言語間比較によるWikipediaの補完情報抽出手法の提案

☆藤原　裕也　(甲南大学)
鈴木 優　(名古屋大学)
小西 幸男　(甲南大学)
灘本 明代　(甲南大学)

背景

- **Wikipedia**
  - 特徴
    - 284以上の多言語版が存在
    - 誰でも記事を編集することが可能
  - 問題
    - ある話題に対しての情報が不足している記事が多く存在

情報補完

    - 他のWebから情報補完する
    - 他の言語版から情報補完する　etc…

言語版によって書いてある内容が異なる

例:日本の文化

検索Query:たこ焼き

英語版

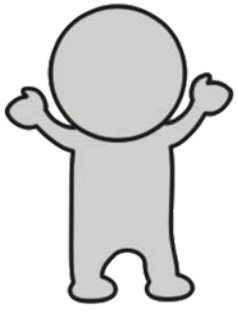
Contents

1 Takoyaki pan
2 See also
3 References

日本語版

目次

1 概要
2 材料
 2.1 生地
 2.2 具
 2.3 味付け
3 食べ方
4 歴史
5 大阪のたこ焼き
6 器具
 6.1 業務用
 6.2 家庭用
7 日本国外におけるたこ焼き
 7.1 朝鮮文化圏
 7.2 中国文化圏
 7.3 東南アジア
8 関連事項



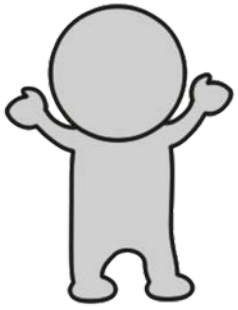
外国人ユーザ

理由
•編集者が少ない
•十分な知識を持っていない

補完

目的

閲覧しているWikipediaに不足している情報を
他言語Wikipediaから補完する



外国人ユーザ

理由
- 編集者が少ない
- 十分な知識を持っていない

英語版

Contents

1 Takoyaki pan
2 See also
3 References

比較

日本語版

目次

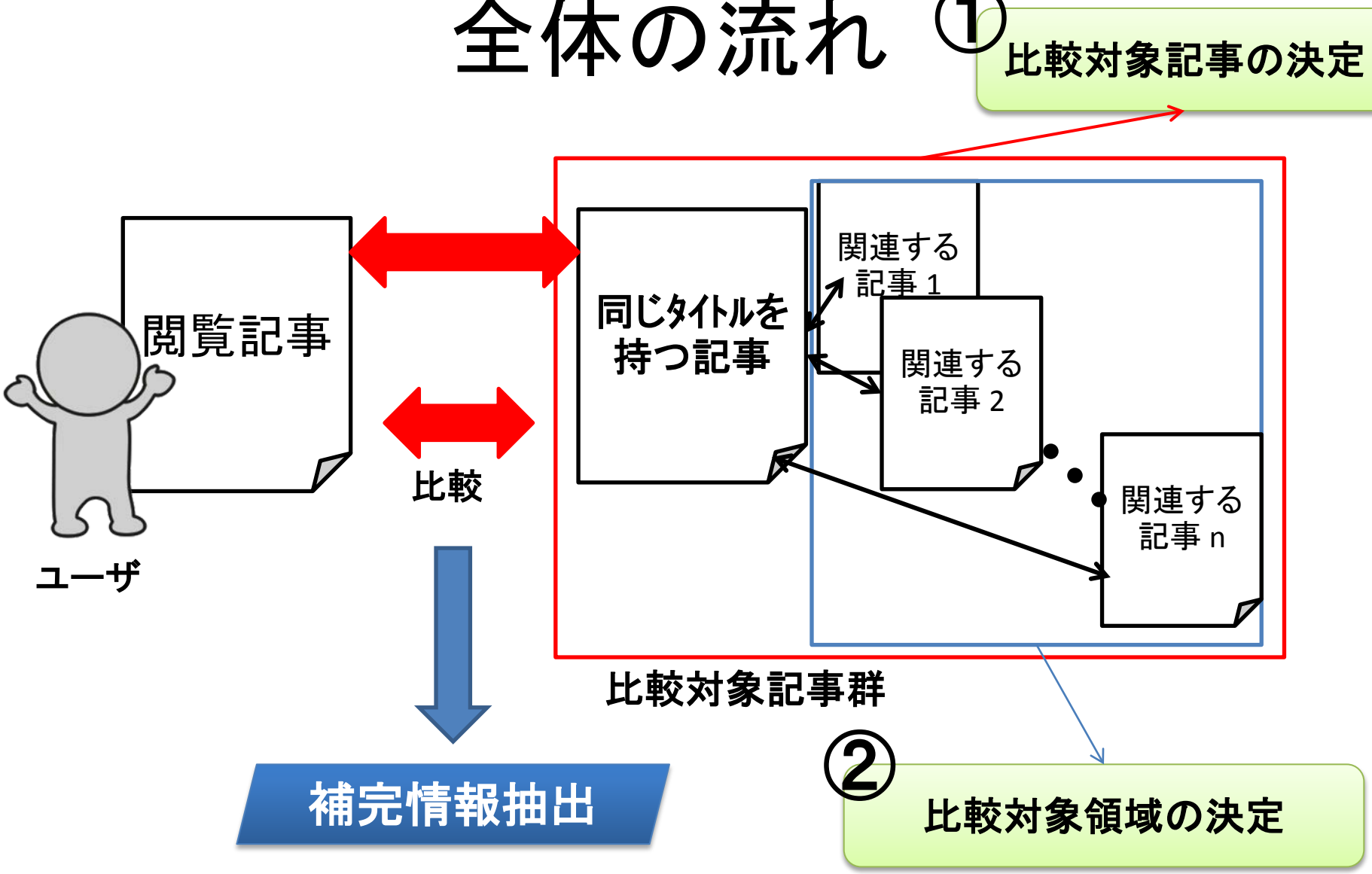
1 概要
2 材料
2.1 生地
2.2 具
2.3 味付け
3 食べ方
4 歴史
5 大阪のたこ焼き
6 器具
6.1 業務用
6.2 家庭用
7 日本国外におけるたこ焼き
7.1 朝鮮文化圏
7.2 中国文化圏
7.3 東南アジア
8 関連事項

補完情報を抽出

補完情報を抽出

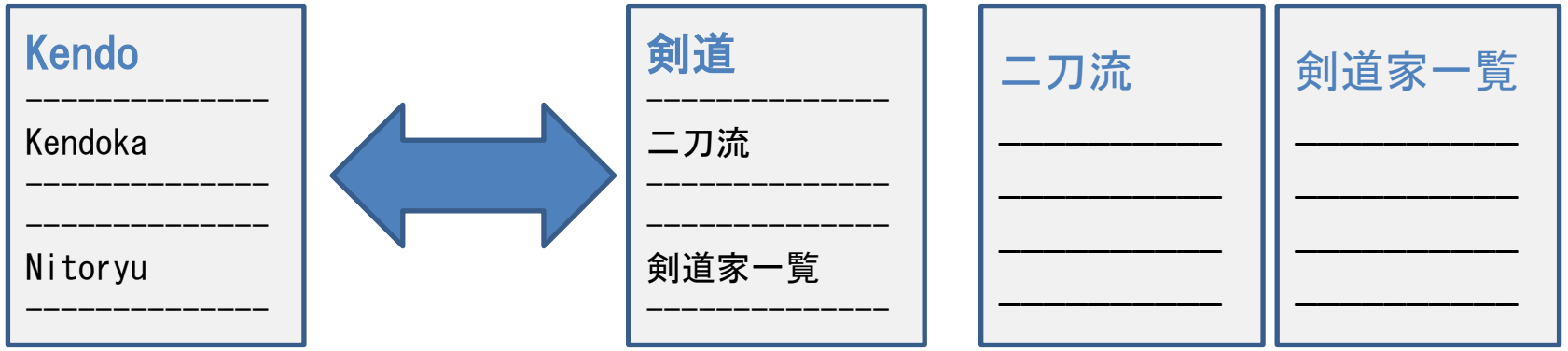
補完

全体の流れ


①
比較対象記事の決定
閲覧記事
ユーザ
比較
同じタイトルを
持つ記事
関連する
記事 1
関連する
記事 2
関連する
記事 n
比較対象記事群
補完情報抽出
②
比較対象領域の決定

①比較対象Wikipediaの記事の決定

- 言語や文化の違いから情報の粒度が異なる
  →対応する記事が複数にまたがる場合がある
  - Ex:「剣道」
    - 英語版:二刀流や剣道家の説明が含まれている
    - 日本語版:二刀流，剣道家一覧の記事が各々存在する



