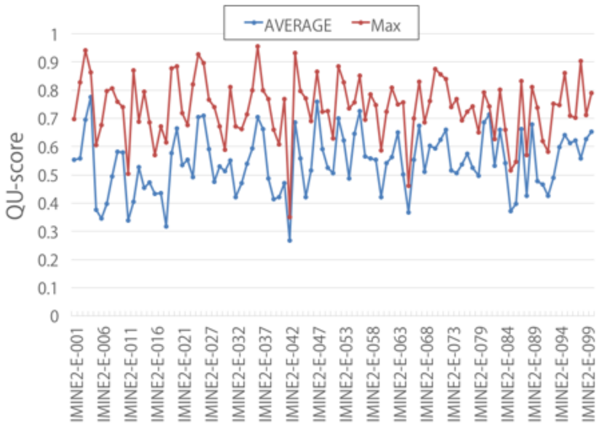


Figure 19. Per-topic QU-score for English Query Understanding.

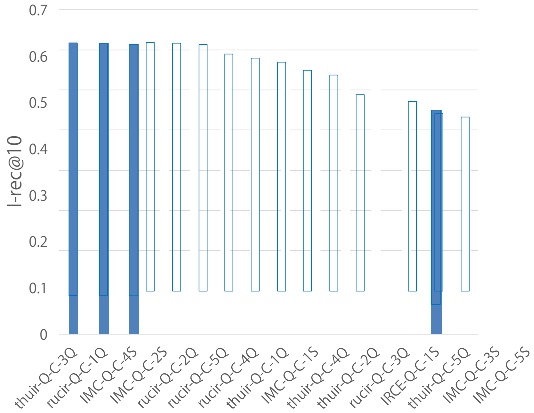


Figure 20. I-rec@10 for 91 unclear topics in Chinese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

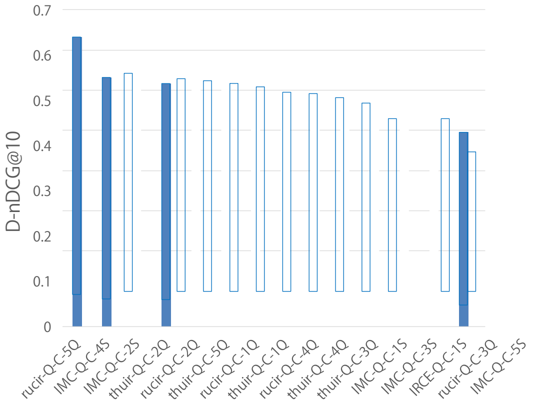


Figure 21. D-nDCG@10 for 91 unclear topics in Chinese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

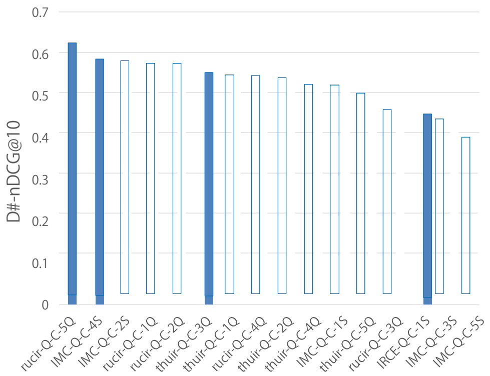


Figure 22. D#-nDCG@10 for 91 unclear topics in Chinese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

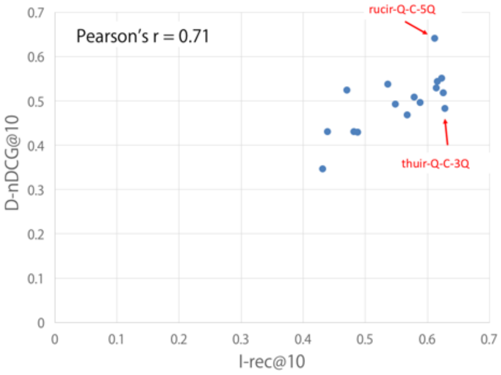


Figure 23. I-rec@10/D-nDCG@10 graph for Chinese Query Understanding.

Figures 24 and 25 show the mean V-score and QU-score performances of the Chinese Query Understanding subtask runs. The significantly different run pairs are also reported in Appendix B. Figure 26 shows the corresponding V-score/D#-nDCG graph. From the figures, we can observe that rucir-Q-C-5Q, which is the top performer in terms of D#-nDCG, is the winner in terms of both V-score and QU-score. However, rucir-Q-C-5Q is statistically indistinguishable from the other runs except for rucir-Q-C-3Q and rucir-Q-C-5Q. Figure 27 shows the per-topic QU-score the Chinese Query Understanding subtask.

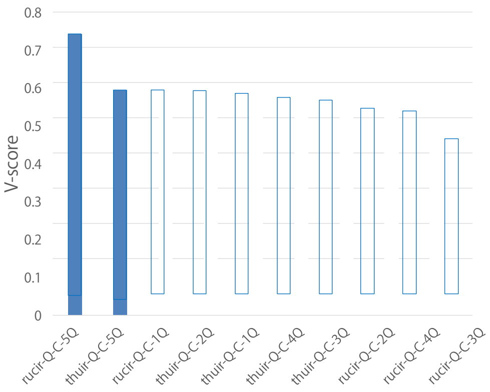


Figure 24. V-score for 91 unclear topics in Chinese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

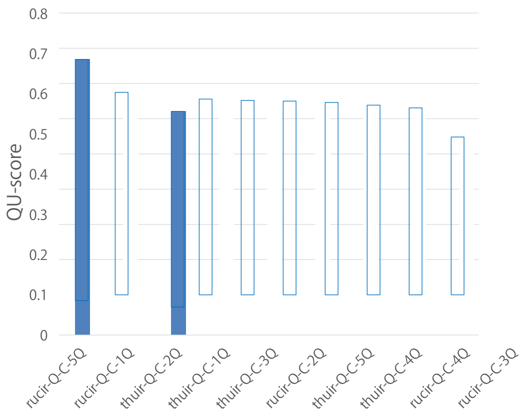


Figure 25. QU-score (official measure) for 91 unclear topics in Chinese Query Understanding subtask (run with the highest performance for each participant team is shown as a colored block).

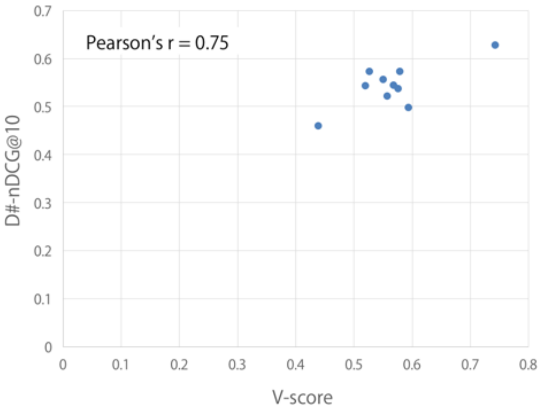


Figure 26. V-score@10/D#-nDCG@10 graph for English Query Understanding.

