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.

Background 2

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.

Query: Fish and Chips

Japanese user

Japanese version

目次

1 概略
2 歴史
3 食べ方
4 関連項目

poor

English version

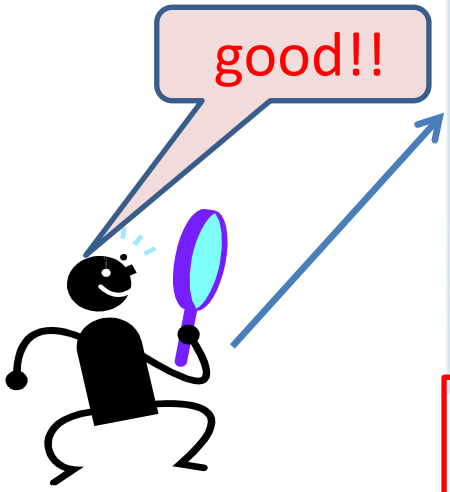
Contents

1 History
1.1 England
1.2 Scotland
1.3 Ireland
2 Composition
2.1 Cooking
2.2 Thickness
2.3 Batter
2.4 Choice of fish
2.5 Accompaniments
3 Vendors
4 Cultural impact
5 Environment
6 See also
7 Footnotes
8 External links

rich

Propose

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

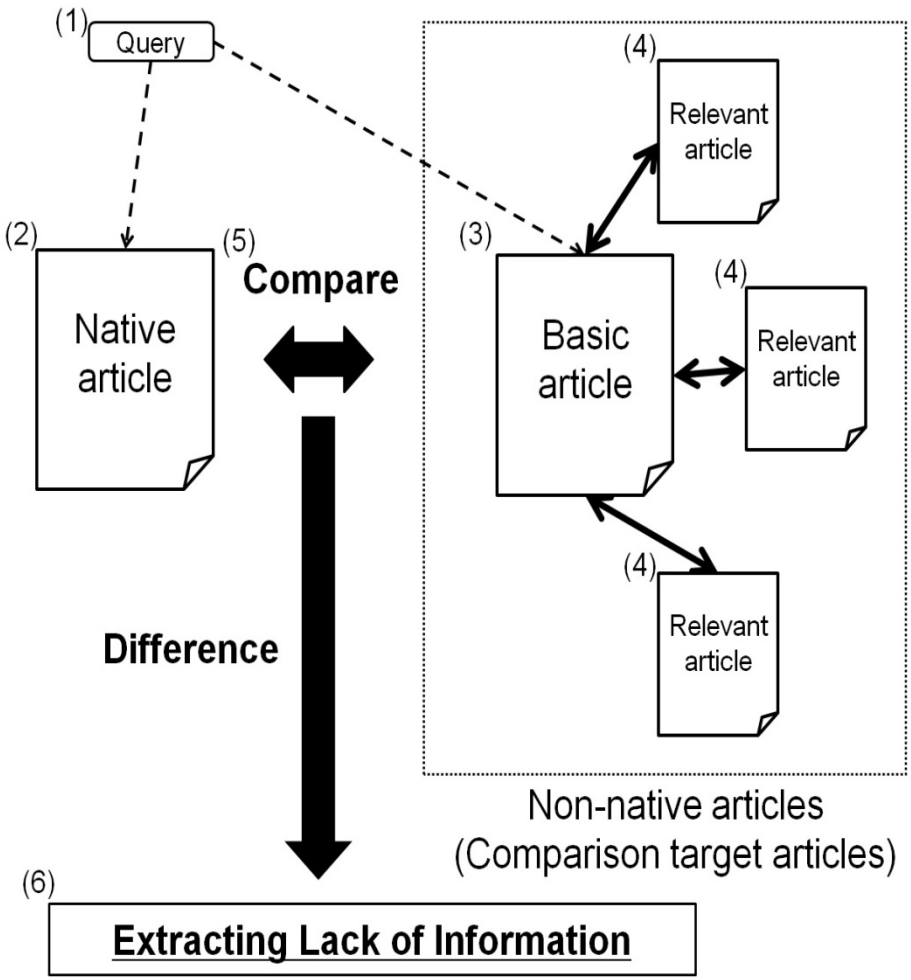


complement

Extracting Lack of information on Wikipedia by Comparing Multilingual Articles

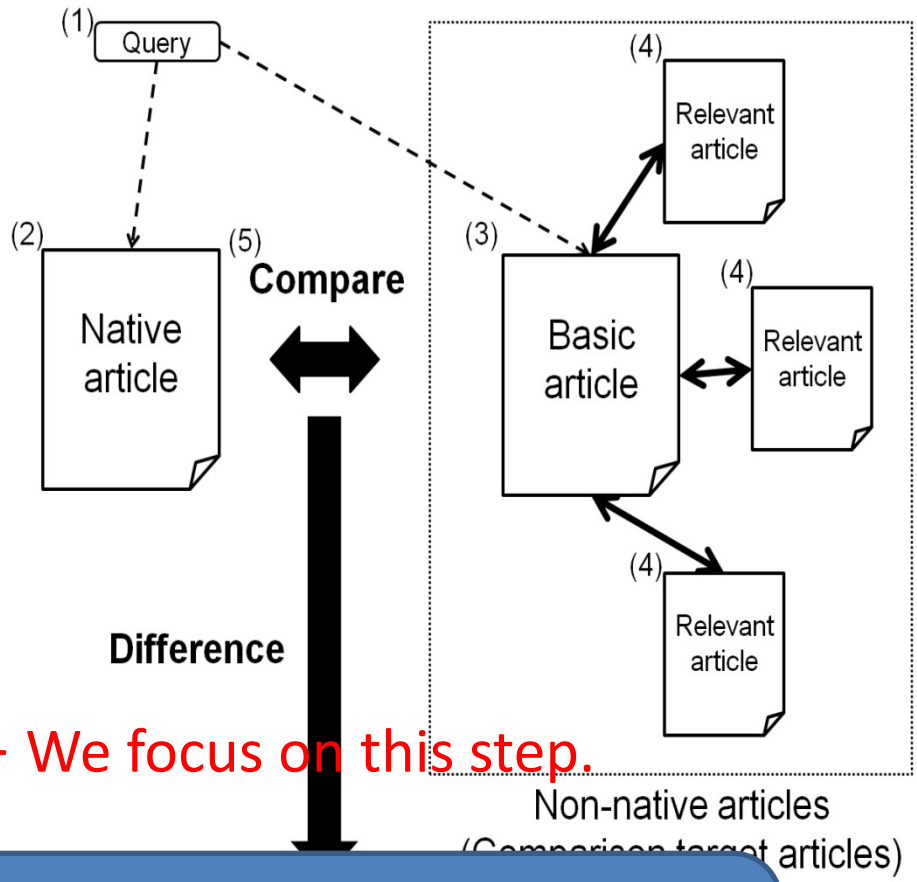
Our Flow

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

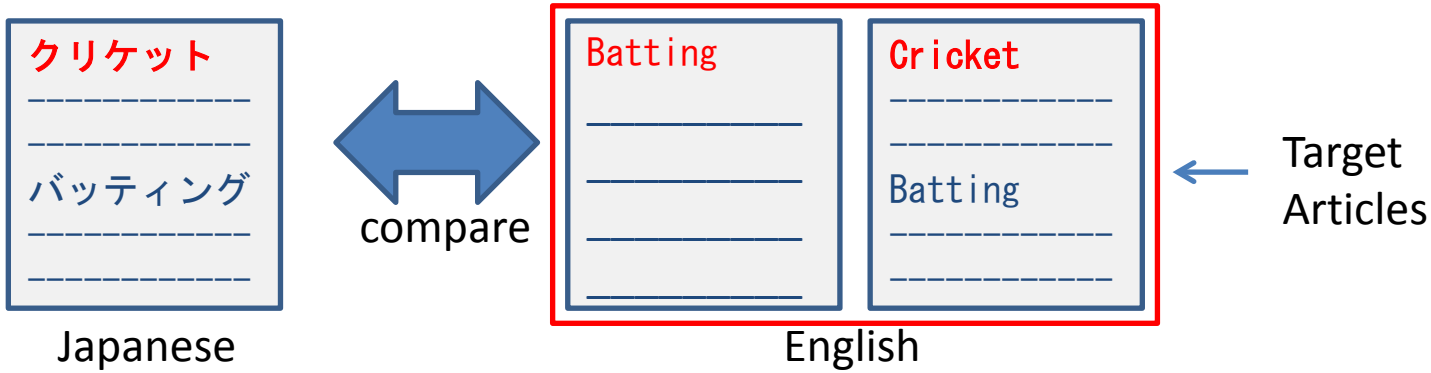