複数ページと比較して補完情報を抽出する必要がある

①比較対象Wikipediaの記事の決定

リンク構造解析

関連している記事同士は
リンク関係にある

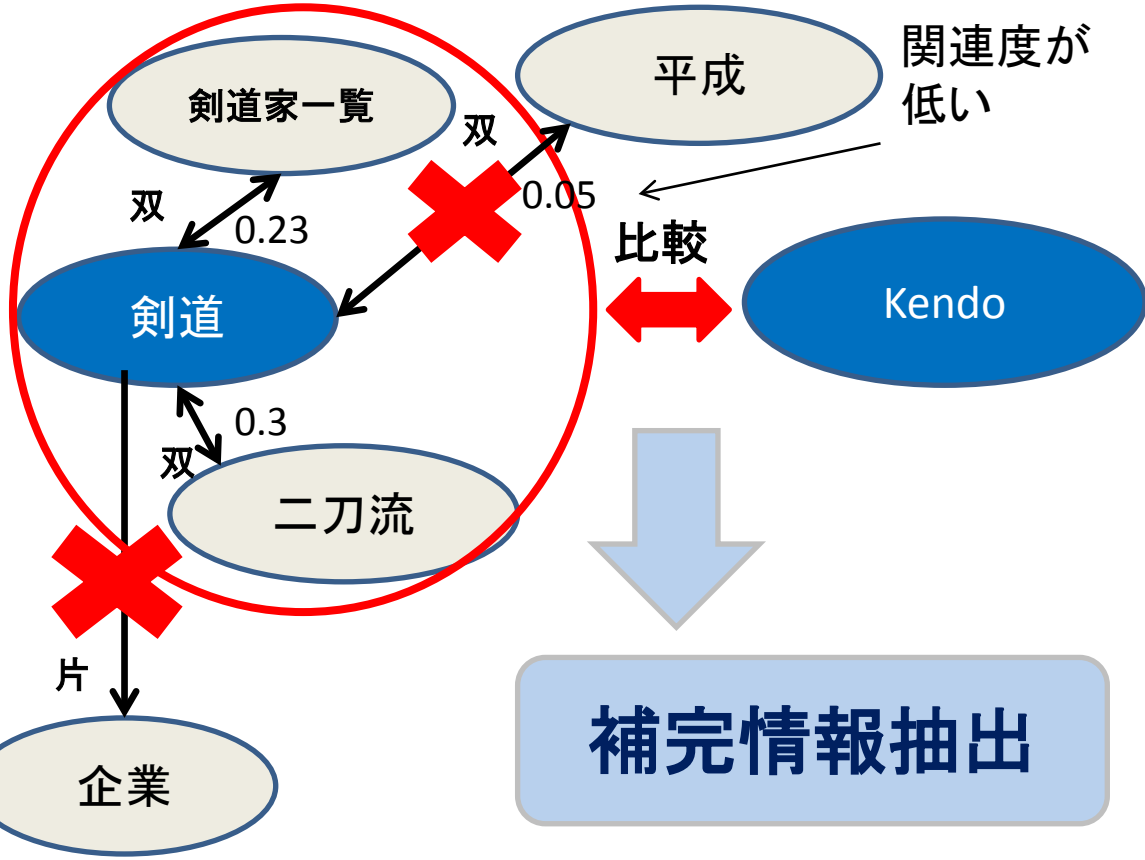
リンクグラフの生成

強連結：関連性が強い

関連度計算

閾値以下を取り除く

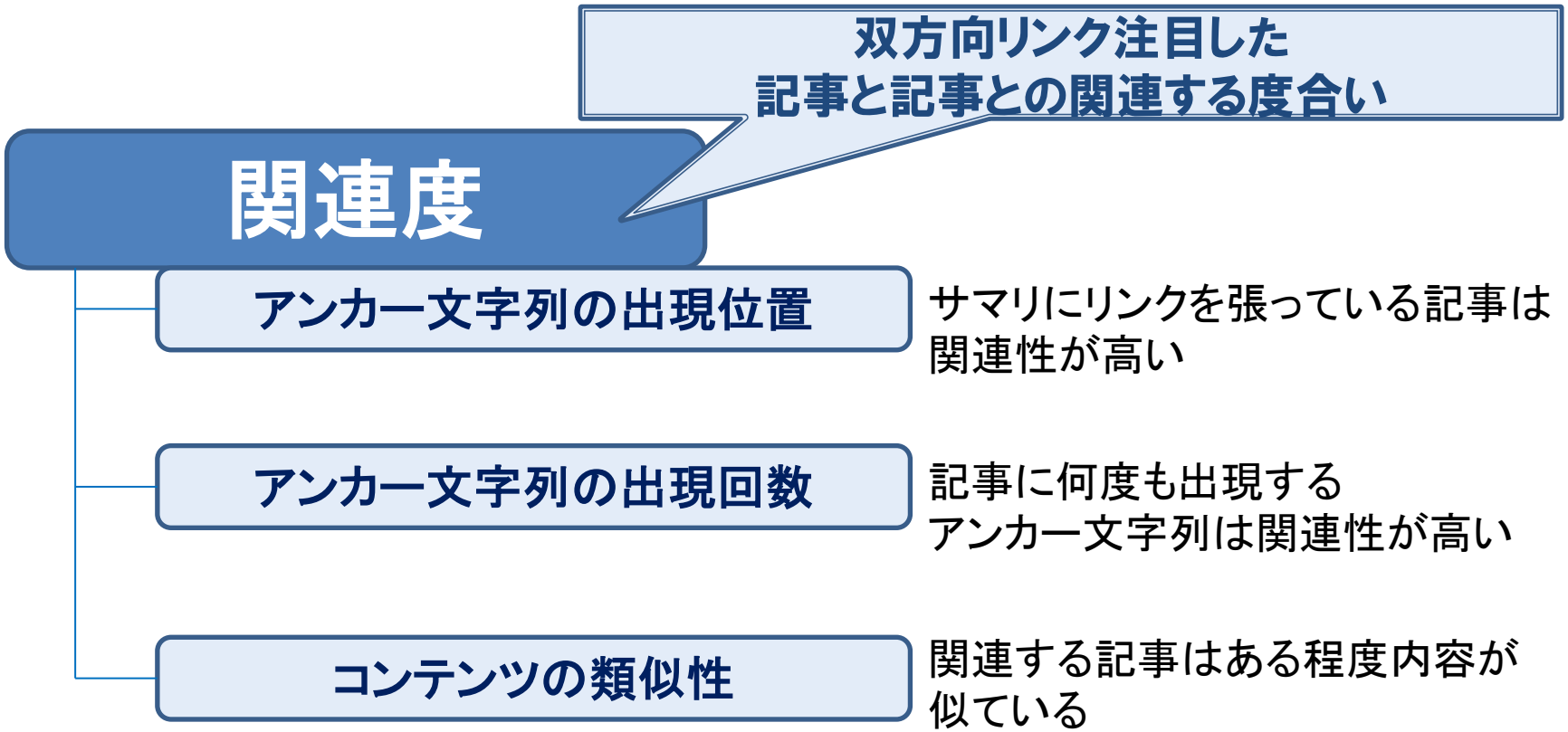
比較対象ページの取得



補完情報抽出

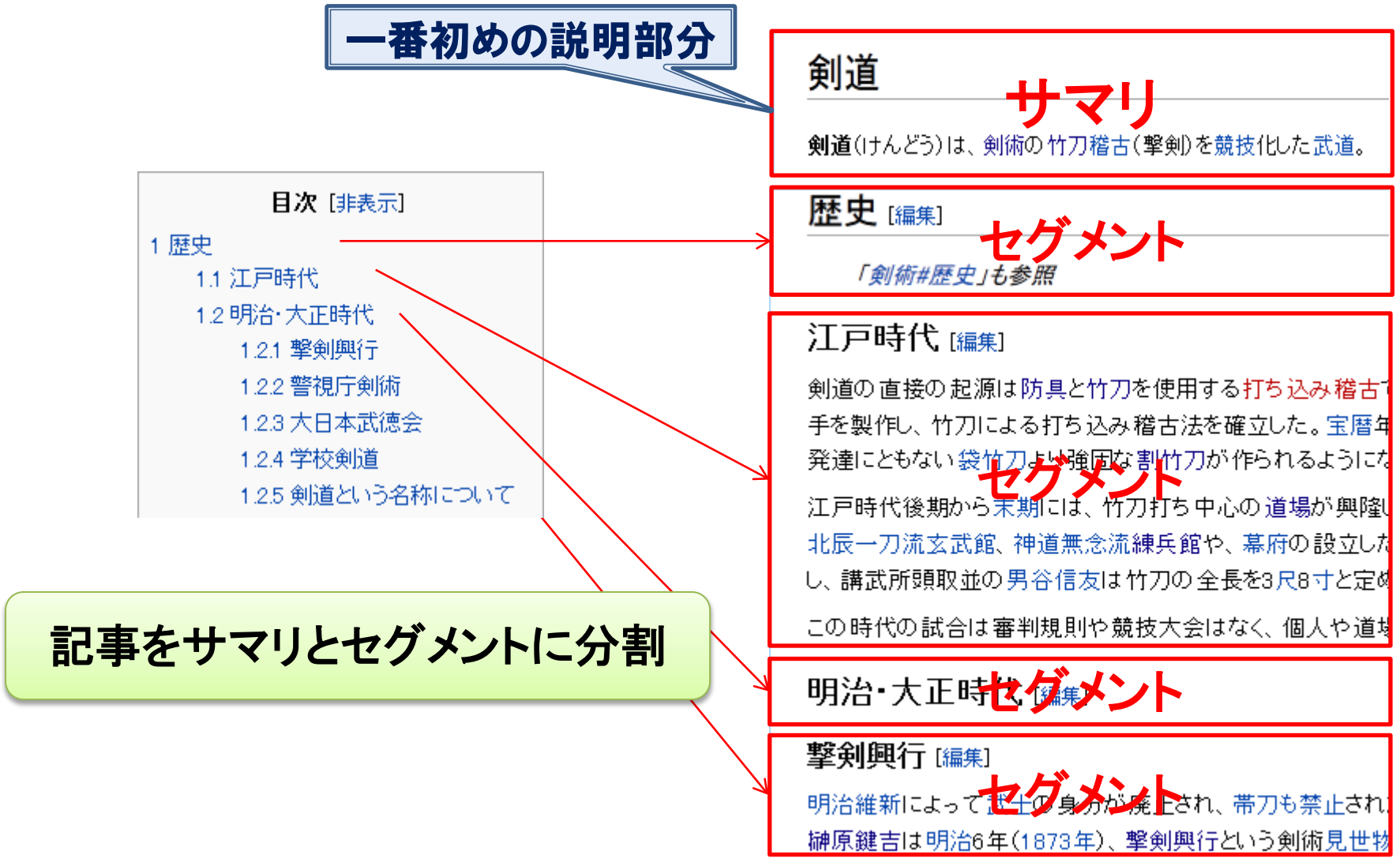
①比較対象Wikipediaの記事の決定

関連度



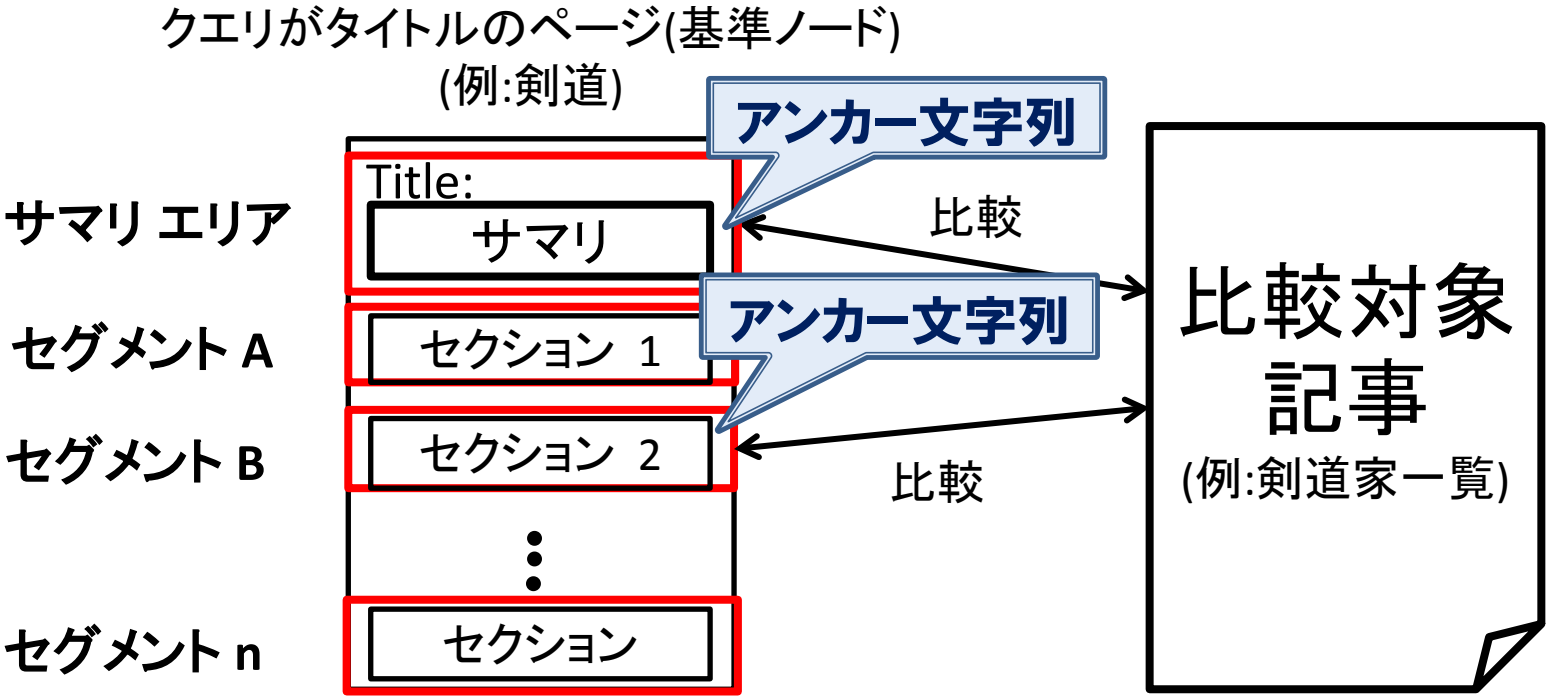
①比較対象Wikipediaの記事の決定

関連度



①比較対象Wikipediaの記事の決定

関連度



比較対象記事のアンカー文字列が基準ノードのサマリ，どのセグメントに出現するかを求める

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

i:比較対象ノード
TF_{sumi}のサマリのアンカー文字列の出現回数
S_{sumi}のサマリと比較対象記事との類似度
TF_{ik}のあるセグメントのアンカー文字列の出現回数
S_{ik}のあるセグメントと比較対象記事との類似度
n:あるIのリンクを張っているセグメントの数
max(R_{im}):比較対象記事群のR_iの最大値

②比較対象領域の決定

- 以前，我々は関連度を用いて抽出した比較対象記事に対し補完情報の抽出を行った
  - → 閲覧記事に関係ない情報が抽出される場合が存在した
  - – 例：剣道と二刀流の場合
    - 剣道の情報は一部分のみ
    - 西洋剣術の二刀流など

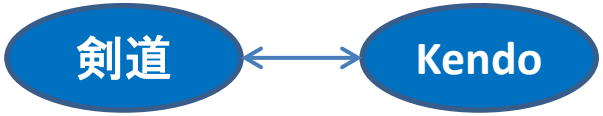
比較対象領域を決定する必要がある

②比較対象領域の決定

得られた比較対象記事に対し記事の分類を行う

- 比較基準記事

  閲覧記事と同じタイトルを持つ記事
  例:「Kendo」に対する「剣道」

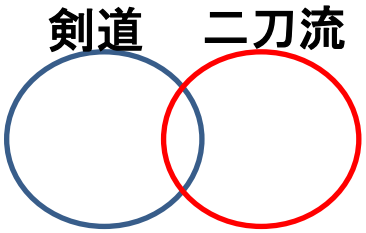


- 包含関係記事

  比較基準記事と包含関係にある記事
  例:「剣道」と「剣道家一覧」の関係

- 部分一致記事

