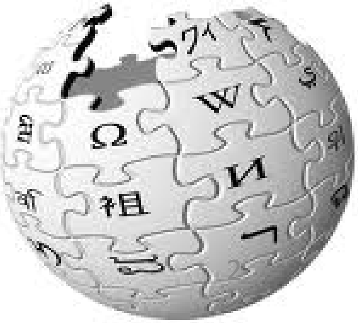
Extracting Difference Information from Multilingual Wikipedia

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

Background 1

- Wikipedia, a large encyclopedia that is accessible using the Internet.
- Wikipedia has two characteristics.
  - Anyone can edit the content .
  - There are 250 over language versions.
- The content of articles differ among about respective language version.

Native language version should have difference information.

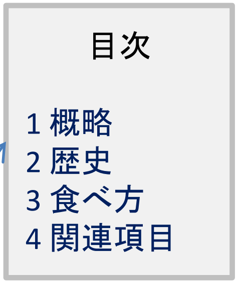


Background 2

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.
In this way , **Native language version has difference information**.

Query: Fish and Chips

Japanese version

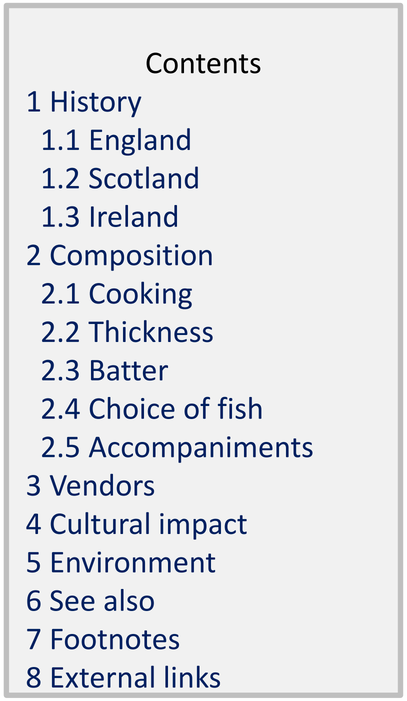


目次

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

Japanese user

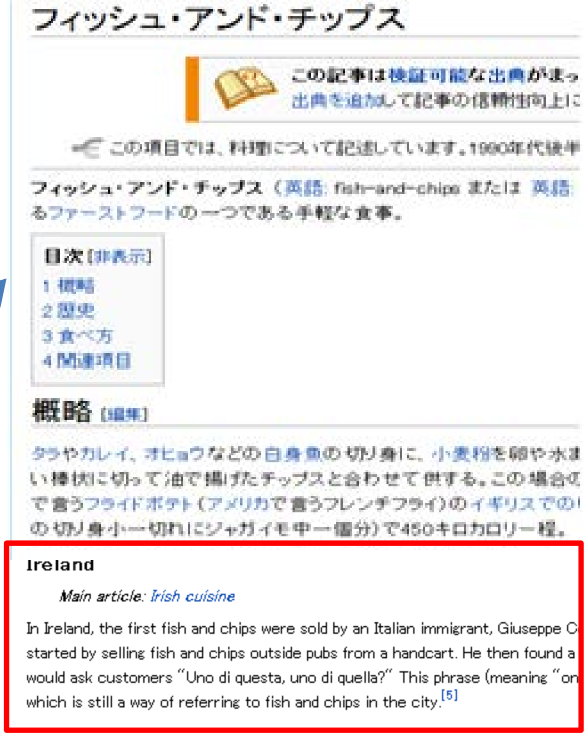
Background 3

- Generally, users browse the Wikipedia of their own native language.
- Occasionally they would refer to the other language versions.
- It is difficult for them to understand the whole content which is written by non-native language.

It may be able to understand a passage of non-native language Wikipedia.

Propose

- We consider that if there are difference information, we add it in browsing language version.



フィッシュ・アンド・チップス

この記事は検証可能な出典がまっ
出典を追加して記事の信頼性向上に

この項目では、料理について記述しています。1990年代後半

フィッシュ・アンド・チップス（英語: fish-and-chips または 英語
るファーストフードの一つである手軽な食事。

目次 [非表示]
1 概略
2 歴史
3 食べ方
4 関連項目

概略 [編集]