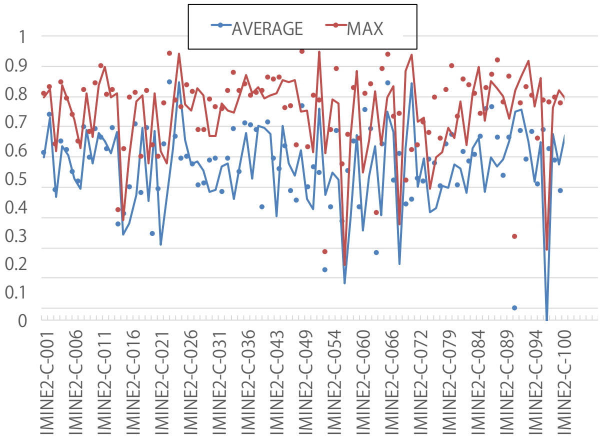


Figure 27. Per-topic QU-score for English Query Understanding.

5.4 English Vertical Incorporating subtask

Next, we report the evaluation results of the Vertical Incorporating subtask. Figures 28, 29, and 30 show the mean I-rec@10, D-nDCG@10, and D#-nDCG@10 performances of the English Vertical Incorporating subtask runs. The significantly different run pairs are also reported in Appendix B. Figure 31 shows the corresponding I-rec@10/D-nDCG@10 graph. Note that, in the results of D#-nDCG shown in Figure 30, the performance of the clear queries is evaluated with nDCG@10. Unfortunately, we received the English Vertical Incorporating subtask runs only from rucir team. From the results, it can be observed that rucir-V-E-1M consistently performs the best in terms of all the metrics. rucir-V-E-1M significantly outperformed the other runs except for rucir-V-E-3M in terms of D#-nDCG@10.

From figure 28, we found that all the runs achieve the quite high intent recall (i.e., I-rec@10); every run achieves more than 0.95 I-rec@10. One possible reason is the effect of virtual documents. From the result assessment, we found that virtual documents (i.e. verticals) tended to be relevant to multiple intents for many topics. Therefore, any run which ranks virtual documents higher is likely to get higher intent recall.

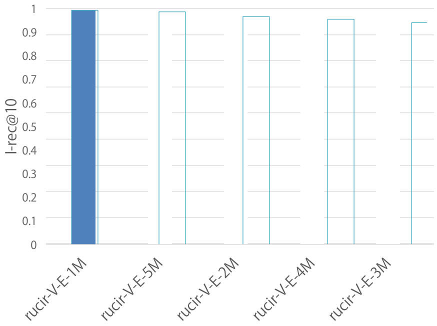


Figure 28. I-rec@10 for 97 unclear queries in English Vertical Incorporating subtask (run with the highest performance for each participant team is shown as a colored block).

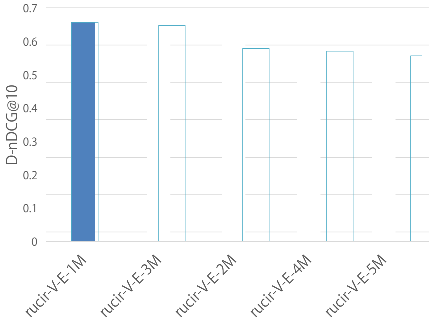


Figure 29. D-nDCG@10 for 97 unclear queries in English Vertical Incorporating subtask (run with the highest performance for each participant team is shown as a colored block).

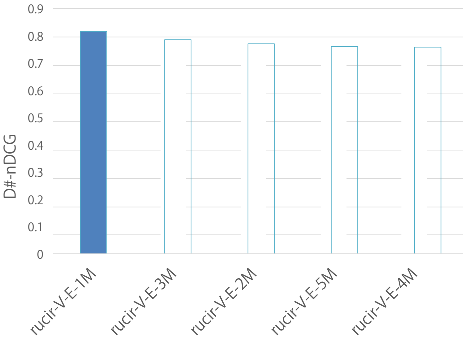


Figure 30. D#-nDCG@10 (official measure) for all queries in English Vertical Incorporating subtask (run with the highest performance for each participant team is shown as a colored block).

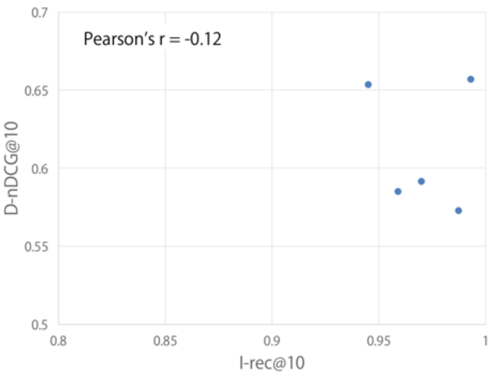


Figure 31. I-rec@10/D-nDCG@10 graph for English Vertical Incorporating subtask.

5.5 Chinese Vertical Incorporating subtask

Figures 32, 33, and 34 show the mean I-rec@10, D-nDCG@10, and D#-nDCG@10 performances of the Chinese Vertical Incorporating subtask runs. The significantly different run pairs are also reported in Appendix B. Figure 35 shows the corresponding I-rec@10/D-nDCG@10 graph. From the results, we can observe that rucir-V-C-1M is the winner in terms of all the metrics. Having that rucir achieves the best performance in QU-score in the Chinese Query Understanding subtask. We believe their strategy to find relevant verticals contributes the performance of the Vertical Incorporating subtask runs.

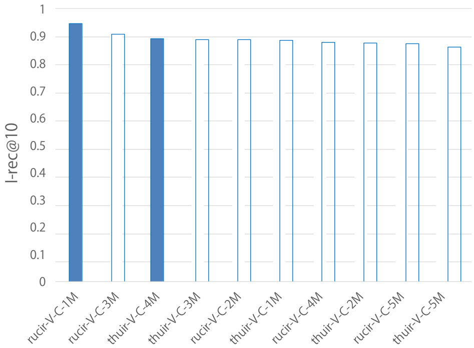


Figure 32. I-rec@10 for 91 unclear queries in Chinese Vertical Incorporating subtask (run with the highest performance for each participant team is shown as a colored block).

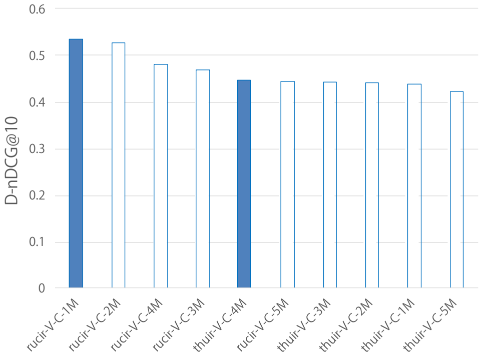


Figure 33. D-nDCG@10 for 91 unclear queries in Chinese Vertical Incorporating subtask (run with the highest performance for each participant team is shown as a colored block).