  記事の一部分が比較基準記事と関係する記事
  例:「剣道」と「二刀流」の関係

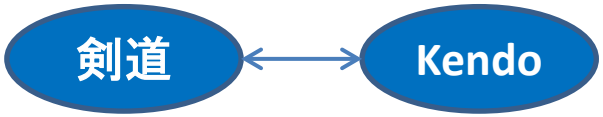


②比較対象領域の決定

得られた比較対象記事に対し記事の分類を行う

- 比較基準記事

  閲覧記事と同じタイトルを持つ記事
  例:「Kendo」に対する「剣道」

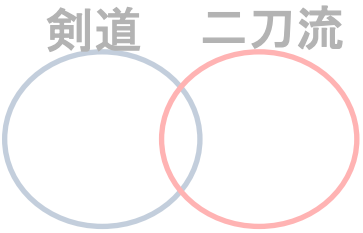


- 包含関係記事

  比較基準記事と包含関係にある記事
  例:「剣道」と「剣道家一覧の関係

- 部分一致記事

  記事の一部分が比較基準記事と関係する記事
  例:「剣道」と「二刀流」の関係



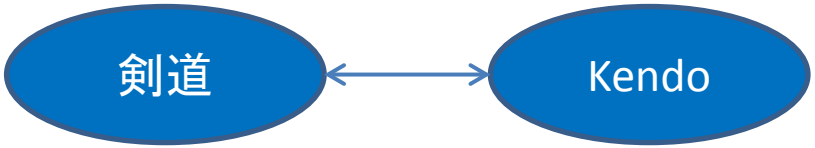
比較基準記事

- 閲覧記事と同じタイトルを持つ記事
  - 例:「Kendo」に対する「剣道」

抽出方法

閲覧記事に存在する言語間リンクを用いて抽出

英語版:Kendo



②比較対象領域の決定

得られた比較対象記事に対し記事の分類を行う

- 比較基準記事

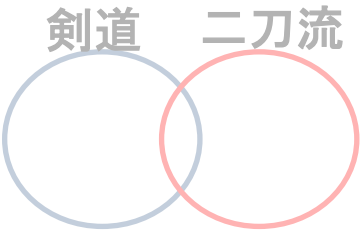
  閲覧記事と同じタイトルを持つ記事
  例:「Kendo」に対する「剣道」

- 包含関係記事

  比較基準記事と包含関係にある記事
  例:「剣道」と「剣道家一覧」の関係

- 部分一致記事

  記事の一部分が比較基準記事と関係する記事
  例:「剣道」と「二刀流」の関係



包含関係記事

- 比較基準記事と包含関係になっている記事
  ->比較基準記事とis-a関係になっている記事
    - 例:「剣道」と「剣道家一覧」の関係

抽出方法

中山らの提案するLSP法を用いて抽出

記事の冒頭文を重要文とし，その文に比較基準記事のアンカー文字列が存在する場合に包含関係記事として抽出

例:**剣道家一覧**

比較基準記事への アンカー文字列

剣道家一覧は、剣道で活躍した人の一覧

参考文献：Nakayama,K.： “Wikipedia Mining for Triple Extraction Enhanced by Co-reference Resolution “ Proceedings of the 1St International Workshop on Social Data on the Web (SDoW’08)