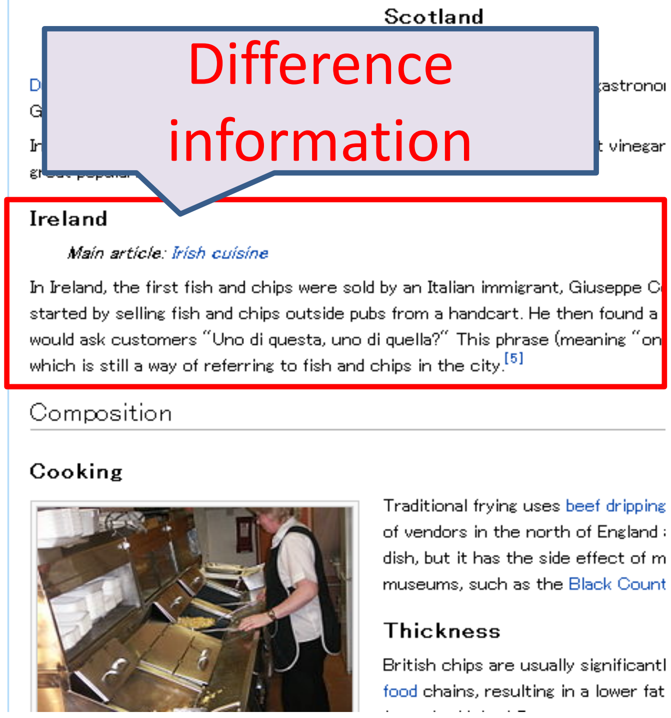
タラやカレイ、オヒョウなどの白身魚の切り身に、小麦粉を卵や水ま
い棒状に切って油で揚げたチップスと合わせて供する。この場合の
で言うフライドポテト（アメリカで言うフレンチフライ）のイギリスでの
の切り身小一切れにジャガイモ中一個分）で450キロカロリー程。

Ireland

Main article: Irish cuisine

In Ireland, the first fish and chips were sold by an Italian immigrant, Giuseppe C
started by selling fish and chips outside pubs from a handcart. He then found a
would ask customers "Uno di questa, uno di quella?" This phrase (meaning "on
which is still a way of referring to fish and chips in the city.[5]

add



Difference information

Scotland

D
G
Ir
great popular

astronoi
t vinegar

Ireland

Main article: Irish cuisine

In Ireland, the first fish and chips were sold by an Italian immigrant, Giuseppe C
started by selling fish and chips outside pubs from a handcart. He then found a
would ask customers "Uno di questa, uno di quella?" This phrase (meaning "on
which is still a way of referring to fish and chips in the city.[5]

Composition

Cooking

Traditional frying uses beef dripping
of vendors in the north of England
dish, but it has the side effect of m
museums, such as the Black Count

Thickness

British chips are usually significantl
food chains, resulting in a lower fat

Extracting Difference information from Multilingual Wikipedia

Naive Method

1. User inputs keyword in Japanese to the system.
2. The system retrieves one article related to keyword from the Japanese version of Wikipedia.
3. The system extracts the English article by using interlanguage link in Wikipedia.
4. The system compares Japanese article with English comparison target articles.
5. The system outputs Japanese article with sections of English articles that do not appear in the Japanese article.

Naive Method

1. User inputs keyword in Japanese to the system.

2. The system retrieves one article related to keyword from the Japanese version of Wikipedia.

3. The system extracts the English article by using interlanguage link in Wikipedia.← We focus on this step.

4. The system compares Japanese article with English

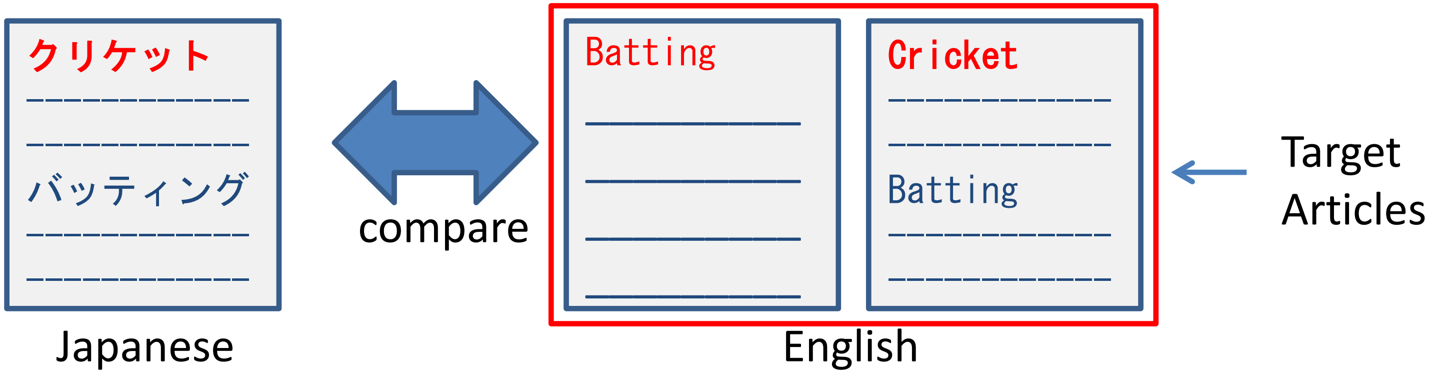
Granularity of information differs between the languages in Wikipedia.

5

le.

– 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.



クリケット

バッティング

compare

Batting

Cricket

Batting

Target Articles

Japanese

English

When we compare a Japanese Wikipedia with English Wikipedia, we have to consider multiple comparison English articles we call these articles “Target Article”.

Our Flow

1. Users input keywords in Japanese to the system.

2. The system retrieves one article related to keywords from the Japanese version of Wikipedia.

3. The system extracts the multiple English comparison target articles.

4. The system compares sections in Japanese article and those in English articles, and detects which sections appeared in English articles but not in Japanese articles.