1. **Users input a query in their native language to the system.**
2. **The system retrieves one native article of which title is the same as the user's input query.**
3. **It translates the query to the non-native language using a language dictionary and retrieves a non-native article of which the title is the same as the user's input query.**
4. **It extracts comparison articles from the non-native articles using a Wikipedia link graph.** ← We focus on this step.
5. **It compares a native article with no ar**
6. **It av**



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.



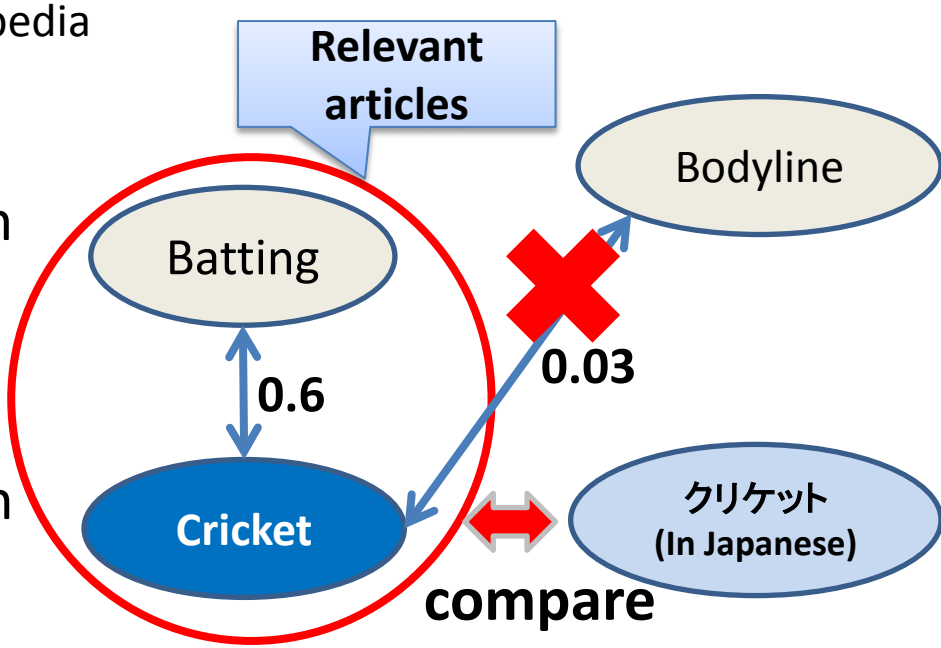
When we compare a native article with non-native articles, we have to consider multiple comparison non-native articles.

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→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 *β* value, then we regard the articles as relevant articles.



:root node(basic article)

:other node

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.

Position of the link anchor	The Important anchor appears in the summery area in Wikipedia.
Number of the link anchor	Important anchors related to the basic article appear many times in the basic article.
Similarity between the content	If articles are similar, relevance degree becomes high.

Calculating Relevance Degree

ex:Cricket

Contents [hide]

Rules and Game-play

Main article: Laws of cricket

Segment

Summary

A cricket match is played between two teams of eleven players each[19][20] on a grassy field, typically 137-150 metres (15
Laws of Cricket do not specify the size or shape of the field[22] but it is often oval.