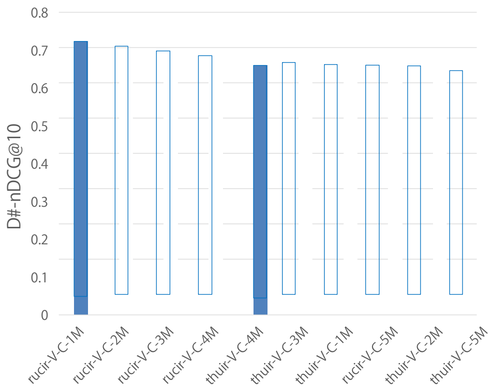


Figure 34. D#-nDCG@10 (official measure) for all queries in English Vertical Incorporating subtask (run with the highest performance for each participant team is shown as a colored block).

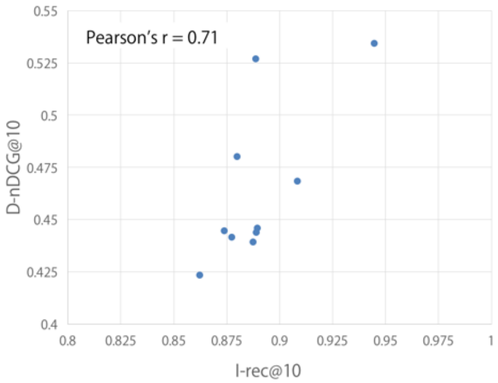


Figure 35. I-rec@10/D-nDCG@10 graph for English Vertical Incorporating subtask.

6. Conclusions

This paper provides an overview of the NTCIR-12 IMine-2 task. The IMine-2 task comprises the Query Understanding subtask and the Vertical Incorporating subtask. In this paper, we mainly explained the task design, data, evaluation methodology, and evaluation results. From the evaluation results we found that:

- In the Query Understanding subtask, NEXTI achieves the best performance in Japanese subtask, and rucir is the top performer in both of the English and Chinese subtasks.
- In the Query Understanding subtask, while the performances of the top runs in terms of D#-nDCG are similar to each other, the differences of their V-score performances larger.
- In the Vertical Incorporating subtask, rucir achieves the top performance in both of the English and Chinese subtasks.
- In the Vertical Incorporating subtask, we found that most runs achieve quite high I-rec@10 performances. This might be mainly because verticals are likely to satisfy multiple intents.

7. ACKNOWLEDGMENTS

We thank the NTCIR-12 IMine-2 participants for their effort in submitting runs. We appreciate efforts made by Yahoo! JAPAN Corporation for providing quite valuable search query data. We also appreciate Sogou for their efforts on providing user behavior data.

8. REFERENCES

[1] J. Arguello, F. Diaz, J. Callan, and J.-F. Crespo: Sources of evidence for vertical selection. In Proceedings of ACM SIGIR2009.

[2] T. Demeester, D. Trieschnigg, D. Nguyen, and D. Hiemstra: Overview of the TREC 2013 Federated Web Search Track. In Proceedings of TREC2013, 2013.

[3] T. Demeester, D. Trieschnigg, D. Nguyen, K. Zhou, and D. Hiemstra: Overview of the TREC 2014 Federated Web Search Track, In Proceedings of TREC2014, 2014.

[4] T. Sakai and R. Song: Evaluating Diversified Search Results Using Per-Intent Graded Relevance, In Proceedings of ACM SIGIR 2011, 2011.

[5] K. Zhou, T. Demeester, D. Nguyen, D. Hiemstra and D. Trieschnigg: Aligning Vertical Collection Relevance with User Intent, In Proceedings of ACM CIKM 2014, 2014.

[6] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, R. Song, R, M.P. Kato, and M. Iwata: Overview of the NTCIR-10 INTENT-2 Task, In Proceedings of NTCIR-10, 2013.

[7] T. Sakai: Designing Test Collections for Comparing Many Systems, In Proceedings of ACM CIKM2014, 2014.

[8] R. Song, M. Zhang, T. Sakai, M.P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii: Overview of the NTCIR-9 INTENT Task, In Proceedings of NTCIR-10, 2011.

[9] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. P. Kato, H. Ohshima and K. Zhou: Overview of the NTCIR-11 IMine Task. In Proceedings of NTCIR-11, 2014.

[10] T. Yamamoto, T. Sakai, M. Iwata, Y. Chen, J.-R. Wen and K. Tanaka: The Wisdom of Advertisers: Mining Subgoals via Query Clustering. In Proceedings of ACM CIKM 2012, 2012.

[11] T. Sakai: Metrics, Statistics, PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173), 2014.

APPENDIX

A. TOPICS

Full lists of English, Chinese, and Japanese query topics used in the IMine-2 task are shown in Tables 5, 6, and 7, respectively. The queries marked “x” in the “Shared” column represent they are shared among English, Chinese and Japanese query topics.

Table 5. NTCIR-12 IMine-2 English queries.