5. The system outputs Japanese article with sections of English articles that do not appear in the Japanese article.

Our Flow

1. Users input keywords in Japanese to the system.
2. The system retrieves one article related to keywords from the Japanese version of Wikipedia.
3. The system extracts the multiple English comparison target articles.
4. The system compares sections in Japanese article and

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

English articles that do not appear in the Japanese article.

Extract comparison target articles

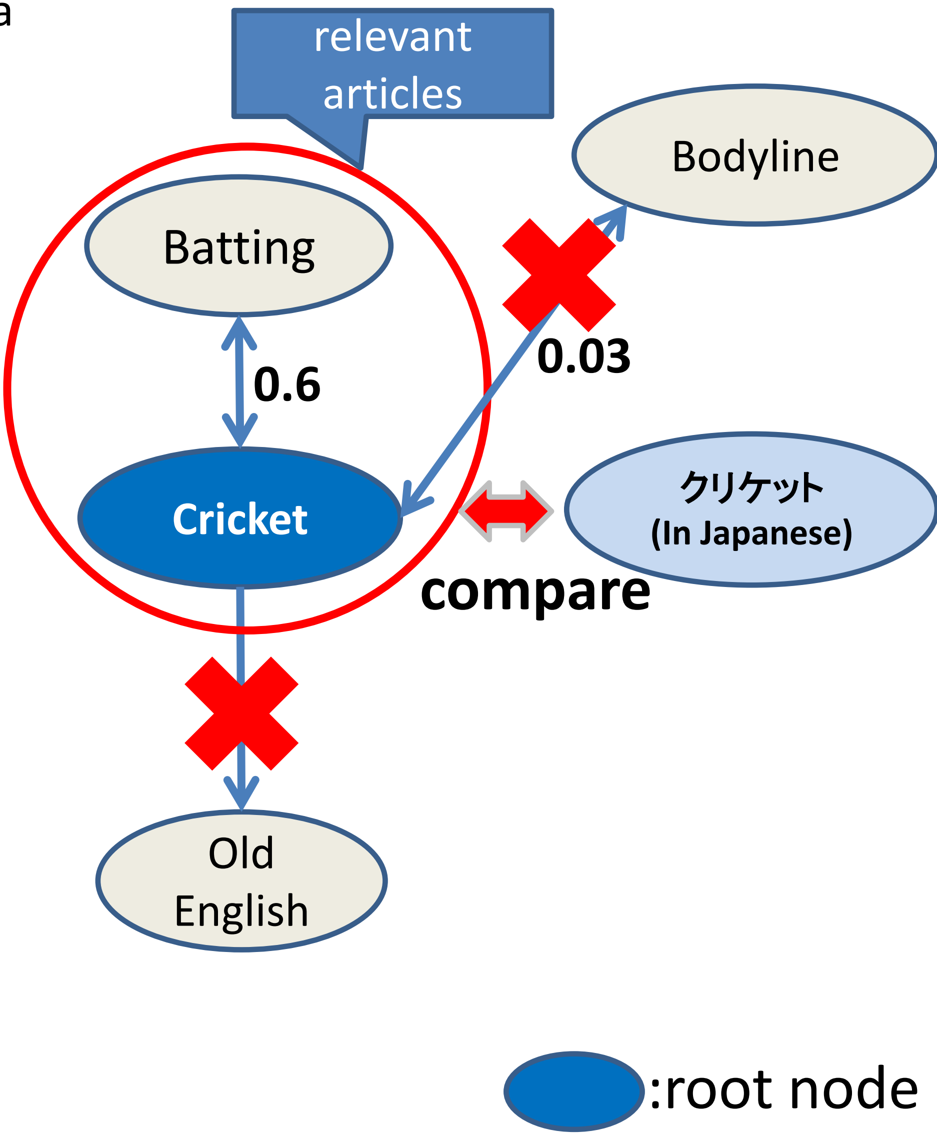
We create a link graph for English Wikipedia based on the user's input keyword.

1. The system extracts English articles that have the same title as the user's input and translated. We designate the English article as the basic article. We regard the basic article as the root node of the link graph.

2. The system extracts all interactive linked articles that are the subjects of link-out and link-in connections with the root node. Then it includes the articles as nodes and there by creates the link graph.

3. The system calculates the relevance degree between the root node and respective articles in the link graph.

4. When the relevance degree is greater than a threshold α value, then the system regards the articles as relevant articles.

