Introduction

- In information retrieval system, users often submit the query which is a short description by natural language, and they decide the relevance of document not based on semantics of query terms in documents, but existence of query terms.
- The key terms extracted from a question in IR4QA can be different with distinct segmentation strategies.
- The query expansion is an effective way to solve term mismatch problem by expanding the key terms with a certain number of other related terms in the initial query.
- Wikipedia is a multilingual, web-based, free-content encyclopedia project, and each of its article provides information to explain the term of the article title.

In order to expand the most relevant terms, we make use of related Wikipedia article as the “seed” document of the related question. And then make question expansion based on the most relevant paragraph of the question.

System Architecture

Query Expansion

1. Question Classification

   - 《千里走单骑》和张艺谋是什么关系？
   - 高仓健是谁？

   RELATIONSHIP
   BIOGRAPHY

2. Article Retrieval

   (1) Use different template to extract name entities from different types of question.
   (2) Take retrieval in the Wikipedia by the name entities to get the related article. If cannot find, relocate the final article by the most relevant title.

3. Paragraph Location

   Question: 第76届奥斯卡最佳男主角是谁？
   
   (1) 高仓健
   (2) 《千里走单骑》
   (3) 《海上钢琴师》
   (4) 《盲战》

   Located Paragraph

   - 高仓健の主演作品
   - 《海上钢琴师》

   4. Query Expansion

   Question: 李永波和中国羽毛球队是什么关系？
   Initial Query Expression: 李永波，中国，羽毛球，球队
   New Query Expression: 李永波，中国，羽毛球，球队，1962年，著名，羽毛球运动，球运，运动员，教练员，辽宁，宁人，现为，国家队，总教练，国家，乒羽，中心，副主任

   Term Weight:

   \[
   w(q|Q) = p \cdot w(q|D) \cdot k \cdot \text{avg(boost)} \cdot \frac{\text{score}(q)}{\text{MaxScore}}
   \]

   Where \( w(q|Q) \) is the weight of key term \( q \) in the original query \( Q \), \( w(q|D) \) is the weight of \( q \) in document \( d \), \( n \) is the number of top selected documents, and \( P \) and \( k \) are experimentally determined positive constants. \( \text{boost} \) is one factor as a multiplier besides the factors \( sf \) and \( idf \) to compute the weight of the query key term in the initial query, and the \( \text{avg(boost)} \) is the average value of them.

Experiments

<table>
<thead>
<tr>
<th>RUN</th>
<th>Analysis File</th>
<th>Mean AP</th>
<th>Mean Q</th>
<th>Mean nDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>WUST-CS-01-T</td>
<td>No</td>
<td>0.2694</td>
<td>0.293</td>
<td>0.4881</td>
</tr>
<tr>
<td>WUST-EN-01-T</td>
<td>No</td>
<td>0.1037</td>
<td>0.1206</td>
<td>0.2815</td>
</tr>
<tr>
<td>WHUQA-EN-CS-03-T.xml</td>
<td>0.1435</td>
<td>0.1564</td>
<td>0.292</td>
<td></td>
</tr>
</tbody>
</table>

Formal run experiment official results (AFTER bug fix)

The system does not achieve a good official result.
- As our system segment the question into words based on the same dictionary in the index processing module. If the key term does exist in the dictionary, our system may not retrieve the related document.
- In our “EN-CS” work, we extract English key terms and then translate them into Chinese by Google translation. So the quality of the translation determines the performance of our EN-CS result.

Conclusions

- In order to solve the problems of inappropriate key terms exacted from the initial question and term mismatch, we apply query expansion technique to get more useful key terms for the query based on related Wikipedia article content.
- We find that when the question types are DEFINITION, BIOGRAPHY, PERSON, ORGANIZATION, LOCATION or DATE, we can find the relevant paragraph easily.
- The Wikipedia is a document sets of description type, so it performs better for the explanation questions.