Topic ID	Query	Query Type	Shared
IMINE2-E-001	cvs	Ambiguous	x
IMINE2-E-002	Bumblebee	Ambiguous	
IMINE2-E-003	Tony Allen	Ambiguous	
IMINE2-E-004	wallpaper	Ambiguous	x
IMINE2-E-005	Opera	Ambiguous	
IMINE2-E-006	ginger	Ambiguous	
IMINE2-E-007	spirit	Ambiguous	
IMINE2-E-008	Pluto	Ambiguous	
IMINE2-E-009	full house	Ambiguous	
IMINE2-E-010	JFK	Ambiguous	
IMINE2-E-011	persona	Ambiguous	x
IMINE2-E-012	Virginia	Ambiguous	
IMINE2-E-013	steam	Ambiguous	
IMINE2-E-014	Borders	Ambiguous	x
IMINE2-E-015	Manchester	Ambiguous	x
IMINE2-E-016	PS	Ambiguous	x
IMINE2-E-017	elegy	Ambiguous	x
IMINE2-E-018	Elizabeth	Ambiguous	x
IMINE2-E-019	Yosemite	Ambiguous	x
IMINE2-E-020	Williams tennis	Ambiguous	
IMINE2-E-021	Nirvana	Ambiguous	
IMINE2-E-022	Tomahawk	Ambiguous	x
IMINE2-E-023	Magnus	Ambiguous	
IMINE2-E-024	KitKat	Ambiguous	
IMINE2-E-025	mahomet high school homepage	Very clear	
IMINE2-E-026	warner brothers	Faceted	
IMINE2-E-027	Socrates	Faceted	
IMINE2-E-028	Robert Kennedy	Faceted	
IMINE2-E-029	fossil	Faceted	
IMINE2-E-030	Star Wars	Faceted	x
IMINE2-E-031	maple trees	Faceted	
IMINE2-E-032	iraq war	Faceted	
IMINE2-E-033	Santa Claus	Faceted	x
IMINE2-E-034	digital art	Faceted	
IMINE2-E-035	moody blues	Faceted	
IMINE2-E-036	Uyghur cuisine	Faceted	x
IMINE2-E-037	bass guitars	Faceted	
IMINE2-E-038	poker	Faceted	
IMINE2-E-039	swallow	Faceted	
IMINE2-E-040	Pokemon	Faceted	x
IMINE2-E-041	gaba	Faceted	
IMINE2-E-042	Cat	Faceted	x
IMINE2-E-043	boy names	Faceted	
IMINE2-E-044	powerpoint	Faceted	
IMINE2-E-045	Denmark	Faceted	x
IMINE2-E-046	Gardening	Faceted	x
IMINE2-E-047	t-test	Faceted	x
IMINE2-E-048	sodium	Faceted	
IMINE2-E-049	spanish recipes	Faceted	
IMINE2-E-050	united airlines phone number	Very clear	
IMINE2-E-051	recover eyesight	Task-oriented	x
IMINE2-E-052	obesity prevention	Task-oriented	x
IMINE2-E-053	hair growth	Task-oriented	x
IMINE2-E-054	quit smoking	Task-oriented	
IMINE2-E-055	grow taller	Task-oriented	
IMINE2-E-056	sleep better	Task-oriented	
IMINE2-E-057	relieve stress	Task-oriented	
IMINE2-E-058	speak French	Task-oriented	
IMINE2-E-059	ride unicycle	Task-oriented	
IMINE2-E-060	run faster	Task-oriented	
IMINE2-E-061	learn Korean	Task-oriented	x
IMINE2-E-062	play piano	Task-oriented	
IMINE2-E-063	become firefigher	Task-oriented	
IMINE2-E-064	make resume	Task-oriented	
IMINE2-E-065	wedding	Task-oriented	
IMINE2-E-066	learn golf	Task-oriented	x
IMINE2-E-067	mastering touch typing	Task-oriented	x
IMINE2-E-068	debt releaf	Task-oriented	
IMINE2-E-069	grow vegetables	Task-oriented	
IMINE2-E-070	whale watching	Task-oriented	
IMINE2-E-071	how to spend Christmas	Task-oriented	x
IMINE2-E-072	home cleaning	Task-oriented	
IMINE2-E-073	travel to Italy	Task-oriented	
IMINE2-E-074	travel Hawaii	Task-oriented	x
IMINE2-E-075	safeco field address	Very clear	
IMINE2-E-076	wallpaper scenery	Vertical-oriented	x
IMINE2-E-077	happy birthday graphics	Vertical-oriented	
IMINE2-E-078	new year card design	Vertical-oriented	x
IMINE2-E-079	drawings of flowers	Vertical-oriented	
IMINE2-E-080	michael jackson photo	Vertical-oriented	
IMINE2-E-081	world news	Vertical-oriented	x
IMINE2-E-082	TPP progress	Vertical-oriented	x
IMINE2-E-083	mlb scores	Vertical-oriented	
IMINE2-E-084	apple latest news	Vertical-oriented	
IMINE2-E-085	obama update	Vertical-oriented	
IMINE2-E-086	what is GPU	Vertical-oriented	x
IMINE2-E-087	bluetooth	Vertical-oriented	
IMINE2-E-088	Construction point	Vertical-oriented	x
IMINE2-E-089	analogy definition	Vertical-oriented	
IMINE2-E-090	parkinson's disease	Vertical-oriented	
IMINE2-E-091	single-lens reflex recommendation	Vertical-oriented	x
IMINE2-E-092	Why white chocolate white	Vertical-oriented	
IMINE2-E-093	Do bananas have seeds	Vertical-oriented	x
IMINE2-E-094	difference between college and university	Vertical-oriented	
IMINE2-E-095	how to fix a broken zipper	Vertical-oriented	
IMINE2-E-096	cheap laptops	Vertical-oriented	
IMINE2-E-097	iPhone case	Vertical-oriented	x
IMINE2-E-098	discount plasma tv	Vertical-oriented	
IMINE2-E-099	mothers day gifts	Vertical-oriented	
IMINE2-E-100	ps3 online shopping	Vertical-oriented	x

Table 6. NTCIR-12 IMine-2 Chinese queries.