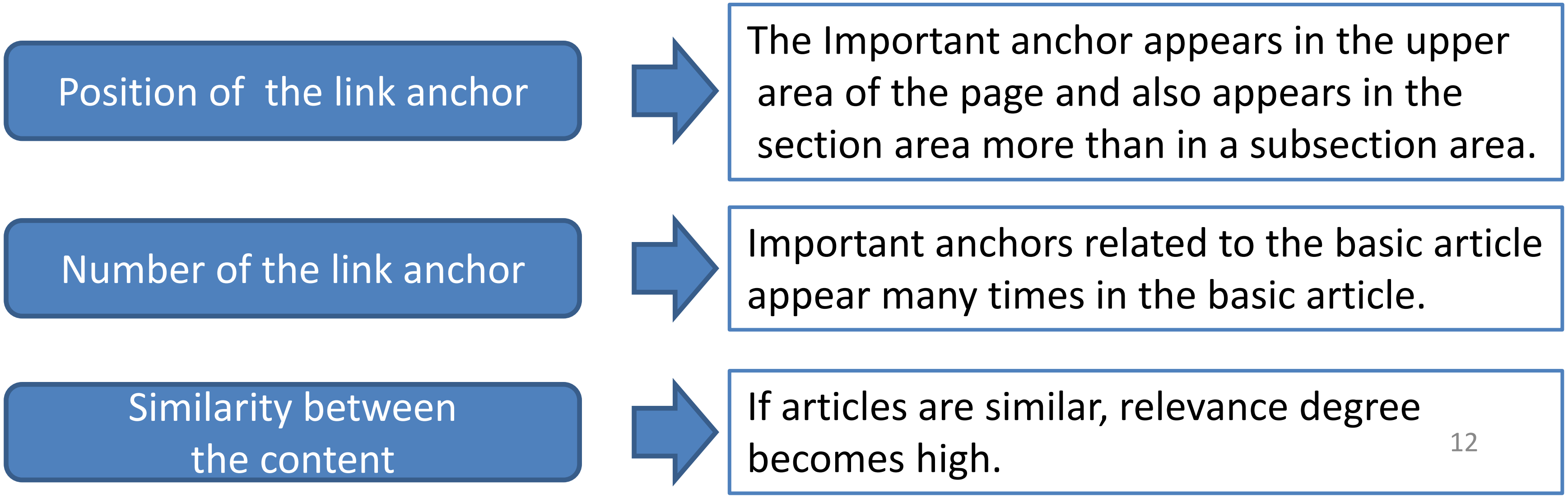


Calculating Relevance Degree

- We proposed extraction of the relevance article using only cosine similarity between root node and the other nodes.

  ⇒The result of recall ratio is not good

- We propose Relevance Degree between root node and the other nodes.