②比較対象領域の決定

得られた比較対象記事に対し記事の分類を行う

- 比較基準記事

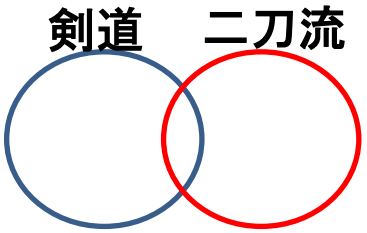
  閲覧記事と同じタイトルを持つ記事
  例:「Kendo」に対する「剣道」

- 包含関係記事

  比較基準記事と包含関係にある記事
  例:「剣道」と「剣道家一覧」の関係

- 部分一致記事

  記事の一部分が比較基準記事と関係する記事
  例:「剣道」と「二刀流」の関係

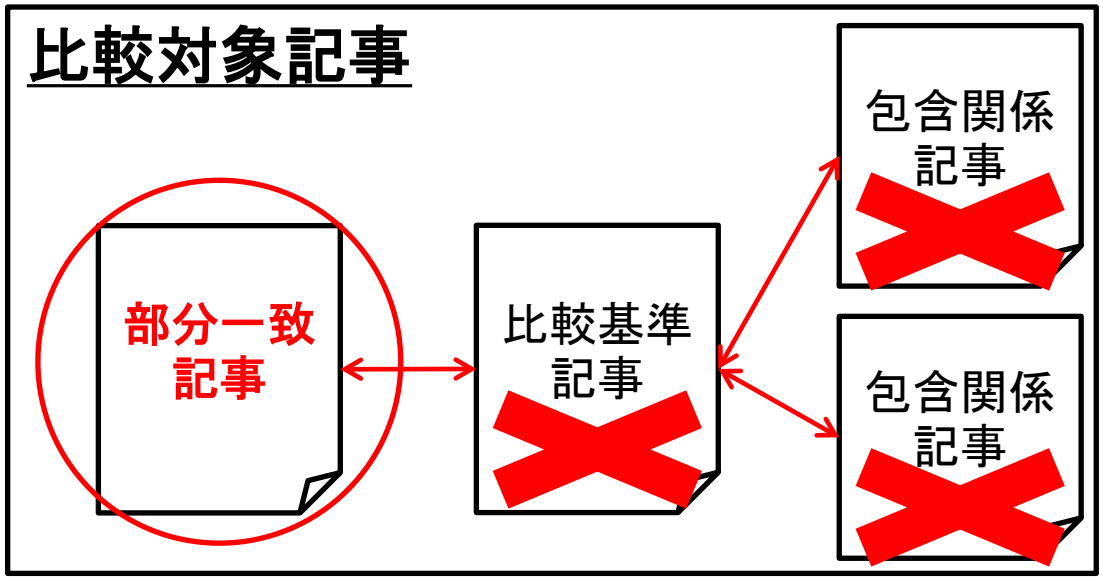


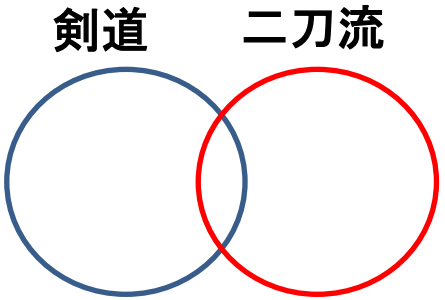
部分一致記事

- 記事の一部分が比較基準記事と関係する記事
  - 例:「剣道」と「二刀流」との関係

抽出方法

比較対象記事から比較基準記事と包含関係記事を除いたすべての記事





②比較対象領域の決定

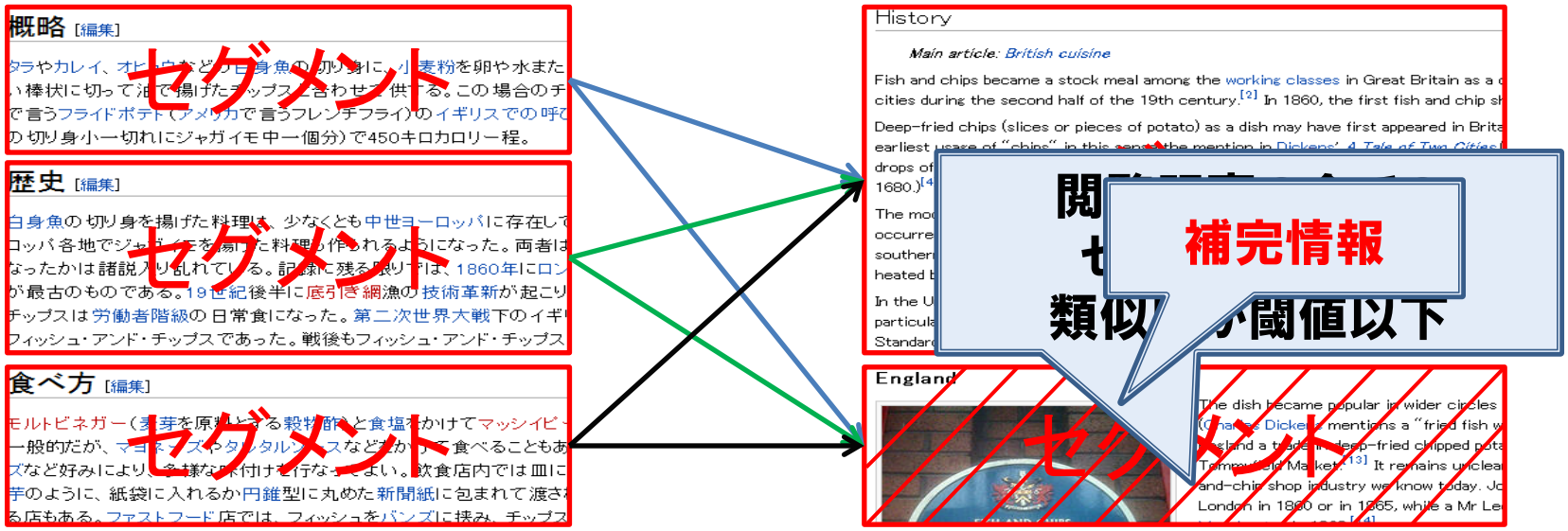
- 分類
  - 比較基準記事
  - 包含関係記事
  - 部分一致記事
- 比較領域の決定
  - 比較基準記事・包含関係記事
    - 記事全体を比較
  - 部分一致記事
    - セクションのタイトルに比較基準記事のアンカー文字列を含む場合
    - サブセクションのタイトルに比較基準記事のアンカー文字列を含む場合
    - 記事本文中に比較基準記事のアンカー文字列を含む場合