A cricket match consists of multiple periods called *innings*. During an innings, one team plays defense (the *fielding* team) ar
team). The two teams switch between fielding and batting after each innings. All eleven members of the fielding team take
of the batting team (two batsmen) are on the field at any given time.

The key action takes place in the pitch, a rectangular strip in the centre of the field. The two batsmen face each other at
fielding team's eleven members stand outside the pitch, spread out across the field.

Behind each batsman is a target called a *wicket*. One designated member of the fielding team, called the *bowler*, is given a b
the ball from one end of the pitch to the wicket behind the batsman on the other side of the pitch. The batsman tries to p
wicket by hitting the ball with a bat. If the bowler succeeds in hitting the wicket, or if the ball, after being hit, is caught by t
touches the ground, the batsman is *dis*[illegible] the field, to be replaced by another batsma

Segment

If the batsman is successful in hitting the ball away from the wicket without the ball being caught before it hits the ground,
score points (*runs*) for their team by running across the pitch, switching positions. Each switch of positions is worth one ru
multiple runs or they may attempt no runs. By attempting runs, the batsmen risk dismissal, which can happen if the fielding
reached that end of the pitch.

If the batsman hits the bowled ball over the field boundary without the ball touching the field, the batting team scores six r
the boundary, the batting team scores four runs and may not attempt more. When the batsmen have finished attempting th
continues to bowl toward the same wicket, regardless of any switch of the batsmen's positions.[23]

After a bowler has bowled six times (an *over*), another member of the fielding team is designated as the new bowler. The ne
may bowl multiple times during an innings, but may not bowl two overs in succession.

The innings is complete when 10 of the 11 members of the batting team have been dismissed or a set number of overs has
depending on the match.