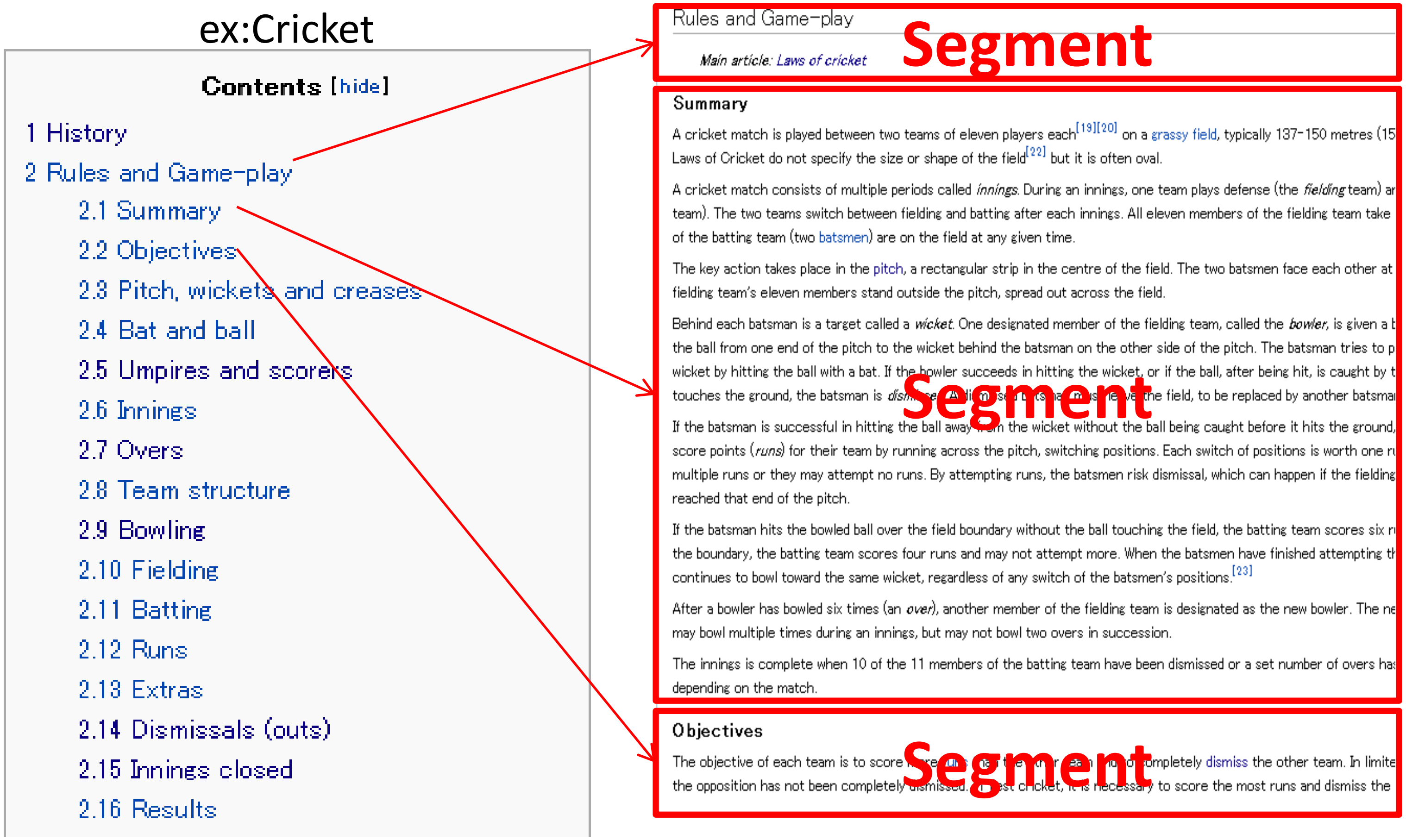


Position of the link anchor → The Important anchor appears in the upper area of the page and also appears in the section area more than in a subsection area.

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



Contents [hide]

1 History
2 Rules and Game-play
    2.1 Summary
    2.2 Objectives
    2.3 Pitch, wickets and creases
    2.4 Bat and ball
    2.5 Umpires and scorers
    2.6 Innings
    2.7 Overs
    2.8 Team structure
    2.9 Bowling
    2.10 Fielding
    2.11 Batting
    2.12 Runs
    2.13 Extras
    2.14 Dismissals (outs)
    2.15 Innings closed
    2.16 Results

Rules and Game-play

Main article: Laws of cricket

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 *dismissed*. A dismissed batsman must leave the field, to be replaced by another batsmai

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 ru
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.

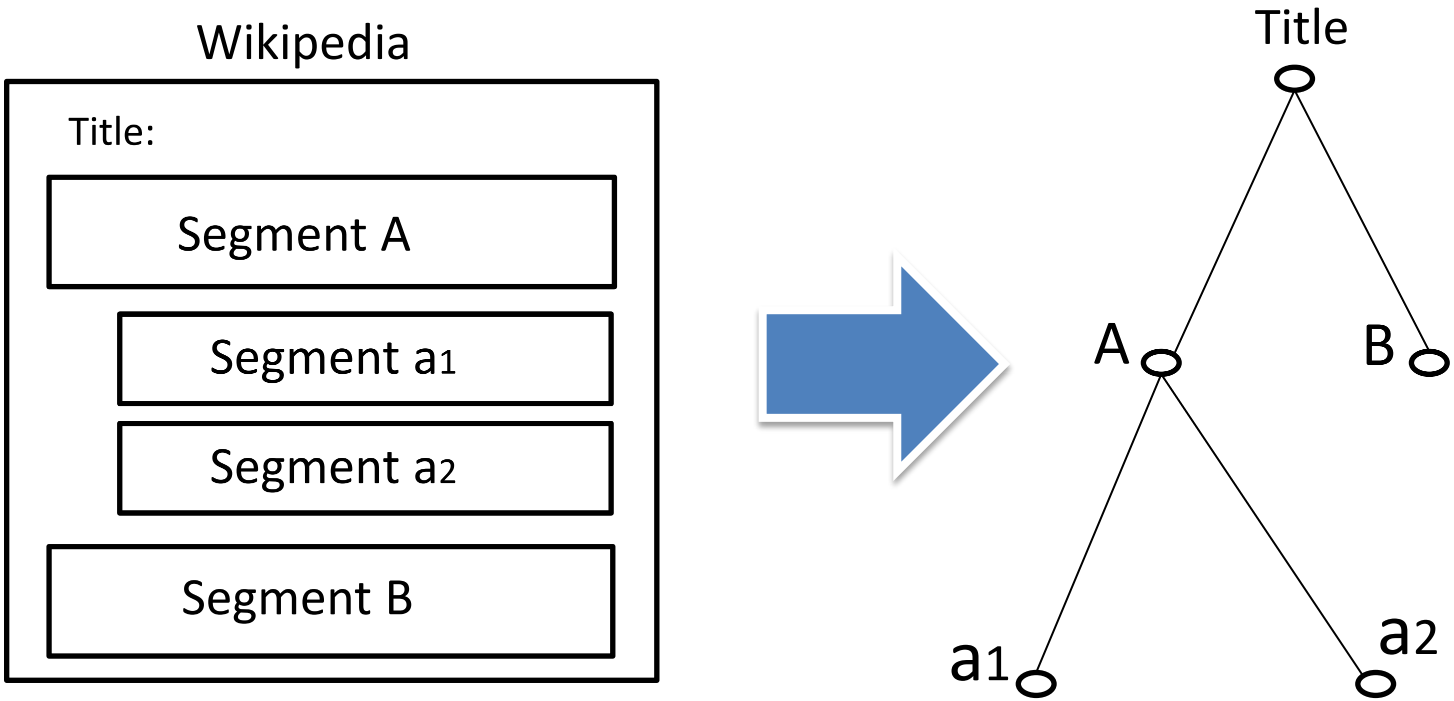
The innings is complete when 10 of the 11 members of the batting team have been dismissed or a set number of overs has
depending on the match.