比較対象領域の決定と補完情報抽出

- 分類ごとに補完情報を抽出するために，閲覧記事と比較対象となる領域を決定する

比較基準記事・包含関係記事

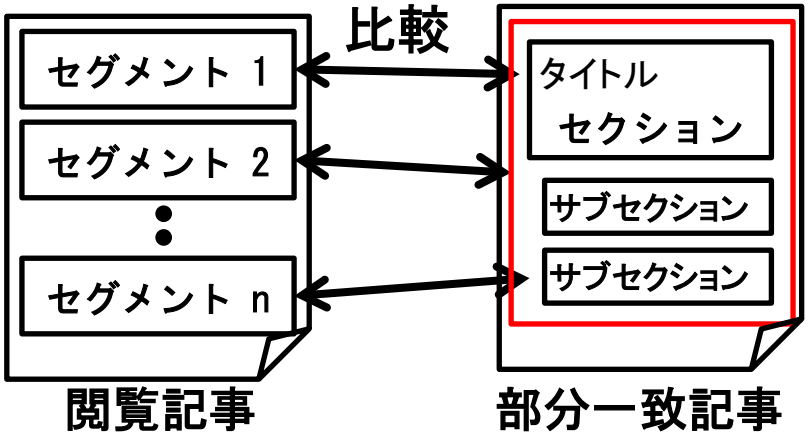
閲覧記事との関係が強いと考え記事全体を比較対象とする



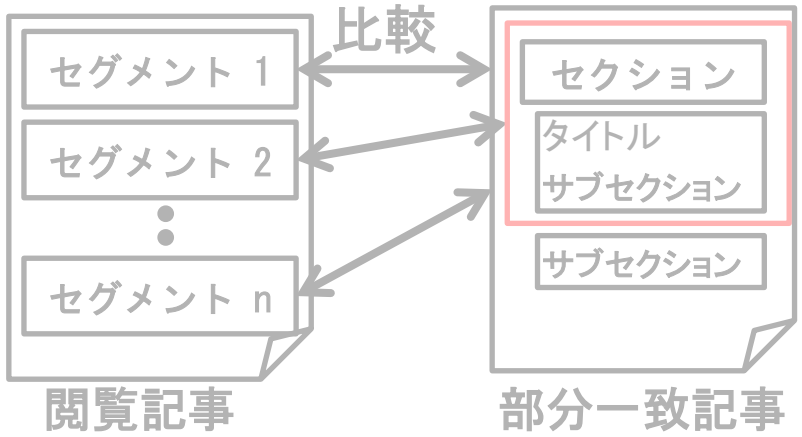
比較対象領域の決定と補完情報抽出

部分一致記事

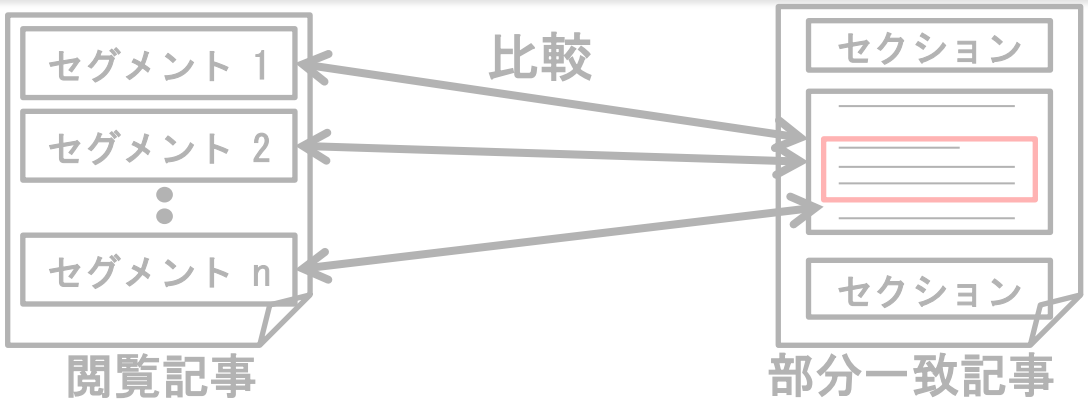
セクションのタイトルに比較基準記事のアンカー文字列を含む場合



サブセクションのタイトルに比較基準記事のアンカー文字列を含む場合



記事本文中に比較基準記事のアンカー文字列を含む場合

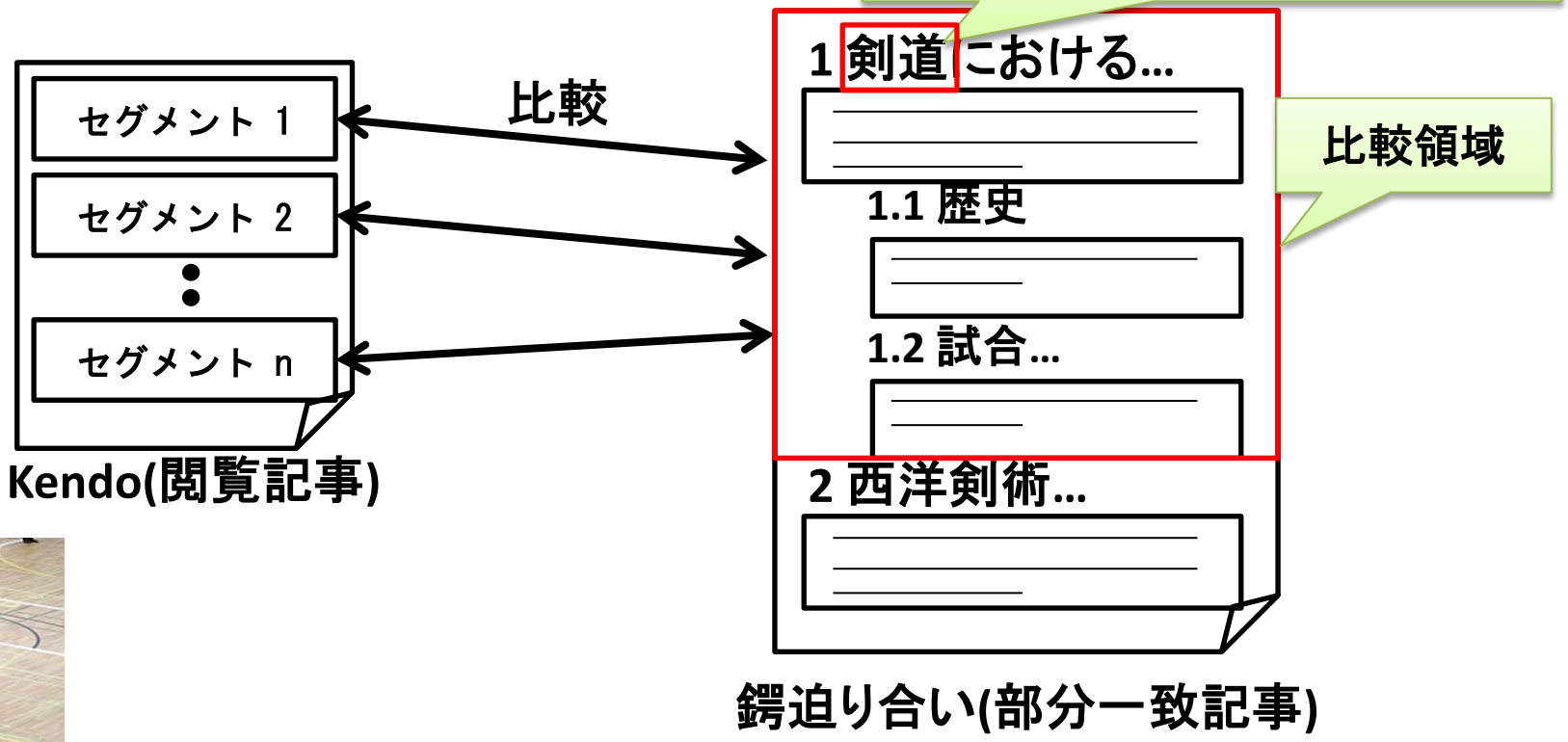


比較対象領域の決定と補完情報抽出

セクションのタイトルに比較基準記事のアンカー文字列を含む場合

サブセクションを含めそのセクション全体を比較対象とする

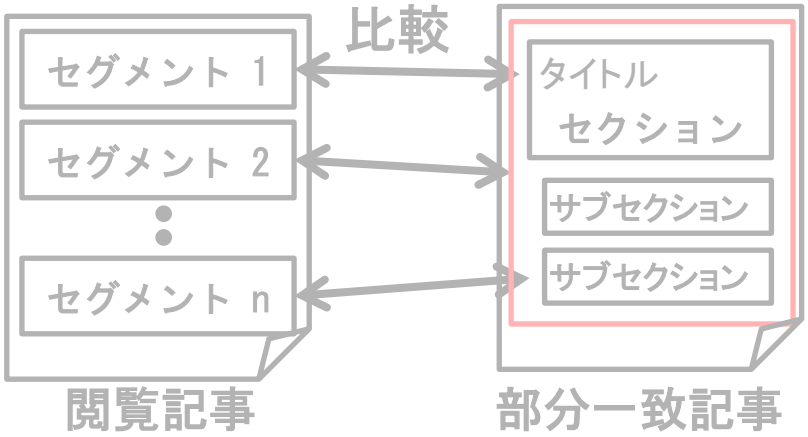
例：剣道(比較基準記事)と鍔迫り合い(部分一致記事)