Topic ID	Query	Query Type	Shared
IMINE2-C-001	cvs	Ambiguous	x
IMINE2-C-002	壁纸	Ambiguous	x
IMINE2-C-003	边界	Ambiguous	x
IMINE2-C-004	曼彻斯特	Faceted	x
IMINE2-C-005	PS	Ambiguous	x
IMINE2-C-006	哀歌	Faceted	x
IMINE2-C-007	伊丽莎白	Ambiguous	x
IMINE2-C-008	优胜美地	Faceted	x
IMINE2-C-009	战斧	Ambiguous	x
IMINE2-C-010	星球大战	Ambiguous	x
IMINE2-C-011	圣诞老人	Faceted	x
IMINE2-C-012	新疆菜	Vertical-oriented	x
IMINE2-C-013	口袋妖怪	Faceted	x
IMINE2-C-014	猫	Ambiguous	x
IMINE2-C-015	丹麦	Faceted	x
IMINE2-C-016	园艺	Faceted	x
IMINE2-C-017	T 检验	Vertical-oriented	x
IMINE2-C-018	预防肥胖	Task-oriented	x
IMINE2-C-019	生发	Task-oriented	x
IMINE2-C-020	韩语学习	Task-oriented	x
IMINE2-C-021	高尔夫学习	Task-oriented	x
IMINE2-C-022	盲打学习	Task-oriented	x
IMINE2-C-023	圣诞节怎么过	Task-oriented	x
IMINE2-C-024	夏威夷旅游	Task-oriented	x
IMINE2-C-025	贺年卡设计	Task-oriented	x
IMINE2-C-026	国际新闻	Vertical-oriented	x
IMINE2-C-027	TPP 进展	Vertical-oriented	x
IMINE2-C-028	GPU 是什么	Vertical-oriented	x
IMINE2-C-029	单反相机推荐	Task-oriented	x
IMINE2-C-030	香蕉有种子么	Vertical-oriented	x
IMINE2-C-031	iphone 保护套	Vertical-oriented	x
IMINE2-C-032	广发聚丰基金今日净值	Vertical-oriented	
IMINE2-C-033	手机游戏排行榜	Vertical-oriented	
IMINE2-C-034	李蒽熙	Faceted	
IMINE2-C-035	cctv5 节目表	Vertical-oriented	
IMINE2-C-036	小石潭记原文及翻译	Vertical-oriented	
IMINE2-C-037	小木虫	Very clear	
IMINE2-C-038	雅诗兰黛	Faceted	
IMINE2-C-039	描写春天的句子	Task-oriented	
IMINE2-C-040	支付宝客服电话	Very clear	
IMINE2-C-041	天天基金净值查询	Task-oriented	
IMINE2-C-042	陈赫电视剧	Vertical-oriented	
IMINE2-C-043	ems 快递单号查询	Task-oriented	
IMINE2-C-044	尼泊尔地图	Vertical-oriented	
IMINE2-C-045	亚投行创始成员国名单	Vertical-oriented	
IMINE2-C-046	中国之声在线收听	Vertical-oriented	
IMINE2-C-047	qq 签名伤感	Vertical-oriented	
IMINE2-C-048	支付宝实名认证	Very clear	
IMINE2-C-049	多啦 a 梦国语版全集	Vertical-oriented	
IMINE2-C-050	公交查询	Task-oriented	
IMINE2-C-051	国税发票查询	Task-oriented	
IMINE2-C-052	国际油价	Vertical-oriented	
IMINE2-C-053	巧虎智力答题	Vertical-oriented	
IMINE2-C-054	去哪儿网机票查询	Vertical-oriented	
IMINE2-C-055	西祠胡同	Very clear	
IMINE2-C-056	灵域	Faceted	
IMINE2-C-057	刘兰芳评书	Vertical-oriented	
IMINE2-C-058	大写数字一到十	Vertical-oriented	
IMINE2-C-059	特殊符号图案大全	Vertical-oriented	
IMINE2-C-060	北京交通违章查询	Task-oriented	
IMINE2-C-061	亚冠赛程	Vertical-oriented	
IMINE2-C-062	汉英在线翻译	Task-oriented	
IMINE2-C-063	速度与激情	Faceted	
IMINE2-C-064	谚语大全	Vertical-oriented	
IMINE2-C-065	chrome 浏览器官方下载	Very clear	
IMINE2-C-066	爱回家粤语	Vertical-oriented	
IMINE2-C-067	芒果台直播	Vertical-oriented	
IMINE2-C-068	高铁网上订票官网	Very clear	
IMINE2-C-069	搬家吉日查询	Task-oriented	
IMINE2-C-070	qq 影音官方下载	Vertical-oriented	
IMINE2-C-071	270005 基金今天净值	Vertical-oriented	
IMINE2-C-072	安卓游戏下载	Vertical-oriented	
IMINE2-C-073	斯巴达克斯第二季	Vertical-oriented	
IMINE2-C-074	白眉大侠单田芳	Vertical-oriented	
IMINE2-C-075	小米 4 怎么样	Vertical-oriented	
IMINE2-C-076	肯德基订餐	Task-oriented	
IMINE2-C-077	uber	Very clear	
IMINE2-C-078	亚航官网	Very clear	
IMINE2-C-079	威客兼职	Task-oriented	
IMINE2-C-080	注册香港公司	Task-oriented	
IMINE2-C-081	完美世界	Faceted	
IMINE2-C-082	双色球	Faceted	
IMINE2-C-083	中国好声音	Vertical-oriented	
IMINE2-C-084	苹果 6	Faceted	
IMINE2-C-085	心花路放	Vertical-oriented	
IMINE2-C-086	张碧晨	Faceted	
IMINE2-C-087	小苹果	Faceted	
IMINE2-C-088	爷们儿电视剧	Vertical-oriented	
IMINE2-C-089	曼联	Vertical-oriented	
IMINE2-C-090	附近的电影院	Vertical-oriented	
IMINE2-C-091	美现首例埃博拉患者	Vertical-oriented	
IMINE2-C-092	月全食	Faceted	
IMINE2-C-093	火影忍者	Faceted	
IMINE2-C-094	节约用水手抄报	Vertical-oriented	
IMINE2-C-095	辽宁号	Vertical-oriented	
IMINE2-C-096	资生堂	Faceted	
IMINE2-C-097	星光大道	Ambiguous	
IMINE2-C-098	顺丰运单查询	Very clear	
IMINE2-C-099	qq 头像	Vertical-oriented	
IMINE2-C-100	苹果手机序列号	Vertical-oriented	

Table 7. NTCIR-12 IMine-2 Japanese queries.

Topic ID	Query	Query Type	Shared
IMINE2-J-001	cvs	Ambiguous	x
IMINE2-J-002	ゆず	Ambiguous	
IMINE2-J-003	フェイト	Ambiguous	
IMINE2-J-004	壁紙	Ambiguous	x
IMINE2-J-005	ワンピース	Ambiguous	
IMINE2-J-006	マック	Ambiguous	
IMINE2-J-007	アルク	Ambiguous	
IMINE2-J-008	ゾロ	Ambiguous	
IMINE2-J-009	スバル	Ambiguous	
IMINE2-J-010	読売	Ambiguous	
IMINE2-J-011	ペルソナ	Ambiguous	x
IMINE2-J-012	青山	Ambiguous	
IMINE2-J-013	なでしこ	Ambiguous	
IMINE2-J-014	ボーダーズ	Ambiguous	x
IMINE2-J-015	マンチェスター	Ambiguous	x
IMINE2-J-016	PS	Ambiguous	x
IMINE2-J-017	エレジー	Ambiguous	x
IMINE2-J-018	エリザベス	Ambiguous	x
IMINE2-J-019	ヨセミテ	Ambiguous	x
IMINE2-J-020	ミンク	Ambiguous	
IMINE2-J-021	ミューレン	Ambiguous	
IMINE2-J-022	トマホーク	Ambiguous	x
IMINE2-J-023	鉄拳	Ambiguous	
IMINE2-J-024	メッセンジャー	Ambiguous	
IMINE2-J-025	フィクサー	Ambiguous	
IMINE2-J-026	一人暮らし	Faceted	
IMINE2-J-027	テレビ	Faceted	
IMINE2-J-028	阪神タイガース	Faceted	
IMINE2-J-029	クレヨンしんちゃん	Faceted	
IMINE2-J-030	スターウォーズ	Faceted	x
IMINE2-J-031	競艇	Faceted	
IMINE2-J-032	シャチ	Faceted	
IMINE2-J-033	サンタクロース	Faceted	x
IMINE2-J-034	名倉潤	Faceted	
IMINE2-J-035	プリウス	Faceted	
IMINE2-J-036	ウイグル料理	Faceted	x
IMINE2-J-037	携帯電話	Faceted	
IMINE2-J-038	しょこたん	Faceted	
IMINE2-J-039	テイルズ	Faceted	
IMINE2-J-040	ポケモン	Faceted	x
IMINE2-J-041	メガネ	Faceted	
IMINE2-J-042	ねこ	Faceted	x
IMINE2-J-043	紙飛行機	Faceted	
IMINE2-J-044	東京　お土産	Faceted	
IMINE2-J-045	デンマーク	Faceted	x
IMINE2-J-046	ガーデニング	Faceted	x
IMINE2-J-047	t 検定	Faceted	x
IMINE2-J-048	神田沙也加	Faceted	
IMINE2-J-049	ヴィレッジヴァンガード	Faceted	
IMINE2-J-050	キティちゃん	Faceted	
IMINE2-J-051	視力 改善	Task-oriented	x
IMINE2-J-052	肥満 予防	Task-oriented	x
IMINE2-J-053	育毛	Task-oriented	x
IMINE2-J-054	のどの痛み 直し方	Task-oriented	
IMINE2-J-055	ホワイトニング	Task-oriented	
IMINE2-J-056	O 脚 直し方	Task-oriented	
IMINE2-J-057	盗聴器 探し方	Task-oriented	
IMINE2-J-058	速読　方法	Task-oriented	
IMINE2-J-059	周期表 覚え方	Task-oriented	
IMINE2-J-060	ruby 勉強	Task-oriented	
IMINE2-J-061	韓国語 学習	Task-oriented	x
IMINE2-J-062	TOEIC 対策	Task-oriented	
IMINE2-J-063	小論文　書き方	Task-oriented	
IMINE2-J-064	マイホーム 購入	Task-oriented	
IMINE2-J-065	ギター 弾く	Task-oriented	
IMINE2-J-066	ゴルフ 上達	Task-oriented	x
IMINE2-J-067	タッチタイピング 習得	Task-oriented	x
IMINE2-J-068	エントリーシート 作成	Task-oriented	
IMINE2-J-069	振り袖 レンタル	Task-oriented	
IMINE2-J-070	油性ペン 落とし方	Task-oriented	
IMINE2-J-071	クリスマス　過ごし方	Task-oriented	x
IMINE2-J-072	京都 観光	Task-oriented	
IMINE2-J-073	ディズニーランド 楽しむ	Task-oriented	
IMINE2-J-074	ハワイ 旅行	Task-oriented	x
IMINE2-J-075	東京　大阪	Task-oriented	
IMINE2-J-076	壁紙　風景	Vertical-oriented	x
IMINE2-J-077	神戸空港 写真	Vertical-oriented	
IMINE2-J-078	年賀状 イラスト	Vertical-oriented	x
IMINE2-J-079	大島優子 画像	Vertical-oriented	
IMINE2-J-080	国家予算　グラフ	Vertical-oriented	
IMINE2-J-081	海外　ニュース	Vertical-oriented	x
IMINE2-J-082	TPP 進展	Vertical-oriented	x
IMINE2-J-083	日米首脳会談	Vertical-oriented	
IMINE2-J-084	大阪都構想 結果	Vertical-oriented	
IMINE2-J-085	女子プロゴルフ　結果	Vertical-oriented	
IMINE2-J-086	GPU とは	Vertical-oriented	x
IMINE2-J-087	アルファリポ酸 定義	Vertical-oriented	
IMINE2-J-088	K 点	Vertical-oriented	x
IMINE2-J-089	宇宙線 wikipedia	Vertical-oriented	
IMINE2-J-090	バセドウ病	Vertical-oriented	
IMINE2-J-091	一眼レフ　おすすめ	Vertical-oriented	x
IMINE2-J-092	一人旅　おすすめ　日帰り	Vertical-oriented	
IMINE2-J-093	バナナの種 どこにある	Vertical-oriented	x
IMINE2-J-094	同級生 同窓生 違い	Vertical-oriented	
IMINE2-J-095	生ビールの生 意味	Vertical-oriented	
IMINE2-J-096	kindle 購入	Vertical-oriented	
IMINE2-J-097	iPhone ケース	Vertical-oriented	x
IMINE2-J-098	ブルーレイディスク	Vertical-oriented	
IMINE2-J-099	母の日　ギフト	Vertical-oriented	
IMINE2-J-100	ps3 通販	Vertical-oriented	x