Objectives

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

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

Calculating Relevance Degree



Wikipedia

Title:

Segment A

Segment a1

Segment a2

Segment B

Title

A

B

a1

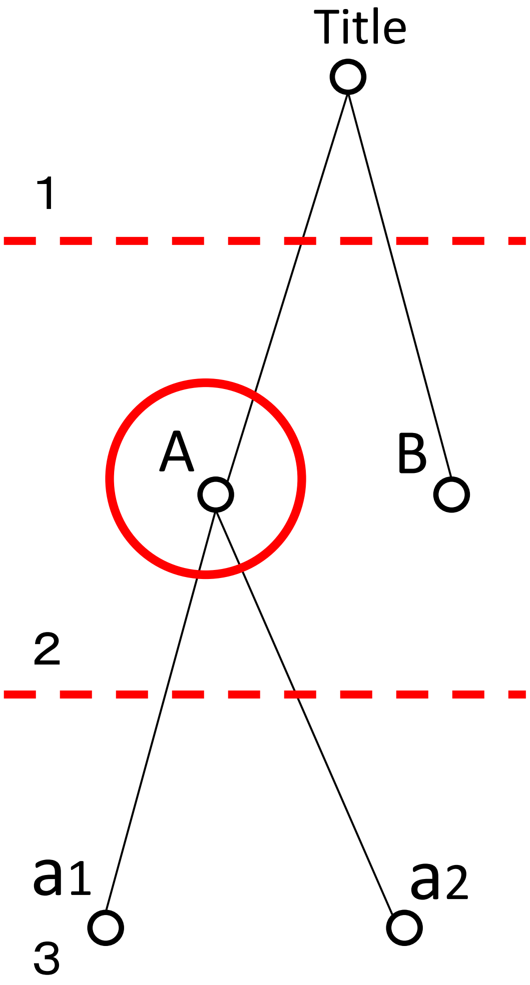
a2

The system creates segment tree for which the root node is the title name of the basic article, child nodes are segments.

In the segment tree, the child node of the root node is a section of the basic article; the grandchild node is a subsection of the parent node.

The left side nodes show younger segment numbers.

Calculating Relevance Degree

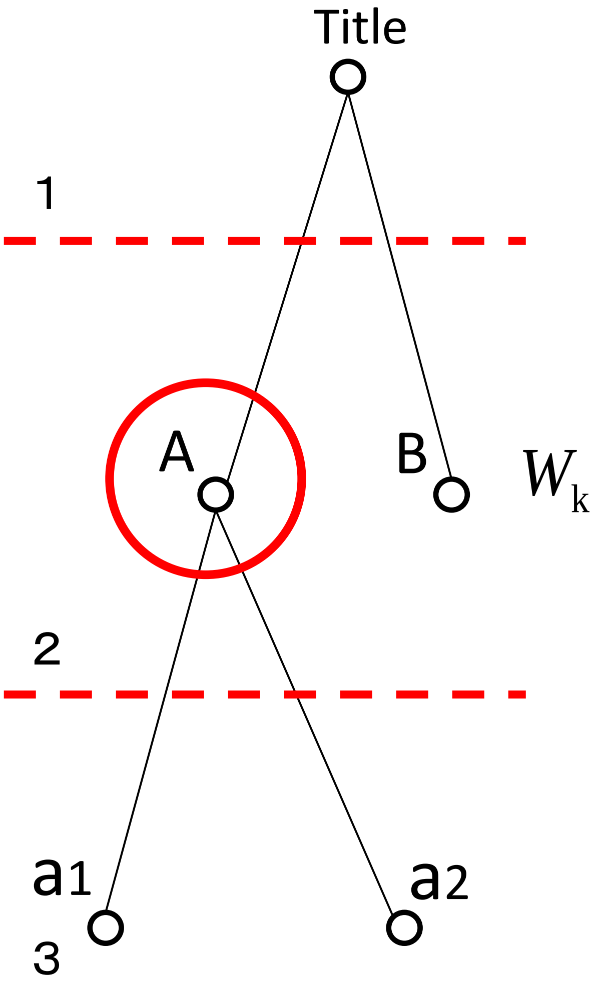


- Our hypothesis is that important anchors related to the basic article appear many times in the basic article.

- They appear in upper areas of the basic article, and also appear in the section area more than subsection area.

- We infer that the relevance degree of the nodes of the low hierarchy and left side in the segment tree are higher than the nodes of the deep hierarchy and right side in the segment tree.