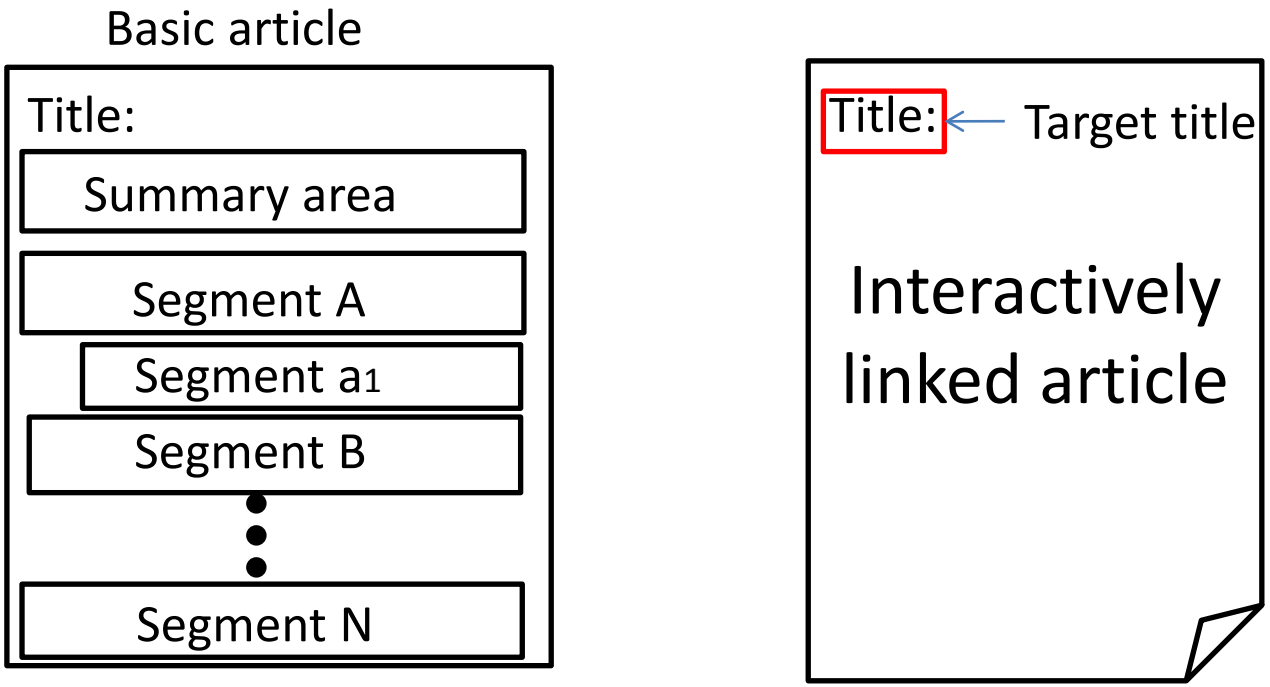
Objectives

The objective of each team is to score [illegible] completely dismiss the other team. In limite
the opposition has not been completely dismissed. [illegible] cricket, it is necessary to score the most runs and dismiss the

Segment

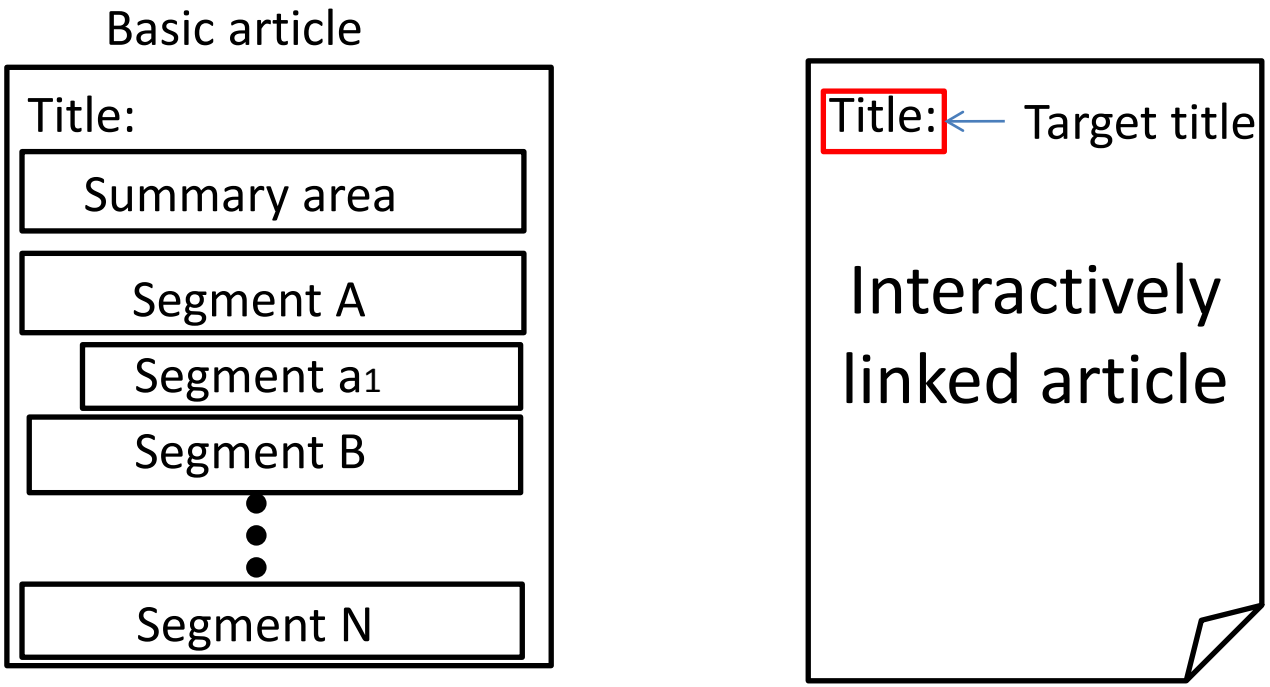
The system divides the basic article according to the structure of the table of contents of the basic article.
The divided parts → segments.