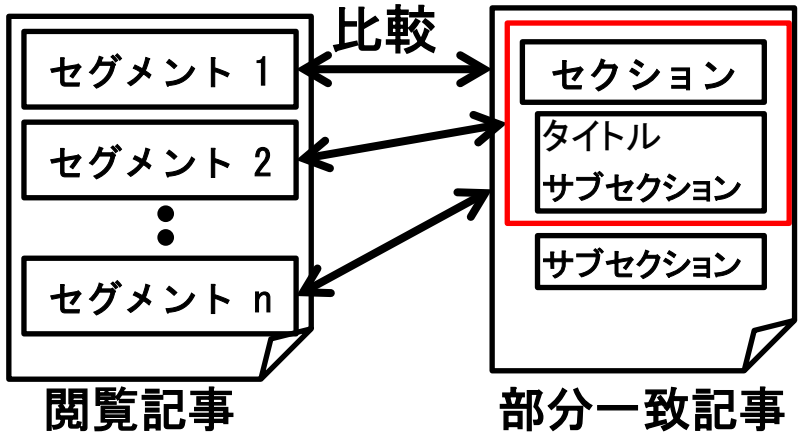
比較対象領域の決定と補完情報抽出

部分一致記事

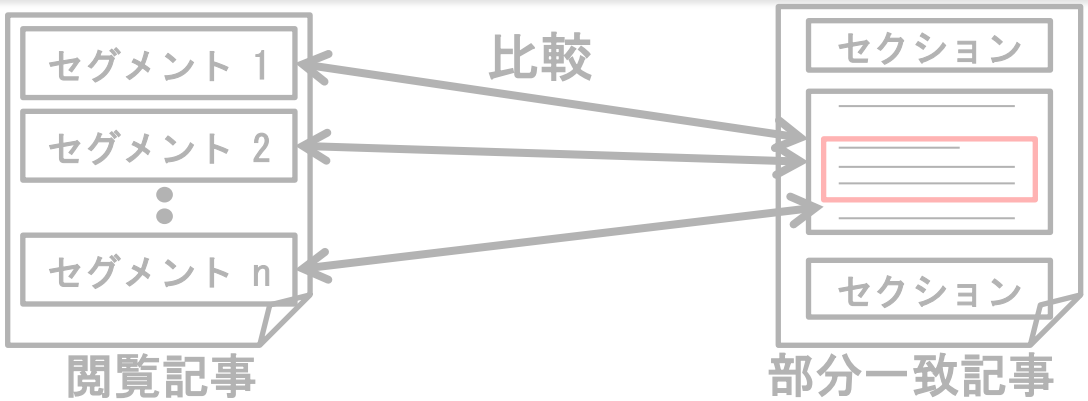
セクションのタイトルに比較基準記事のアンカー文字列を含む場合



サブセクションのタイトルに比較基準記事のアンカー文字列を含む場合



記事本文中に比較基準記事のアンカー文字列を含む場合

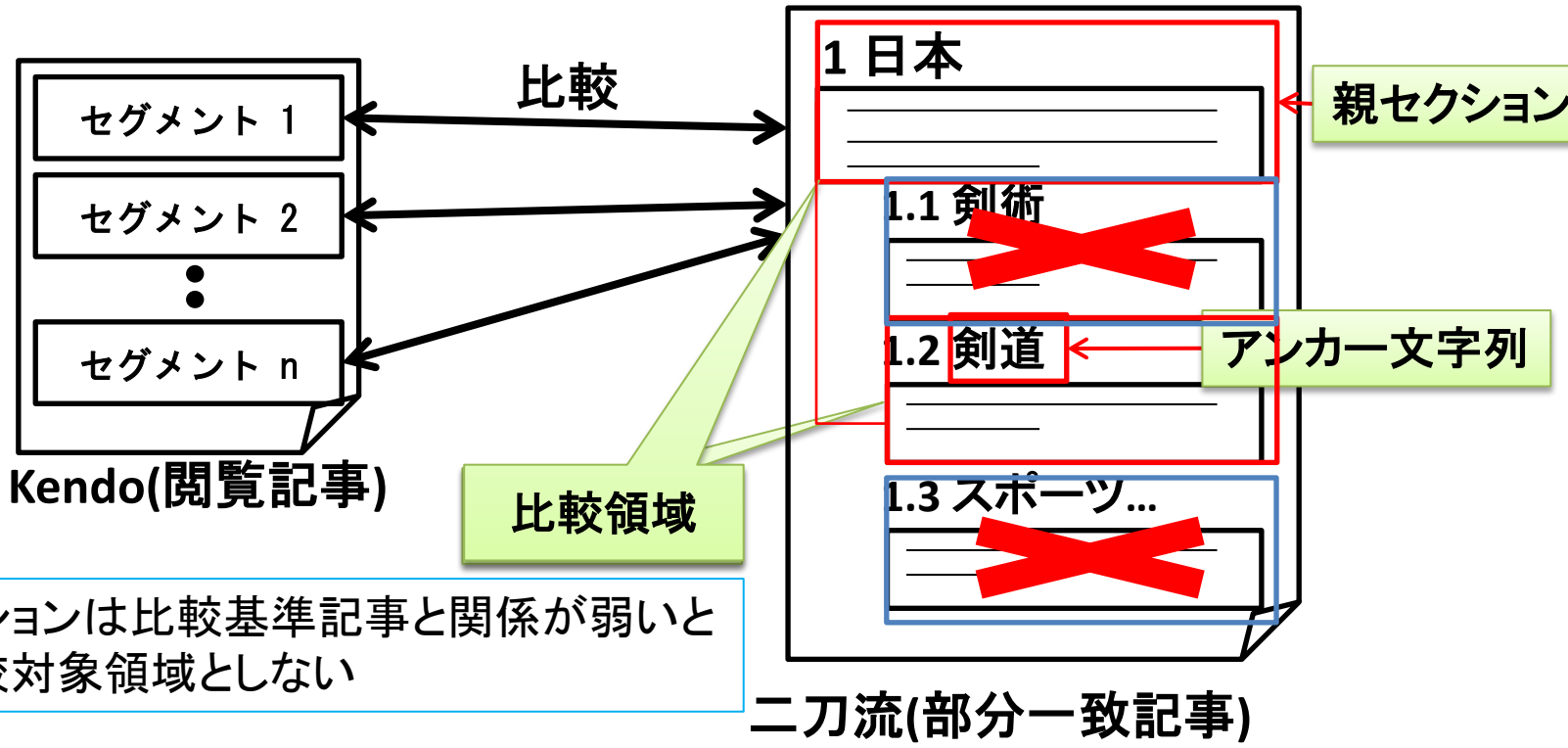


比較対象領域の決定と補完情報抽出

サブセクションのタイトルに比較基準記事のアンカー文字列を含む場合

親セクションとそのサブセクションを比較対象とする

例：剣道(比較基準記事)と二刀流(部分一致記事)

