Extracting Lack of Information on Wikipedia by Comparing Multilingual Articles

☆Yuya Fujiwara      Konan University(Japan)
Yu Suzuki            Nagoya University(Japan)
Yukio Konishi       Konan University(Japan)
Akiyo Nadamoto Konan University(Japan)
Background 1

• Many people all over the world use Wikipedia on the Internet.
• An important policy of Wikipedia is that the contents of articles is the same for all language version.
• Articles of the same topic of any language version are expected to have exactly identical contents except for language.

This policy is not obeyed, especially for culture-related topics.
The content of article about “Fish and Chips” is very rich in the English version, but poor in the Japanese version. Because “Fish and Chips” is a very popular dish in the U.K., but not in Japan. *There are some lack of information on one language Wikipedia, however there may be rich information on other language Wikipedia.*
Propose

• If there are lack of information in Wikipedia article, we complement it in native language version.
Our Flow

1. Users input a query in their native language to the system.
2. The system retrieves one native article of which title is the same as the user’s input query.
3. It translates the query to the non-native language using a language dictionary and retrieves a non-native article whose the title is the same as the user’s input query.
4. It extracts comparison articles from the non-native articles using a Wikipedia link graph.
5. It compares a native article with non-native articles extracted in 4. and extracts lack of information.
6. It browses lack of information available on the web.
Our Flow

1. Users input a query in their native language to the system.
2. The system retrieves one native article of which title is the same as the user’s input query.
3. It translates the query to the non-native language using a language dictionary and retrieves a non-native article of which the title is the same as the user’s input query.
4. It extracts comparison articles from the non-native articles using a Wikipedia link graph.
5. It compares a native article with non-native articles extracted in 4. and extracts lack of information.
6. It browses lack of information available on the web.

← We focus on this step.

Granularity of information differs between the languages in Wikipedia.
– For example:

• Article of "Cricket" is written about Batting of Cricket both Japanese Wikipedia and English Wikipedia.

• In English Wikipedia, there are another page about detail of Batting of Cricket.

When we compare a native article with non-native articles, we have to consider multiple comparison non-native articles.

We extract target articles based on the Wikipedia link graph and our proposed relevance degree.
Extract comparison target articles

We create a link graph for non-native Wikipedia based on the user’s input query.

1. We extracts an articles having the same title as the user’s input from the comparison Wikipedia. Basic article $\rightarrow$ root node

2. We extracts all interactive linked articles and they become nodes in link graph.

3. We calculates the relevance degree between the root node and the other nodes in the link graph.

4. When the relevance degree is greater than a threshold $\theta$ value, then we regard the articles as relevant articles.

：root node (basic article)

：other node

<table>
<thead>
<tr>
<th>Batting</th>
<th>Cricket</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.6$</td>
<td>compare</td>
</tr>
</tbody>
</table>

クリケット (In Japanese)
Calculating Relevance Degree

• Extraction of the relevance article using only cosine similarity between root node and the other nodes.
  ⇒ The result of recall ratio is **not good**

• **Relevance Degree** between root node and the other nodes.

<table>
<thead>
<tr>
<th>Position of the link anchor</th>
<th>The Important anchor appears in the summary area in Wikipedia.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of the link anchor</td>
<td>Important anchors related to the basic article appear many times in the basic article.</td>
</tr>
<tr>
<td>Similarity between the content</td>
<td>If articles are similar, relevance degree becomes high.</td>
</tr>
</tbody>
</table>
Calculating Relevance Degree

The system divides the basic article according to the structure of the table of contents of the basic article. The divided parts → segments.
Calculating Relevance Degree

• The system extracts a title from a node, which is the interactively linked article.
• The title becomes a keyword used for extraction of the anchor text from the basic article.
• The title → target title.
The system counts the anchor text of the target title in the summary area of the basic article.

It also counts the anchors in each segment of the basic article.
The system calculates the similarity between interactively linked article and the summary area of basic article.

- The system calculates the similarity between interactively linked article and the segment area.
Calculating Relevance Degree

\[ R_i = \{\alpha \cdot (TF_{sum_i} \cdot S_{sum_i}) + \sum_{k=1}^{n} (TF_{ik} \cdot S_{ik})\} / \max(R_{im}) \]

- \( i \): the identification number of the interactively linked article
- \( R_i \): Relevance Degree of article \( i \)
- \( TF_{sum_i} \): number of the anchor in summary area
- \( S_{sum_i} \): the similarity between \( i \) and the summary area
- \( TF_{ik} \): number of the anchor in the segment \( k \)
- \( S_{ik} \): the similarity between \( i \) and the segment \( k \)
- \( K \): the segment number
- \( N \): the number of segments in the basic article
- \( \alpha \rightarrow 3.0 \quad \beta \rightarrow 0.2 \)

Positions:
- Link anchor position
- Number of the link anchor
- Similarity between the content
1. Users input a query in their native language to the system.

2. The system retrieves one native article of which title is the same as the user’s input query.

3. It translates the query to the non-native language using a language dictionary and retrieves a non-native article of which the title is the same as the user’s input query.