Calculating Relevance Degree



Position of the link anchor

Number of the link anchor

Similarity between the content

$$W_{kl} = af \times S_{kl} + \sum_{i=1}^{af} \left\{ \left(\frac{1}{d_i} \right)^{n_i} \times \left(n_i - o_i + 1 \right) \right\} / \max(W_{km})$$

W_{kl}:Relevance Degree of article k and l

af:The number of the anchor

d_i:depth

n_i:sibling node of i

o_i:sequence number of l in the same depth

S_{kl}:The content is similarity between article of k and l

Our Flow

1. Users input keywords in Japanese to the system.

2. The system retrieves one article related to keywords from the Japanese version of Wikipedia.

3. The system extracts the multiple English comparison target articles.

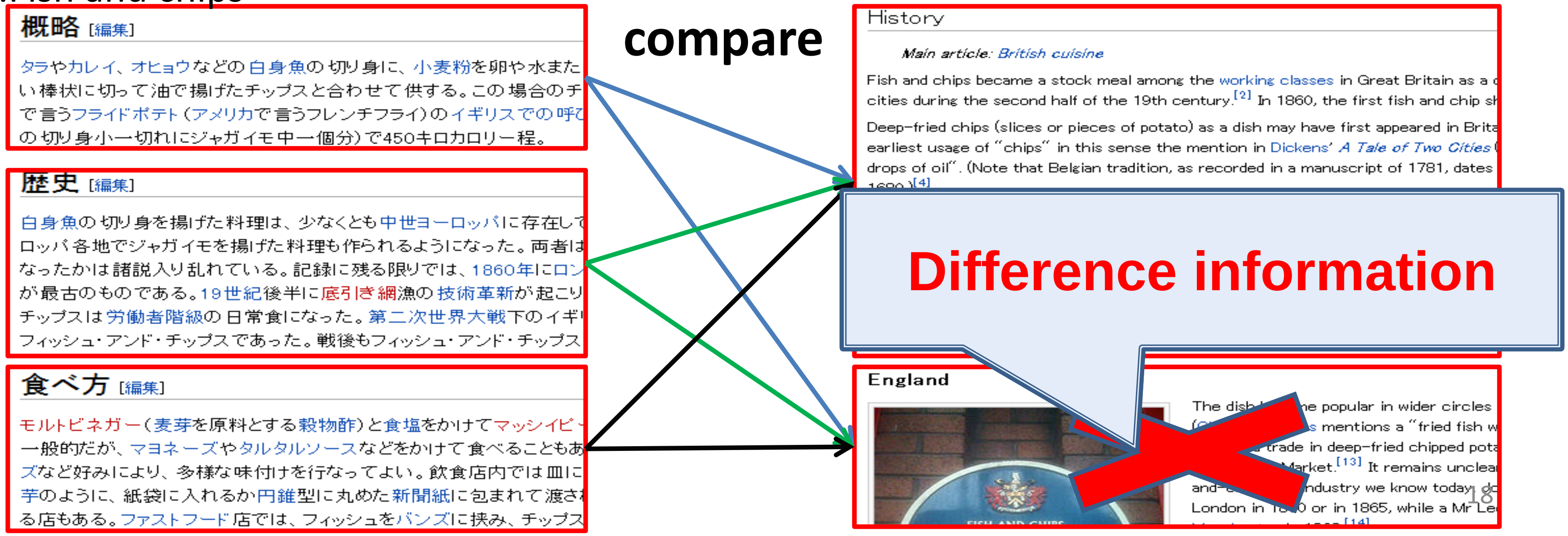
4. The system compares sections in Japanese article and those in English articles, and detects which sections appeared in English articles but not in Japanese articles.

5. The system outputs Japanese article with sections of English articles that do not appear in the Japanese article.

Comparison between Japanese article and English 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 to all content, we extract the content as difference information.

Ex:Fish and chips



compare

概略 [編集]

タラやカレイ、オヒョウなどの白身魚の切り身に、小麦粉を卵や水また
い棒状に切って油で揚げたチップスと合わせて供する。この場合のチ
で言うフライドポテト（アメリカで言うフレンチフライ）のイギリスでの呼び
の切り身小一切れにジャガイモ中一個分）で450キロカロリー程。

歴史 [編集]

白身魚の切り身を揚げた料理は、少なくとも中世ヨーロッパに存在して
ロッパ各地でジャガイモを揚げた料理も作られるようになった。両者は
なったかは諸説入り乱れている。記録に残る限りでは、1860年にロン
が最古のものである。19世紀後半に底引き網漁の技術革新が起こり
チップスは労働者階級の日常食になった。第二次世界大戦下のイギ
フィッシュ・アンド・チップスであった。戦後もフィッシュ・アンド・チップス

食べ方 [編集]

モルトビネガー（麦芽を原料とする穀物酢）と食塩をかけてマッシイビ
一般的だが、マヨネーズやタルタルソースなどをかけて食べることもあ
ズなど好みにより、多様な味付けを行なってよい。飲食店内では皿に
芋のように、紙袋に入れるか円錐型に丸めた新聞紙に包まれて渡さ
る店もある。ファストフード店では、フィッシュをバンズに挟み、チップス

History

Main article: British cuisine

Fish and chips became a stock meal among the working classes in Great Britain as a
cities during the second half of the 19th century.[2] In 1860, the first fish and chip sh

Deep-fried chips (slices or pieces of potato) as a dish may have first appeared in Brita
earliest usage of "chips" in this sense the mention in Dickens' *A Tale of Two Cities*
drops of oil". (Note that Belgian tradition, as recorded in a manuscript of 1781, dates
1680.)[4]