Calculating Relevance Degree



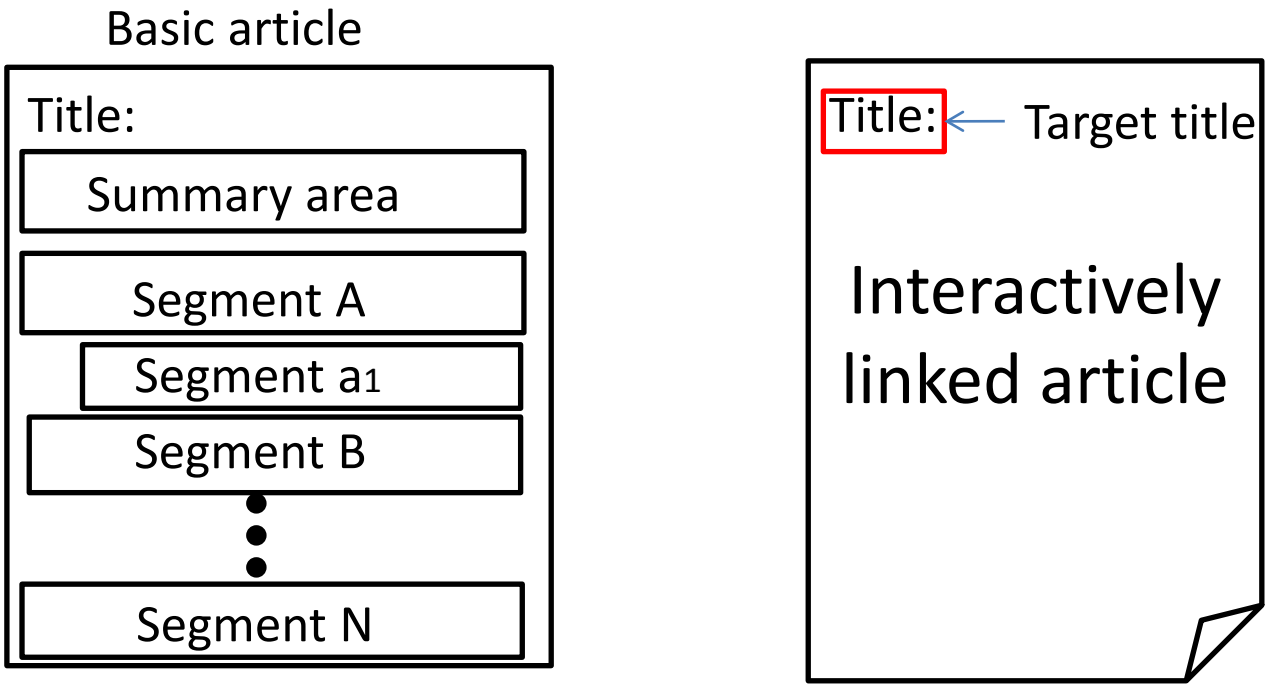
- The system extracts a title from a node, which is the interactively linked article.
- The title becomes a keyword used for extraction of the anchor text from the basic article.
- The title → target title.

Calculating Relevance Degree



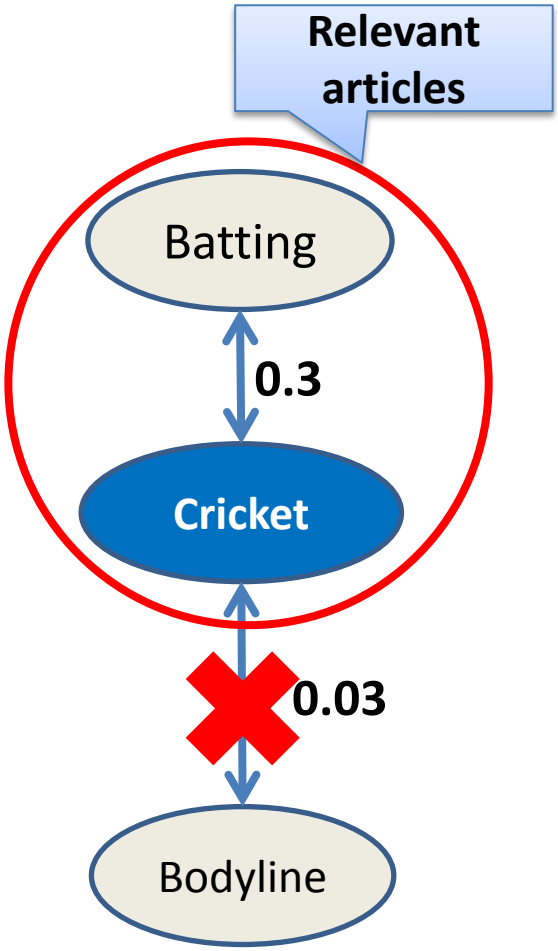
- The system counts the anchor text of the target title in the summary area of the basic article.
- It also counts the anchors in each segment of the basic article.