4. It extracts comparison articles from the non-native articles using a Wikipedia link graph.

5. **It compares a native article with non-native articles extracted in 4. and extracts lack of information.**

6. It browses lack of information available on the web.
Comparison between native article and non-native articles

- Almost all Wikipedia articles are divided into segments based on the table of contents meaning that the segments are divided semantically.

- When comparing the similarity of multilingual Wikipedia, we examine the segment of the table of contents of Wikipedia.

- If the similarity of a content is lower than all content, we extract the content as lack of information.

Ex: Fish and chips

Compare

Segment

History

Main article: British cuisine

Fish and chips became a stock meal among the working classes in Great Britain as city dwellers during the second half of the 19th century. In 1850, the first fish and chip shop in England was opened. Deep-fried chips (slices or pieces of potato) as a dish may have first appeared in Britain, earliest usage of “chips” in this sense the mention in Dickens' A Tale of Two Cities (1859) "drops of oil". (Note: that British tradition, as presented in an manuscript of 1781, dates 1859.)

Lack of information

Segment
Experiment 1

• We confirmed the availability of extracting relevant articles in non-native articles.
  – We compare our method with baseline.
  – The baseline is the cosine similarity.
  – using precision, recall, and F-measure by comparing our proposed method with the baseline.
## Result of Experiment 1

<table>
<thead>
<tr>
<th>Query</th>
<th>#</th>
<th>Baseline</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision(%)</td>
<td>Recall(%)</td>
</tr>
<tr>
<td>Bannock (food)</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Warwick Castle</td>
<td>2</td>
<td>15</td>
<td>100</td>
</tr>
<tr>
<td>Black dog (ghost)</td>
<td>7</td>
<td>67</td>
<td>29</td>
</tr>
<tr>
<td>Fish and chips</td>
<td>4</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>Goodwood Festival of Speed</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bowls</td>
<td>2</td>
<td>33</td>
<td>100</td>
</tr>
<tr>
<td>Burlesque</td>
<td>3</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>Flag of Scotland</td>
<td>6</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Gaelic handball</td>
<td>4</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Kipper</td>
<td>3</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td>National Gallery of Scotland</td>
<td>12</td>
<td>72</td>
<td>67</td>
</tr>
<tr>
<td>Lipton</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Average</td>
<td>-</td>
<td>37</td>
<td>45</td>
</tr>
</tbody>
</table>

 #:Number of correct results
Experiment 2

• We confirmed the accuracy of extracting lack of information.
  – We use English Wikipedia as native article and Japanese Wikipedia as non-native articles.
  – The correct answer was judged by a bilingual person.
### Result of Experiment 2

<table>
<thead>
<tr>
<th>Query</th>
<th>#</th>
<th>Precision(%)</th>
<th>Recall(%)</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bannock (food)</td>
<td>2</td>
<td>33</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Warwick Castle</td>
<td>12</td>
<td>79</td>
<td>92</td>
<td>85</td>
</tr>
<tr>
<td>Black dog (ghost)</td>
<td>32</td>
<td>89</td>
<td>78</td>
<td>83</td>
</tr>
<tr>
<td>Fish and chips</td>
<td>11</td>
<td>45</td>
<td>82</td>
<td>58</td>
</tr>
<tr>
<td>Goodwood Festival of Speed</td>
<td>10</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Bowls</td>
<td>9</td>
<td>50</td>
<td>100</td>
<td>67</td>
</tr>
<tr>
<td>Burlesque</td>
<td>22</td>
<td>71</td>
<td>45</td>
<td>56</td>
</tr>
<tr>
<td>Flag of Scotland</td>
<td>56</td>
<td>98</td>
<td>88</td>
<td>92</td>
</tr>
<tr>
<td>Gaelic handball</td>
<td>16</td>
<td>68</td>
<td>94</td>
<td>79</td>
</tr>
<tr>
<td>Kipper</td>
<td>16</td>
<td>88</td>
<td>94</td>
<td>91</td>
</tr>
<tr>
<td>National Gallery of Scotland</td>
<td>4</td>
<td>57</td>
<td>100</td>
<td>72</td>
</tr>
<tr>
<td>Lipton</td>
<td>8</td>
<td>71</td>
<td>63</td>
<td>67</td>
</tr>
<tr>
<td>Average</td>
<td>–</td>
<td>67</td>
<td>79</td>
<td>71</td>
</tr>
</tbody>
</table>

#:Number of correct results
Discussion of Experiment 2

• They are almost good result.

• Bad result case
  – When we target on tea brand of “Lipton”, we extract “Thomas Lipton” as a relevant article. He created the Lipton tea brand. He is also famous for sportsman. It is not related to the tea brand of “Lipton”. But we extract it as a lack of information for the tea brand of “Lipton”.

• Other case is attributable to a translation problem.
Conclusion and Future work

• We proposed a method for extracting information that exists in one language version, but which does not exist in another language version.

• Two points
  – Examine the link graph of Wikipedia and structure of and article of Wikipedia.
    • Extract comparison target articles of Wikipedia using our proposed degree of relevance.
  – Compare between native article and non-native articles.

• Future work
  – Considering word sense disambiguation.
  – Comparing another languages (ex. Chinese, Korean etc...)