B. SIGNIFICANTLY DIFFERENT RUN PAIRS

Significantly different run pairs found by the two-sided randomized Tukey's HSD at significant level $\alpha = 0.05$ are shown in Figures 36-56.

HUKB-Q-J-1Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

HUKB-Q-J-2Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

HUKB-Q-J-3Q with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, IRCE-Q-J-3S, IRCE-Q-J-5S, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

HUKB-Q-J-4Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

HUKB-Q-J-5Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

IRCE-Q-J-1S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-2S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-3S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-3Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-4S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-5S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-3Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

NEXTI-Q-J-1Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-1Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-2Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-3Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-4Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-5Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

Figure 36. Japanese Query Understanding subtask: significantly different pairs in terms of I-rec@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

HUKB-Q-J-1Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

HUKB-Q-J-2Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

HUKB-Q-J-3Q with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, IRCE-Q-J-1S, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

HUKB-Q-J-4Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

HUKB-Q-J-5Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

IRCE-Q-J-1S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-3Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-2S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-3S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-4S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-5S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

NEXTI-Q-J-1Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-1Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-2Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-3Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-4Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-5Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

Figure 37. Japanese Query Understanding subtask: significantly different pairs in terms of D-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

HUKB-Q-J-1Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

HUKB-Q-J-2Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

HUKB-Q-J-3Q with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, IRCE-Q-J-1S, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

HUKB-Q-J-4Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

HUKB-Q-J-5Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

IRCE-Q-J-1S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-3Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-2S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-3S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-4S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

IRCE-Q-J-5S with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

NEXTI-Q-J-1Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-1Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-2Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-3Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-4Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

YJST-Q-J-5Q with HUKB-Q-J-3Q, IRCE-Q-J-1S, IRCE-Q-J-2S, IRCE-Q-J-3S, IRCE-Q-J-4S, IRCE-Q-J-5S

Figure 38. Japanese Query Understanding subtask: significantly different pairs in terms of D#-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

HUKB-Q-J-1Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q, YJST-Q-J-4Q

HUKB-Q-J-2Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

HUKB-Q-J-3Q with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-5Q

HUKB-Q-J-4Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q, YJST-Q-J-4Q

HUKB-Q-J-5Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

NEXTI-Q-J-1Q with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-3Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

YJST-Q-J-1Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q, YJST-Q-J-4Q

YJST-Q-J-2Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

YJST-Q-J-3Q with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-4Q, YJST-Q-J-5Q

YJST-Q-J-4Q with HUKB-Q-J-1Q, HUKB-Q-J-4Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-3Q

YJST-Q-J-5Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

Figure 39. Japanese Query Understanding subtask: significantly different pairs in terms of V-score@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

HUKB-Q-J-1Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

HUKB-Q-J-2Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

HUKB-Q-J-3Q with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-4Q, YJST-Q-J-5Q

HUKB-Q-J-4Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q, YJST-Q-J-4Q

HUKB-Q-J-5Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

NEXTI-Q-J-1Q with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-3Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-3Q, YJST-Q-J-4Q, YJST-Q-J-5Q

YJST-Q-J-1Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

YJST-Q-J-2Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

YJST-Q-J-3Q with HUKB-Q-J-1Q, HUKB-Q-J-2Q, HUKB-Q-J-4Q, HUKB-Q-J-5Q, NEXTI-Q-J-1Q, YJST-Q-J-1Q, YJST-Q-J-2Q, YJST-Q-J-5Q

YJST-Q-J-4Q with HUKB-Q-J-3Q, HUKB-Q-J-4Q, NEXTI-Q-J-1Q

YJST-Q-J-5Q with HUKB-Q-J-3Q, NEXTI-Q-J-1Q, YJST-Q-J-3Q

Figure 40. Japanese Query Understanding subtask: significantly different pairs in terms of QU-score@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