Calculating Relevance Degree



- 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



:root node

:other node

Position of the link anchor

Number of the link anchor

Similarity between the content

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

i: the identification number of the interactively linked article

Ri: Relevance Degree of article i

TFsumi: number of the anchor in summary area

Ssumi: the similarity between *i* and the summary area

TFik: number of the anchor in the segment k

Sik: the similarity between i and the segment k

K: the segment number

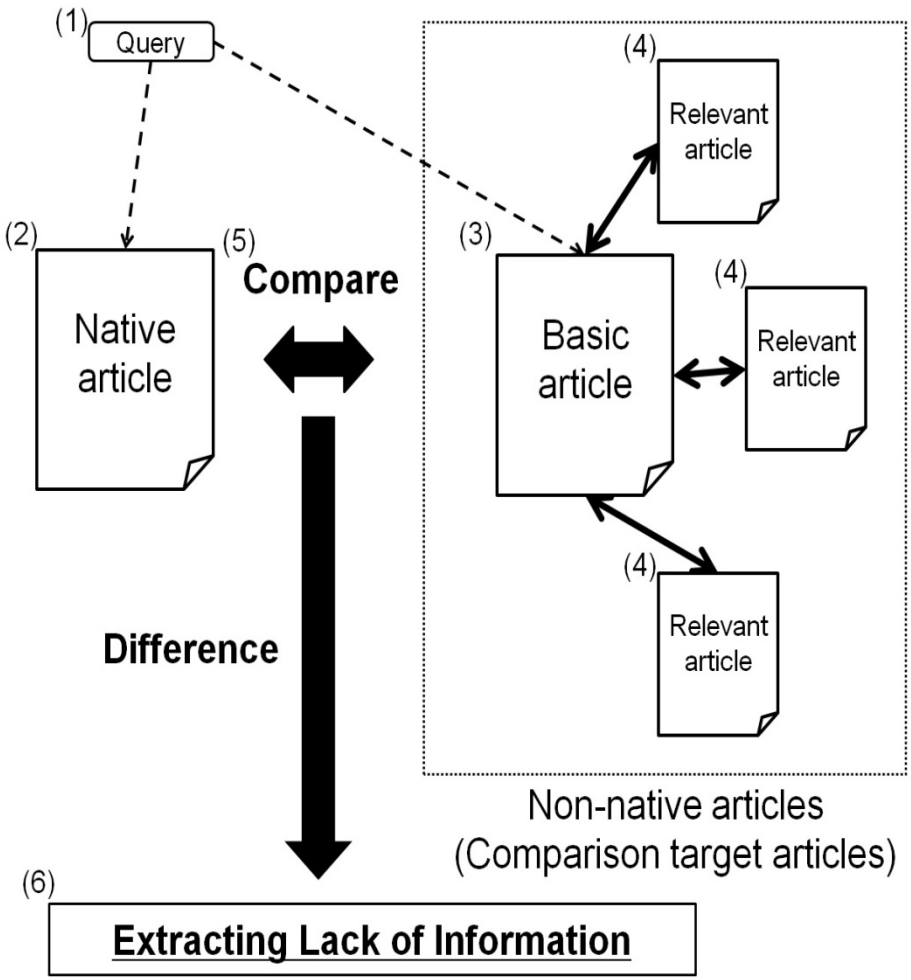
N: the number of segments in the basic article