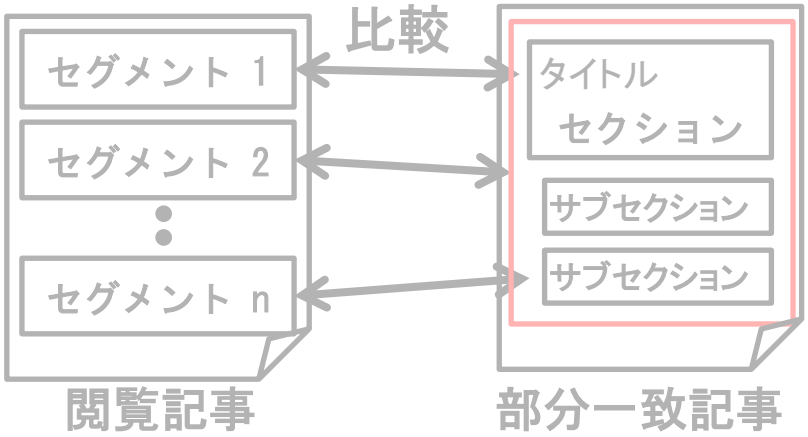


兄弟セクションは比較基準記事と関係が弱いと考え，比較対象領域としない

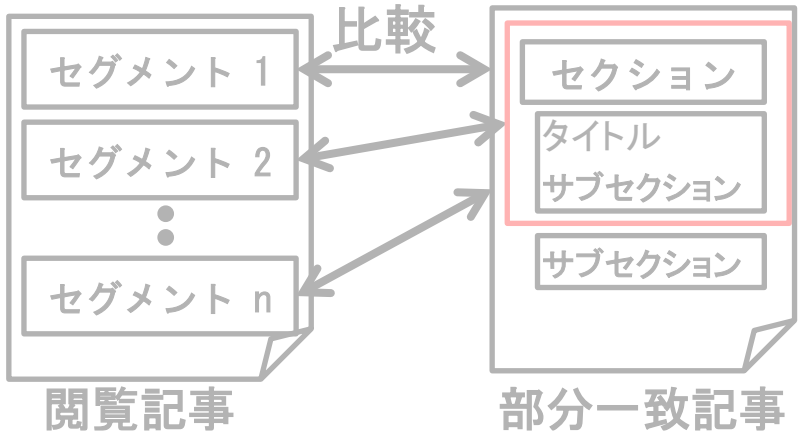
比較対象領域の決定と補完情報抽出

部分一致記事

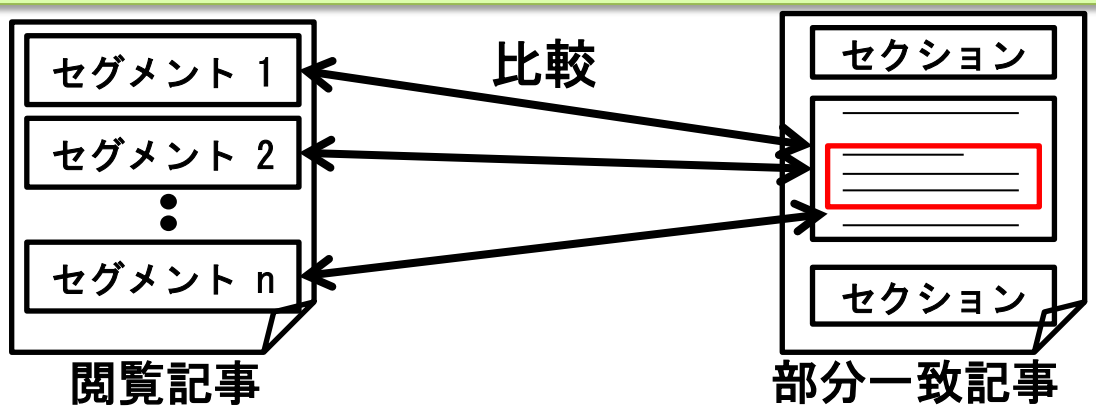
セクションのタイトルに比較基準記事のアンカー文字列を含む場合



サブセクションのタイトルに比較基準記事のアンカー文字列を含む場合



記事本文中に比較基準記事のアンカー文字列を含む場合

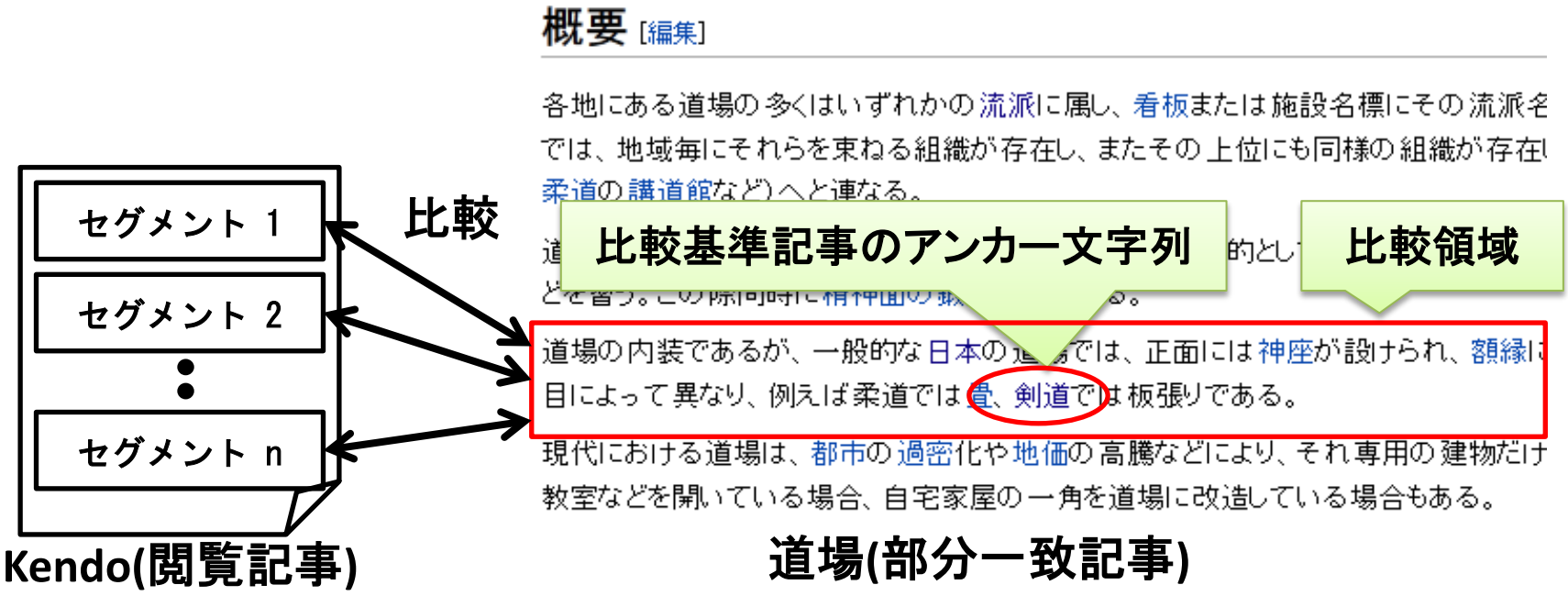


比較対象領域の決定と補完情報抽出

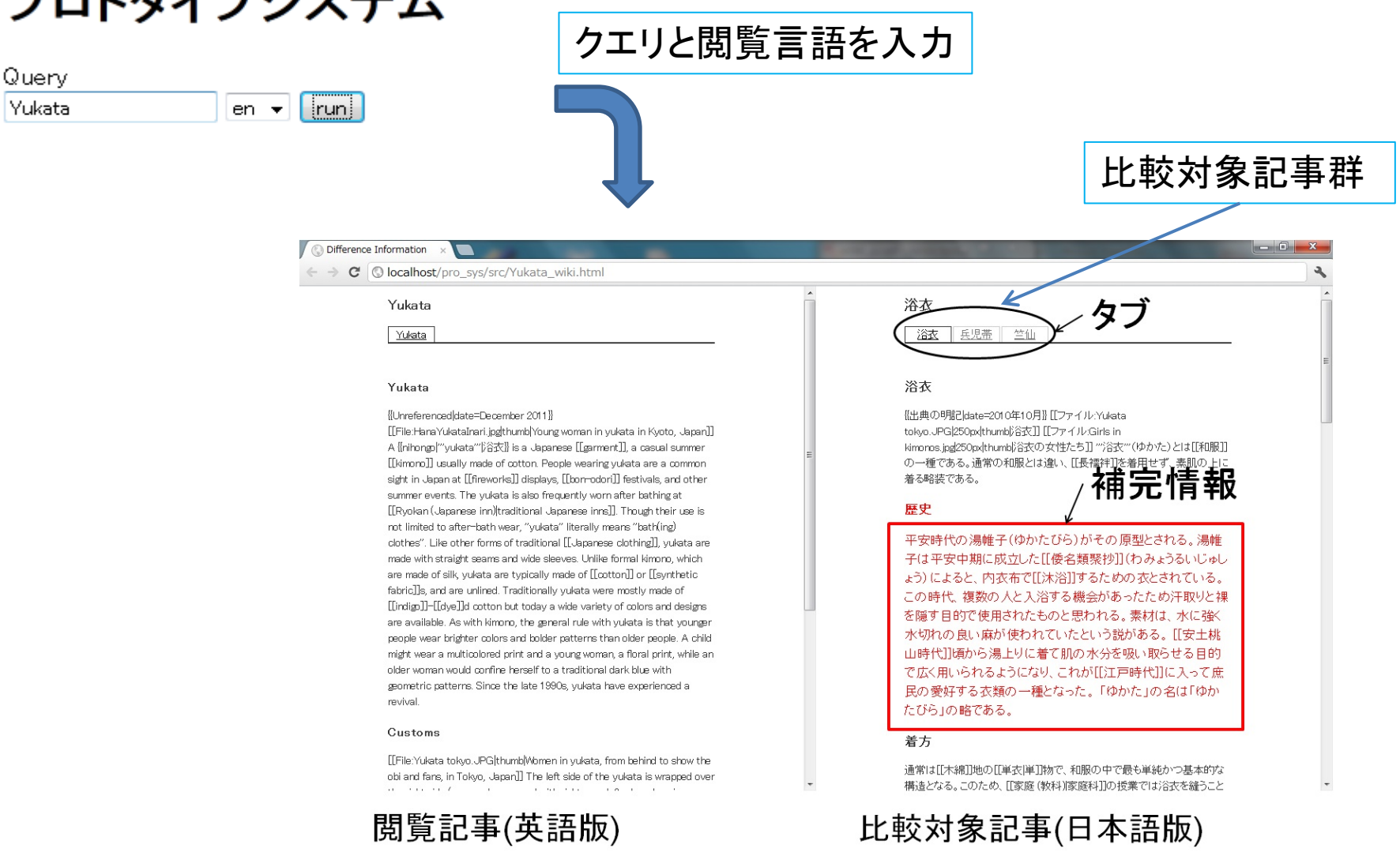
記事本文中に比較基準記事のアンカー文字列を含む場合

アンカー文字列の含まれている段落のみを比較対象とする

例：剣道(比較基準記事)と道場(部分一致記事)