HULTECH-Q-E-1Q with rucir-Q-E-3Q

KDEIM-Q-E-1S with rucir-Q-E-3Q

KDEIM-Q-E-2Q with rucir-Q-E-3Q

KDEIM-Q-E-3Q with rucir-Q-E-3Q

KDEIM-Q-E-4S with rucir-Q-E-3Q

rucir-Q-E-1Q with rucir-Q-E-3Q

rucir-Q-E-2Q with rucir-Q-E-3Q

rucir-Q-E-3Q with HULTECH-Q-E-1Q, KDEIM-Q-E-1S, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, KDEIM-Q-E-4S, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-4Q, rucir-Q-E-5Q

rucir-Q-E-4Q with rucir-Q-E-3Q

rucir-Q-E-5Q with rucir-Q-E-3Q

Figure 41. English Query Understanding subtask: significantly different pairs in terms of I-rec@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

HULTECH-Q-E-1Q with KDEIM-Q-E-4S, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

KDEIM-Q-E-1S with KDEIM-Q-E-4S, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

KDEIM-Q-E-2Q with KDEIM-Q-E-4S, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

KDEIM-Q-E-3Q with KDEIM-Q-E-4S, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

KDEIM-Q-E-4S with HULTECH-Q-E-1Q, KDEIM-Q-E-1S, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-1Q with HULTECH-Q-E-1Q, KDEIM-Q-E-1S, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-2Q with HULTECH-Q-E-1Q, KDEIM-Q-E-1S, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-3Q with HULTECH-Q-E-1Q, KDEIM-Q-E-1S, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, KDEIM-Q-E-4S, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-4Q, rucir-Q-E-5Q

rucir-Q-E-4Q with HULTECH-Q-E-1Q, KDEIM-Q-E-1S, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-5Q with KDEIM-Q-E-4S, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

Figure 42. English Query Understanding subtask: significantly different pairs in terms of D-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

HULTECH-Q-E-1Q with rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q

KDEIM-Q-E-1S with rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

KDEIM-Q-E-2Q with rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

KDEIM-Q-E-3Q with rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q

KDEIM-Q-E-4S with rucir-Q-E-3Q

rucir-Q-E-1Q with HULTECH-Q-E-1Q, KDEIM-Q-E-1S, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-2Q with HULTECH-Q-E-1Q, KDEIM-Q-E-1S, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-3Q with HULTECH-Q-E-1Q, KDEIM-Q-E-1S, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, KDEIM-Q-E-4S, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-4Q, rucir-Q-E-5Q

rucir-Q-E-4Q with KDEIM-Q-E-1S, KDEIM-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-5Q with rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

Figure 43. English Query Understanding subtask: significantly different pairs in terms of D#-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

HULTECH-Q-E-1Q with rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-4Q, rucir-Q-E-5Q

KDEIM-Q-E-2Q with rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q, rucir-Q-E-5Q

KDEIM-Q-E-3Q with rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q, rucir-Q-E-5Q

rucir-Q-E-1Q with HULTECH-Q-E-1Q, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-2Q with HULTECH-Q-E-1Q, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-3Q with KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-2Q, rucir-Q-E-5Q

rucir-Q-E-4Q with HULTECH-Q-E-1Q, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-5Q with HULTECH-Q-E-1Q, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

Figure 44. English Query Understanding subtask: significantly different pairs in terms of V-score@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

HULTECH-Q-E-1Q with rucir-Q-E-3Q, rucir-Q-E-5Q

KDEIM-Q-E-2Q with rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-5Q

KDEIM-Q-E-3Q with rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-1Q with KDEIM-Q-E-3Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-2Q with KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-3Q with HULTECH-Q-E-1Q, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-4Q, rucir-Q-E-5Q

rucir-Q-E-4Q with rucir-Q-E-3Q, rucir-Q-E-5Q

rucir-Q-E-5Q with HULTECH-Q-E-1Q, KDEIM-Q-E-2Q, KDEIM-Q-E-3Q, rucir-Q-E-1Q, rucir-Q-E-2Q, rucir-Q-E-3Q, rucir-Q-E-4Q

Figure 45. English Query Understanding subtask: significantly different pairs in terms of QU-score@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

IMC-Q-C-1S with IMC-Q-C-3S, IMC-Q-C-5S, thuir-Q-C-5Q
IMC-Q-C-2S with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, thuir-Q-C-5Q
IMC-Q-C-3S with IMC-Q-C-1S, IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q
IMC-Q-C-4S with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, thuir-Q-C-5Q
IMC-Q-C-5S with IMC-Q-C-1S, IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q
IRCE-Q-C-1S with IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-3Q
rucir-Q-C-1Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, thuir-Q-C-5Q
rucir-Q-C-2Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, thuir-Q-C-5Q
rucir-Q-C-3Q with IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-3Q
rucir-Q-C-4Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, thuir-Q-C-5Q
rucir-Q-C-5Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, thuir-Q-C-5Q
thuir-Q-C-1Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, thuir-Q-C-5Q
thuir-Q-C-2Q with IMC-Q-C-3S, IMC-Q-C-5S
thuir-Q-C-3Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, thuir-Q-C-5Q
thuir-Q-C-4Q with IMC-Q-C-3S, IMC-Q-C-5S
thuir-Q-C-5Q with IMC-Q-C-1S, IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-3Q

Figure 46. Chinese Query Understanding subtask: significantly different pairs in terms of I-rec@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

IMC-Q-C-1S with IMC-Q-C-5S, rucir-Q-C-5Q
IMC-Q-C-2S with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, rucir-Q-C-5Q
IMC-Q-C-3S with IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-2Q, rucir-Q-C-5Q, thuir-Q-C-2Q, thuir-Q-C-5Q
IMC-Q-C-4S with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, rucir-Q-C-5Q
IMC-Q-C-5S with IMC-Q-C-1S, IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q, thuir-Q-C-5Q
IRCE-Q-C-1S with IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-2Q, rucir-Q-C-5Q, thuir-Q-C-2Q, thuir-Q-C-5Q
rucir-Q-C-1Q with IMC-Q-C-5S, rucir-Q-C-5Q
rucir-Q-C-2Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, rucir-Q-C-5Q
rucir-Q-C-3Q with IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-2Q, rucir-Q-C-5Q, thuir-Q-C-2Q, thuir-Q-C-5Q
rucir-Q-C-4Q with IMC-Q-C-5S, rucir-Q-C-5Q
rucir-Q-C-5Q with IMC-Q-C-1S, IMC-Q-C-2S, IMC-Q-C-3S, IMC-Q-C-4S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-3Q, rucir-Q-C-4Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q, thuir-Q-C-5Q
thuir-Q-C-1Q with IMC-Q-C-5S, rucir-Q-C-5Q
thuir-Q-C-2Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-3Q with IMC-Q-C-5S, rucir-Q-C-5Q
thuir-Q-C-4Q with IMC-Q-C-5S, rucir-Q-C-5Q
thuir-Q-C-5Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, rucir-Q-C-5Q

Figure 47. Chinese Query Understanding subtask: significantly different pairs in terms of D-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