Difference information

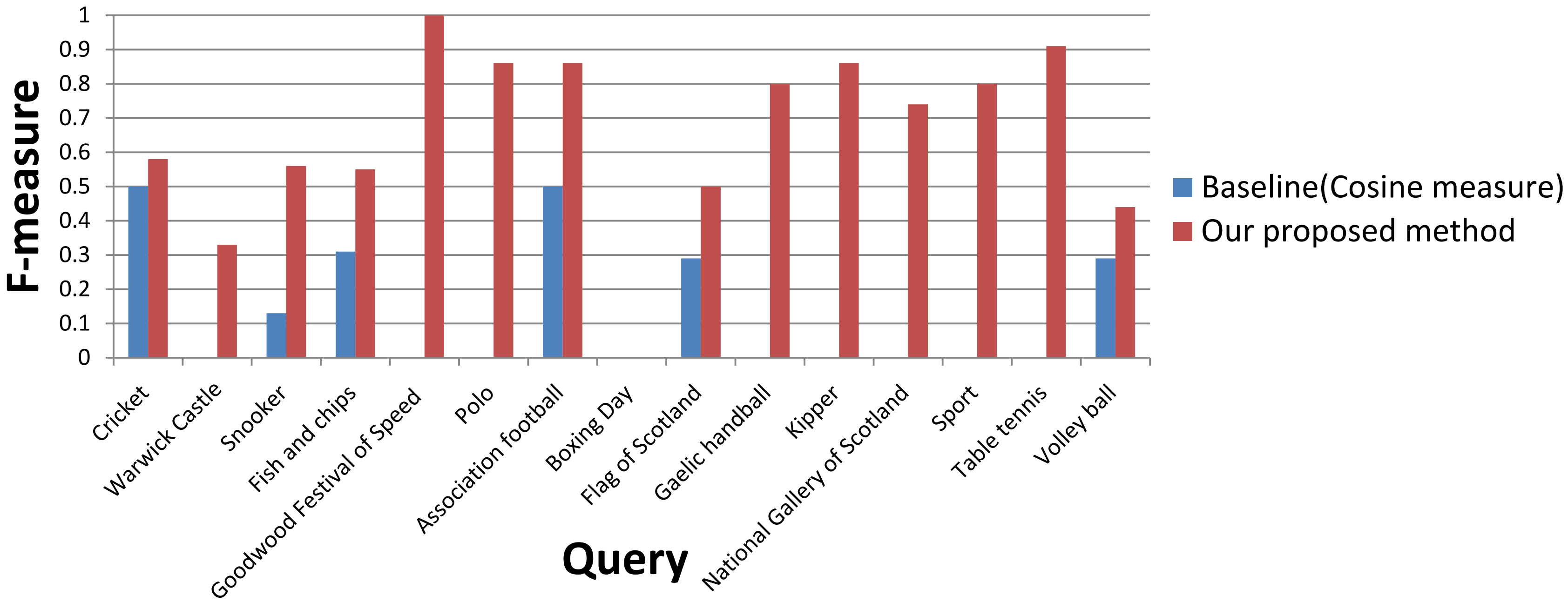
England

The dish ne popular in wider circles
(s mentions a "fried fish w
trade in deep-fried chipped pota
Market.[13] It remains unclea
and- ndustry we know today,
London in 0 or in 1865, while a Mr Le

Experiment

- We confirmed the accuracy of difference information extraction methods.
  - Using precision, recall and F-measure by comparison of our proposed method with the baseline, which is only the cosine similarity between basic articles and other linked articles.
  - Compared Wikipedia
    - We used Japanese and English Wikipedia for experiment
  - We set the threshold of relevance degree to 0.2 based on the result of our pre-experiment.

Experiment



When we compare the results of baseline with our proposed method,
we can observe that the average of precision ratio improves from 0.37 to 0.68,
and that the average of recall ratio improves from 0.1 to 0.61.

Moreover, the average of *F-measure improves* from 0.13 to 0.61.

From these improvements, we confirmed that our proposed method can extract appropriate parts of articles from English Wikipedia articles.

Result and discussion

- Average of precision ratio: 0.37⇒0.68
- Average of recall ratio: 0.1⇒0.61
- Average of F-measure: 0.13⇒0.61

- Discussion
  - when we use the queries “Association Football” and “Table Tennis,” which are generally used terms, our proposed system is more effective than when we use terms for specific fields, because when we use general terms, the relevant terms are numerous. Also, the links to Wikipedia articles are numerous.
  - Therefore, the relevance of correct sections is large.

Conclusion

- We proposed a method for extracting difference information.

- 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 Japanese article and English articles.

Future work

- Calculating credibility of Wikipedia articles.
  - Wikipedia credibility is not always good.
  - For that reason, we must assess Wikipedia credibility in future studies.
- Considering word sense disambiguation.
  - Many instances of word sense disambiguation exist, but we ignored such cases in this time.
  - We intend to consider word sense disambiguation in later investigations.