Max(Rim): the maximum value in all Ri

$\alpha \rightarrow 3.0$ $\beta \rightarrow 0.2$

Our Flow

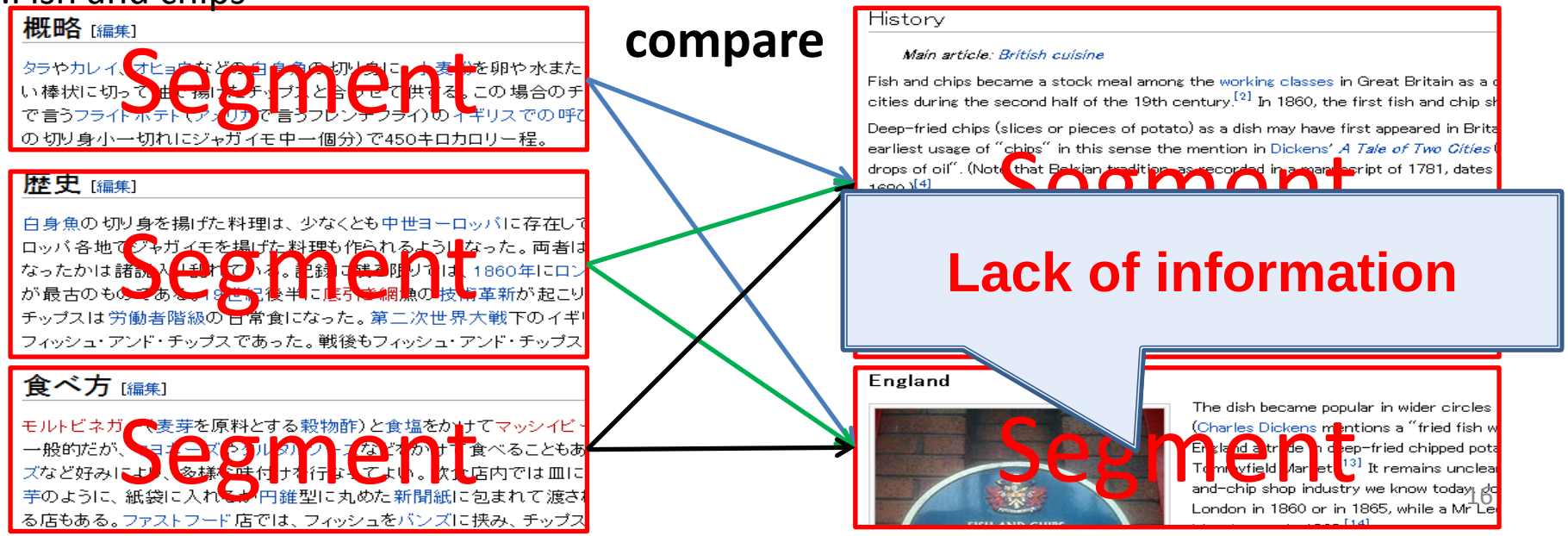
1. Users input a query in their native language to the system.
2. The system retrieves one native article of which title is the same as the user's input query.
3. It translates the query to the non-native language using a language dictionary and retrieves a non-native article of which the title is the 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.



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



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

Query	#	Baseline			Proposed		
		Precision(%)	Recall(%)	F-measure	Precision(%)	Recall(%)	F-measure
Bannock (food)	2	0	0	0	20	50	28
Warwick Castle	2	15	100	27	25	100	40
Black dog (ghost)	7	67	29	40	100	29	44
Fish and chips	4	40	50	44	50	75	60
Goodwood Festival of Speed	2	0	0	0	50	50	50
Bowls	2	33	100	50	10	50	14
Burlesque	3	60	50	55	100	67	80
Flag of Scotland	6	50	50	50	67	33	44
Gaelic handball	4	25	25	25	80	100	89
Kipper	3	67	67	67	100	67	80
National Gallery of Scotland	12	72	67	70	75	50	60
Lipton	1	0	0	0	25	100	40
Average	–	37	45	36	59	64	52

#: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

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

#: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…)