IMC-Q-C-1S with IMC-Q-C-3S, IMC-Q-C-5S, rucir-Q-C-5Q
IMC-Q-C-2S with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, thuir-Q-C-5Q
IMC-Q-C-3S with IMC-Q-C-1S, IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q
IMC-Q-C-4S with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, thuir-Q-C-5Q
IMC-Q-C-5S with IMC-Q-C-1S, IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q, thuir-Q-C-5Q
IRCE-Q-C-1S with IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q
rucir-Q-C-1Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q
rucir-Q-C-2Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q
rucir-Q-C-3Q with IMC-Q-C-2S, IMC-Q-C-4S, rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q
rucir-Q-C-4Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, rucir-Q-C-5Q
rucir-Q-C-5Q with IMC-Q-C-1S, IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, rucir-Q-C-4Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-4Q, thuir-Q-C-5Q
thuir-Q-C-1Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-2Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-3Q with IMC-Q-C-3S, IMC-Q-C-5S, IRCE-Q-C-1S, rucir-Q-C-3Q
thuir-Q-C-4Q with IMC-Q-C-3S, IMC-Q-C-5S, rucir-Q-C-5Q
thuir-Q-C-5Q with IMC-Q-C-2S, IMC-Q-C-4S, IMC-Q-C-5S, rucir-Q-C-5Q

Figure 48. Chinese Query Understanding subtask: significantly different pairs in terms of D#-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

rucir-Q-C-1Q with rucir-Q-C-3Q, rucir-Q-C-5Q
rucir-Q-C-2Q with rucir-Q-C-3Q, rucir-Q-C-5Q
rucir-Q-C-3Q with rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q, thuir-Q-C-5Q
rucir-Q-C-4Q with rucir-Q-C-3Q, rucir-Q-C-5Q
rucir-Q-C-5Q with rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-3Q, rucir-Q-C-4Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q, thuir-Q-C-5Q
thuir-Q-C-1Q with rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-2Q with rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-3Q with rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-4Q with rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-5Q with rucir-Q-C-3Q, rucir-Q-C-5Q

Figure 49. Chinese Query Understanding subtask: significantly different pairs in terms of V-score@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

rucir-Q-C-1Q with rucir-Q-C-3Q, rucir-Q-C-5Q
rucir-Q-C-2Q with rucir-Q-C-3Q, rucir-Q-C-5Q
rucir-Q-C-3Q with rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-4Q, rucir-Q-C-5Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q, thuir-Q-C-5Q
rucir-Q-C-4Q with rucir-Q-C-3Q, rucir-Q-C-5Q
rucir-Q-C-5Q with rucir-Q-C-1Q, rucir-Q-C-2Q, rucir-Q-C-3Q, rucir-Q-C-4Q, thuir-Q-C-1Q, thuir-Q-C-2Q, thuir-Q-C-3Q, thuir-Q-C-4Q, thuir-Q-C-5Q
thuir-Q-C-1Q with rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-2Q with rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-3Q with rucir-Q-C-3Q, rucir-Q-C-5Q
thuir-Q-C-4Q with rucir-Q-C-3Q, rucir-Q-C-5Q

thuir-Q-C-5Q with rucir-Q-C-3Q, rucir-Q-C-5Q

Figure 50. Chinese Query Understanding subtask: significantly different pairs in terms of QU-score@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

rucir-V-E-1M with rucir-V-E-3M
rucir-V-E-3M with rucir-V-E-1M, rucir-V-E-5M
rucir-V-E-5M with rucir-V-E-3M

Figure 51. English Vertical Incorporating subtask: significantly different pairs in terms of I-rec@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

rucir-V-E-1M with rucir-V-E-2M, rucir-V-E-4M, rucir-V-E-5M
rucir-V-E-2M with rucir-V-E-1M, rucir-V-E-3M
rucir-V-E-3M with rucir-V-E-2M, rucir-V-E-4M, rucir-V-E-5M
rucir-V-E-4M with rucir-V-E-1M, rucir-V-E-3M
rucir-V-E-5M with rucir-V-E-1M, rucir-V-E-3M

Figure 52. English Vertical Incorporating subtask: significantly different pairs in terms of D-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

rucir-V-E-1M with rucir-V-E-2M, rucir-V-E-4M, rucir-V-E-5M
rucir-V-E-2M with rucir-V-E-1M, rucir-V-E-3M
rucir-V-E-3M with rucir-V-E-2M, rucir-V-E-4M, rucir-V-E-5M
rucir-V-E-4M with rucir-V-E-1M, rucir-V-E-3M
rucir-V-E-5M with rucir-V-E-1M, rucir-V-E-3M

Figure 53. English Vertical Incorporating subtask: significantly different pairs in terms of D#-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

rucir-V-C-1M with rucir-V-C-4M, rucir-V-C-5M, thuir-V-C-2M, thuir-V-C-5M
rucir-V-C-4M with rucir-V-C-1M
rucir-V-C-5M with rucir-V-C-1M
thuir-V-C-2M with rucir-V-C-1M
thuir-V-C-5M with rucir-V-C-1M

Figure 54. Chinese Vertical Incorporating subtask: significantly different pairs in terms of I-rec@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

rucir-V-C-1M with rucir-V-C-5M, thuir-V-C-1M, thuir-V-C-2M, thuir-V-C-3M, thuir-V-C-4M, thuir-V-C-5M
rucir-V-C-2M with rucir-V-C-5M, thuir-V-C-1M, thuir-V-C-2M, thuir-V-C-3M, thuir-V-C-4M, thuir-V-C-5M
rucir-V-C-5M with rucir-V-C-1M, rucir-V-C-2M
thuir-V-C-1M with rucir-V-C-1M, rucir-V-C-2M
thuir-V-C-2M with rucir-V-C-1M, rucir-V-C-2M
thuir-V-C-3M with rucir-V-C-1M, rucir-V-C-2M
thuir-V-C-4M with rucir-V-C-1M, rucir-V-C-2M
thuir-V-C-5M with rucir-V-C-1M, rucir-V-C-2M

Figure 55. Chinese Vertical Incorporating subtask: significantly different pairs in terms of D-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).

rucir-V-C-1M with rucir-V-C-5M, thuir-V-C-1M, thuir-V-C-2M, thuir-V-C-3M, thuir-V-C-4M, thuir-V-C-5M
rucir-V-C-2M with rucir-V-C-5M, thuir-V-C-2M, thuir-V-C-5M
rucir-V-C-3M with thuir-V-C-5M
rucir-V-C-5M with rucir-V-C-1M, rucir-V-C-2M
thuir-V-C-1M with rucir-V-C-1M
thuir-V-C-2M with rucir-V-C-1M, rucir-V-C-2M
thuir-V-C-3M with rucir-V-C-1M
thuir-V-C-4M with rucir-V-C-1M
thuir-V-C-5M with rucir-V-C-1M, rucir-V-C-2M, rucir-V-C-3M

Figure 56. Chinese Vertical Incorporating subtask: significantly different pairs in terms of D#-nDCG@10 (two-sided randomized Tukey's HSD at $\alpha = 0.05$).