RUCIR at NTCIR-12 IMINE-2 Task

Ming Yue, Zhicheng Dou, Sha Hu, Jinxiu Li, Xiaojie Wang, Ji-Rong Wen

Renmin University of China, Beijing, China

{yomin, dou, wangxiaojie}@ruc.edu.cn {sallyshahu, jinxiu2216, jirong.wen}@gmail.com

Query Understanding Subtask

Step 1: Subtopic Candidate Retrieval

- Query Suggestion (Bing, Yahoo!, Baidu, Google)
- Disambiguation page: Wikipedia, Baidu Baike

Step 2: Subtopic Candidate Clustering

- Use top 300 search results to represent a subtopic
- Generate a tf-idf vector for each subtopic candidate
- Clustering: K-medoids, Quality Threshold (QT)
- Average linkage
- A cluster -> a subtopic (the centroid)

Step 3: Subtopic Ranking and Diversification

- Rank the subtopics using MMR

$$d_{i+1} = \arg\max\{\lambda Rel(d) + (1-\lambda) Div(d, D_i)\}$$

- Relevance: beta*NumOfCandidates+(1-beta)*IAL
- Novelty: cosine similarity of tf-idf vectors

Step 4: Vertical Intent Classification

- (1) Rule-based
  - Keyword-based rules (true if x% results contain specific keywords), such as:
    - what/how -> Encyclopedia
    - New, latest, daily -> News
    - Sale, deal, coupon -> Shopping
- (2) Trained Classifier
  - Use IMine-1 topics and their query suggestions as training data
  - Use Bing to generate labels: for a query, check whether Bing's SERP include answers/results from a specific vertical
  - Use word occurrences as features
  - SVM

Run Name	System Description	QU-score
rucir-Q-C/E-1Q	Suggestions + Wikipedia, k-medoids, trained classifier	**0.5757** **0.5613**
rucir-Q-C/E-2Q	Suggestions + Wikipedia, k-medoids, rule-based classifier	0.5495 0.5904
rucir-Q-C/E-3Q	Suggestions + Wikipedia, QT clustering, trained classifier	0.4489 0.4166
rucir-Q-C/E-4Q	Suggestions , k-medoids, trained classifier	0.5311 0.5583
rucir-Q-C/E-5Q	Suggestions + Ranking + Diversification	0.6849 0.6911

Results

- (1) Trained classifiers >Rules
- (2) Clustering: K-medoids > QT
- (3) Query suggestions + Wikipedia > Query suggestions

Vertical Incorporating Subtask

Subtopic expansion

Subtopic
PlayStation 4

⇩

Bing's Related Suggestions
PlayStation 4 new video game
PlayStation 4 best price
PlayStation 4 controller

⇩

Expanded Subtopic
PlayStation 4 video game best price controller

Key Choices

- Subtopic is short, hence we expand it to a longer query containing more keywords
- Use BM25 as the basic ranking function
- Assume that the virtual vertical result is highly relevant to the corresponding subtopic, and it is treated as a normal document.

Diversification model – PM2

$$q_i = \frac{w_i}{2s_i + 1}$$

$$d^* = \arg\max_{d \in D}[\lambda \cdot q^* \cdot rel(d, t^*) + (1-\lambda) \cdot \sum_{t_i \neq t^*} q_i \cdot rel(d, t_i)]$$

$$s_i = s_i + \frac{rel(d^*, t_i)}{\sum_{rel(d,t_j)>0} rel(d^*, t_j)}$$

In an iterative procedure, first we compute the quotient q_i for each subtopic t_i. Then we check the unselected documents to select the next best document d*. And finally we update the occupied seat s_i.

Experimental Results

	Chinese Unclear Queries		
Run Name	D♯-nDCG@10	D-nDCG@10	I-Recall
rucir-V-C/E-1M* [SExp+QU]	**0.7395**$^{\star\circ\Delta\dagger}$	**0.5342**$^{\circ\Delta\dagger}$	**0.9449**$^{\star\circ\Delta\dagger}$
rucir-V-C/E-2M* [SExp+Sug]	0.7079†	0.5268$^{\circ\Delta\dagger}$	0.8890
rucir-V-C/E-3M° [noSExp+QU]	0.6884	0.4682	0.9086
rucir-V-C/E-4M$^{\Delta}$ [noSExp+Sug]	0.6801	0.4799	0.8802
rucir-V-C/E-5M† [Baseline]	0.6593	0.4444	0.8742

	English Unclear Queries		
Run Name	D♯-nDCG@10	D-nDCG@10	I-Recall
rucir-V-C/E-1M* [SExp+QU]	**0.8249**$^{\star\circ\Delta\dagger}$	**0.6565**$^{\star\Delta\dagger}$	**0.9933**$^{\circ\Delta}$
rucir-V-C/E-2M* [SExp+Sug]	0.7807	0.5912	0.9701
rucir-V-C/E-3M° [noSExp+QU]	0.7994	0.6534$^{\star\Delta\dagger}$	0.9454
rucir-V-C/E-4M$^{\Delta}$ [noSExp+Sug]	0.7719	0.5847	0.9591
rucir-V-C/E-5M† [Baseline]	0.7800	0.5723	0.9876

Results

- (1) Using subtopics mined by **rucir-Q-C/E-1Q** is better than directly using suggestions
- (2) Subtopic expansion works well