# IRCE at the NTCIR-12 IMine-2 Task

Query Understanding (QU) Subtask


**Ximei Song**
University of Tsukuba

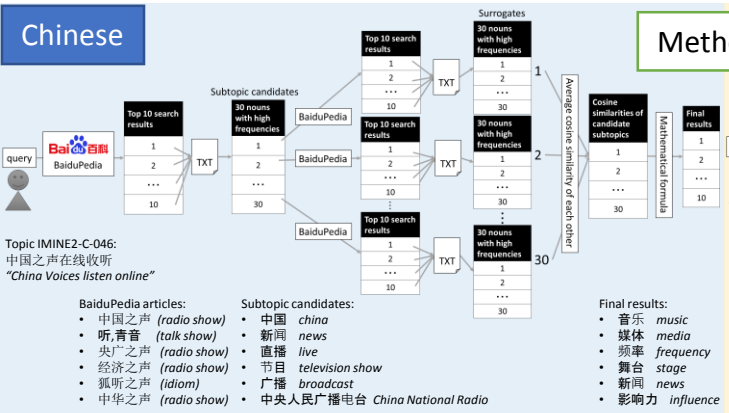
**Yuka Egusa**
National Institute for Educational Policy Research

**Hitomi Saito**
Aichi University of Education

**Masao Takaku**
University of Tsukuba


## Methods

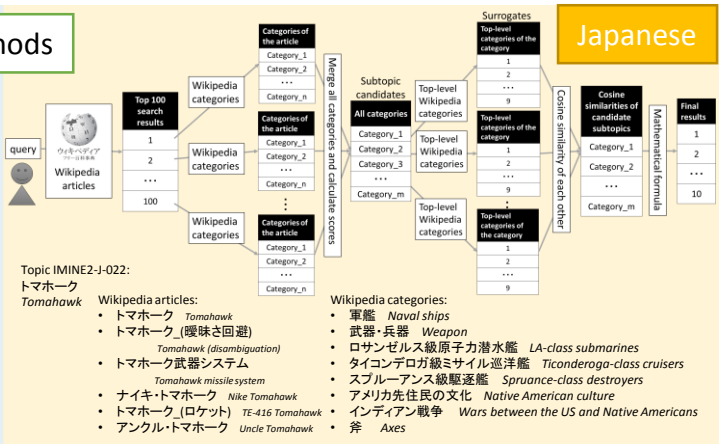
### Chinese




[Dataset] **BaiduPedia**, as an online resource

- Top 30 nouns with high frequencies as subtopic candidates
- Average cosine similarity of each other as seeds for diversification

### Japanese




[Dataset] **Japanese Wikipedia** data dump, as of December 2013

- Wikipedia categories as subtopic candidates
- Nine top-level categories as seeds for diversification
  - ✓ 学問(Academia), 技術(Technology), 自然(Nature), 社会(Society), 地理(Geography), 人間(Humans), 文化(Culture), 歴史(History), 総記(Generals)
- Ranking: $Score = \alpha \times Score_{surrogate} + (1 - \alpha) \times Score_{category}$

## Results

Evaluation results of the Chinese language run

| | I-rec@10 | D-nDCG@10 | D#-nDCG@10 |
|---|---|---|---|
| IRCE-QU-C-1S | **0.4827** | **0.4290** | **0.4558** |

Evaluation results of D#-nDCG@10 per topic type

| | Ambiguous | Faceted | Task-oriented | Vertical-oriented |
|---|---|---|---|---|
| IRCE-QU-C-1S | 0.4456 | 0.4450 | 0.4408 | 0.4698 |

Failure Analysis:

- Although the judged subtopics of the topic IMINE2-C-006 "哀歌" were songs, network novels, published books, songs' information, the Bible, and movies, the majority of the subtopics results from our run was dominated by a particular person's name. Just three subtopic candidates of the original 30 candidates from BaiduPedia covered a subtopic of 歌曲 (songs), another subtopic candidate covered a subtopic of 歌曲资源信息 (songs' information), and the others were not covered. Because these four subtopic candidates were similar to each other, they were ranked lower in the final results.
- In the case of topic IMINE2-C-074 "白眉大侠单田芳", the judged subtopics were downloads, Chinese storytelling, videos, listening to recordings online, adapted dramas, and resources. Just one subtopic candidate of the original 30 candidates from BaiduPedia covered a subtopic of 评书 (Chinese storytelling).
- In the case of topic IMINE2-C-023 "圣诞节怎么过", the judged subtopics were regions, methods, romances, lovers, decorations, event marketing, and gifts. The subtopic candidates for the topic from our run did not cover these subtopics at all.
- In the case of topic IMINE2-C-066 "爱回家粤语", the judged subtopics were downloads, watching videos online, video tapes, and related information. The subtopic candidates for the topic from our run covered only the subtopic of 在线观看 (watching videos online).

Evaluation results of the Japanese language runs

| | I-rec@10 | D-nDCG@10 | D#-nDCG@10 | |
|---|---|---|---|---|
| IRCE-QU-J-1S | 0.4102 | 0.2706 | 0.3404 | ($\alpha = 0.8$) |
| IRCE-QU-J-2S | 0.4043 | 0.3167 | 0.3605 | ($\alpha = 0.2$) |
| IRCE-QU-J-3S | 0.3900 | 0.3300 | 0.3600 | ($\alpha = 0.5$) |
| IRCE-QU-J-4S | **0.4169** | 0.3100 | 0.3634 | (-) |
| IRCE-QU-J-5S | 0.3903 | **0.3387** | **0.3644** | ($\alpha = 0.0$) |

Evaluation results of D#-nDCG@10 per topic type

| | Ambiguous | Faceted | Task-oriented | Vertical-oriented |
|---|---|---|---|---|
| IRCE-QU-J-1S | 0.4233 | 0.3722 | 0.2376 | 0.3388 |
| IRCE-QU-J-2S | 0.4572 | **0.4178** | 0.2491 | 0.3362 |
| IRCE-QU-J-3S | 0.4894 | 0.3977 | 0.2265 | **0.3397** |
| IRCE-QU-J-4S | 0.4736 | 0.3954 | **0.2502** | 0.3335 |
| IRCE-QU-J-5S | **0.4901** | 0.4090 | 0.2343 | 0.3305 |

## Summary

- Extraction of subtopic candidates seems to be insufficient.
  - ✓ The performance of the methods seems to depend on richness and granularity of the original resources.
  - ✓ Adding other resources, such as query suggestions, remains as future works.


Contact: songximei@slis.tsukuba.ac.jp / masao@slis.tsukuba.ac.jp

Github: https://github.com/cres-project/irce-wikipedia