プロトタイプシステム


プロトタイプシステム
Query
Yukata
en
run
クエリと閲覧言語を入力
比較対象記事群
Difference Information
localhost/pro_sys/src/Yukata_wiki.html
Yukata
Yukata
Yukata
{{Unreferenced|date=December 2011}}
[[File:HanaYukataInari.jpg|thumb|Young woman in yukata in Kyoto, Japan]]
A [[nihongo|"yukata"|浴衣]] is a Japanese [[garment]], a casual summer [[kimono]] usually made of cotton. People wearing yukata are a common sight in Japan at [[fireworks]] displays, [[bon-odori]] festivals, and other summer events. The yukata is also frequently worn after bathing at [[Ryokan (Japanese inn)|traditional Japanese inns]]. Though their use is not limited to after-bath wear, "yukata" literally means "bath(ing) clothes". Like other forms of traditional [[Japanese clothing]], yukata are made with straight seams and wide sleeves. Unlike formal kimono, which are made of silk, yukata are typically made of [[cotton]] or [[synthetic fabric]]s, and are unlined. Traditionally yukata were mostly made of [[indigo]]-[[dye]]d cotton but today a wide variety of colors and designs are available. As with kimono, the general rule with yukata is that younger people wear brighter colors and bolder patterns than older people. A child might wear a multicolored print and a young woman, a floral print, while an older woman would confine herself to a traditional dark blue with geometric patterns. Since the late 1990s, yukata have experienced a revival.
Customs
[[File:Yukata tokyo.JPG|thumb|Women in yukata, from behind to show the obi and fans, in Tokyo, Japan]] The left side of the yukata is wrapped over
浴衣
タブ
浴衣
兵児帯
竿仙
浴衣
{{出典の明記|date=2010年10月}} [[ファイル:Yukata tokyo.JPG|250px|thumb|浴衣]] [[ファイル:Girls in kimonos.jpg|250px|thumb|浴衣の女性たち]] '''浴衣'''（ゆかた）とは[[和服]]の一種である。通常の和服とは違い、[[長襦袢]]を着用せず、素肌の上に着る略装である。
補完情報
歴史
平安時代の湯帷子（ゆかたびら）がその原型とされる。湯帷子は平安中期に成立した[[倭名類聚抄]]（わみょうるいじゅしょう）によると、内衣布で[[沐浴]]するための衣とされている。この時代、複数の人と入浴する機会があったため汗取りと裸を隠す目的で使用されたものと思われる。素材は、水に強く水切れの良い麻が使われていたという説がある。[[安土桃山時代]]頃から湯上りに着て肌の水分を吸い取らせる目的で広く用いられるようになり、これが[[江戸時代]]に入って庶民の愛好する衣類の一種となった。「ゆかた」の名は「ゆかたびら」の略である。
着方
通常は[[木綿]]地の[[単衣|単]]物で、和服の中で最も単純かつ基本的な構造となる。このため、[[家庭（教科）|家庭科]]の授業では浴衣を縫うこと
閲覧記事(英語版)
比較対象記事(日本語版)

実験

- 提案手法の有用性を示す実験を行った
  - 実験内容
    - 提案手法とBaselineの比較
    - Baseline:比較対象領域の決定を行わない場合
    - 適合率，再現率，F値の比較
    - 比較言語版
      - 閲覧言語：英語版
      - 比較対象言語：日本語版
    - 設定(前実験より)
      - 関連度の式のα：3.0
      - 関連度の閾値β:0.2
      - コンテンツの比較の閾値γ：0.2

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

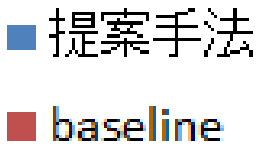
実験条件

Query	実験条件		
	閲覧記事の セクションの数	比較対象 記事の数	比較対象記事の セクションの数
My Neighbor Totoro (となりのトトロ)	9	2	25
Doraemon (ドラえもん)	9	5	38
Iaido (居合道)	10	4	24
Manzai (漫才)	3	4	22
Yukata (浴衣)	1	3	7
Urashima Taro (浦島太郎)	5	5	39
Pikachu (ピカチュウ)	6	2	25
Kinkaku-ji (鹿苑寺)	24	6	28
Hello Kitty (ハローキティ)	15	2	60
Kyudo (弓道)	16	3	47

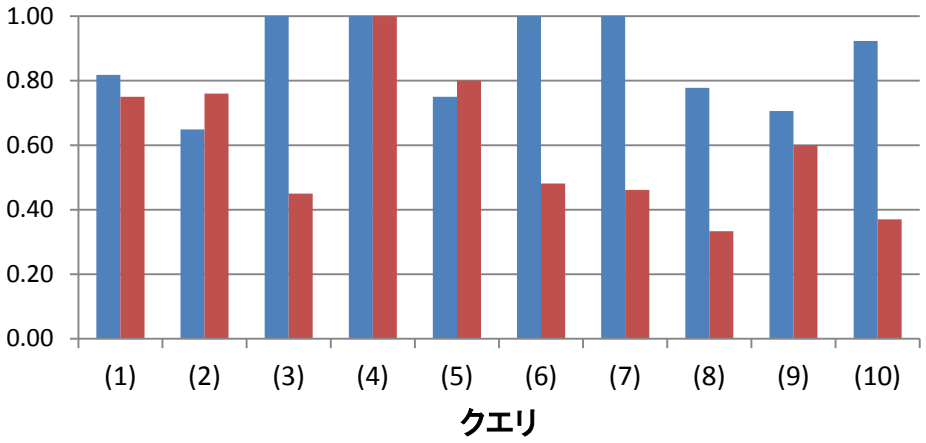
$$\text{適合率} = \frac{\text{抽出した補完情報} \cap \text{正解データ}}{\text{抽出した補完情報}} \qquad \text{再現率} = \frac{\text{抽出した補完情報} \cap \text{正解データ}}{\text{正解データ}}$$

正解データ：閲覧記事に対し補完情報となる比較対象記事のセクションまたは段落

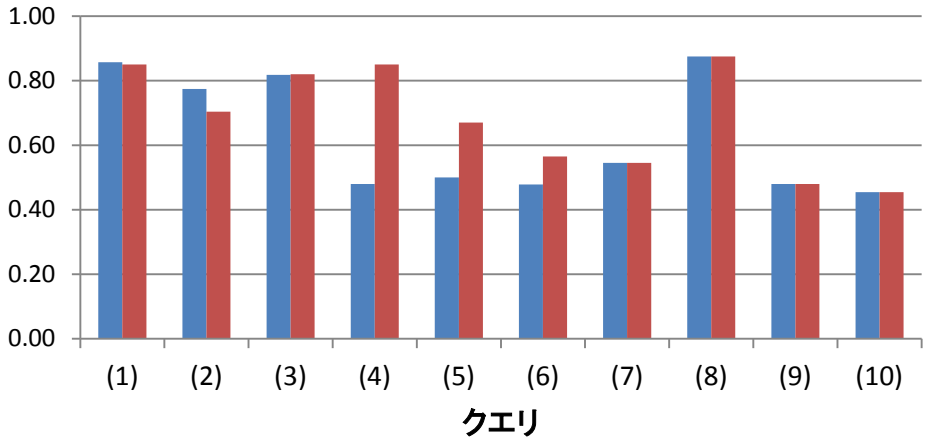
結果



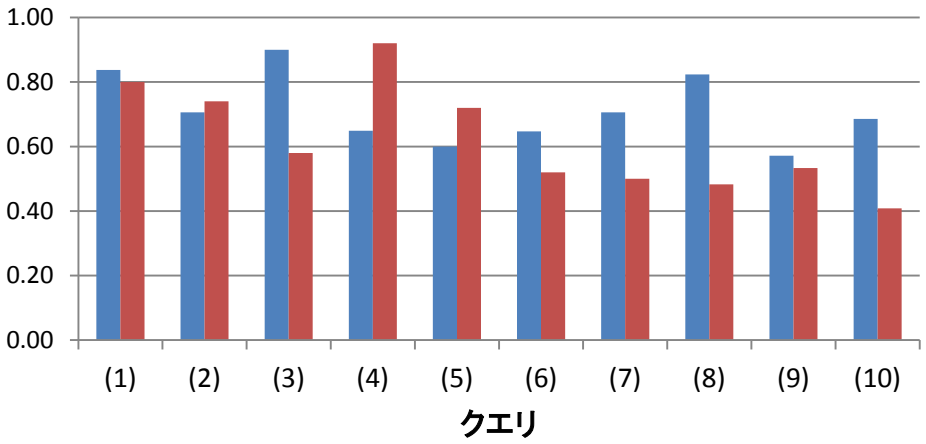
適合率



再現率



F値



Number	クエリ
(1)	My Neighbor Totoro(となりのトトロ)
(2)	Doraemon(ドラえもん)
(3)	Iaido(居合道)
(4)	Manzai(漫才)
(5)	Yukata(浴衣)
(6)	Urashima Taro(浦島太郎)
(7)	Pikachu(ピカチュウ)
(8)	Kinkaku-ji(鹿苑寺)
(9)	Hello_Kitty(ハローキティ)
(10)	Kyudo(弓道)

平均(適合率:0.60->0.86, 再現率:0.68->0.62, F値:0.62->0.71)

結果の良い例

- 居合道
  - Baseline
    - 部分一致記事である武道の称号の範士が抽出，しかし柔道や弓道の範士のように居合道に関係のない情報が抽出された
  - 提案手法
    - 範士，その中でも居合道の情報のみが抽出できた
- となりのトトロ
  - Baseline
    - 部分一致記事である狭山丘陵，地理情報などとなりのトトロと関係のない情報が抽出された
  - 提案手法
    - 狭山丘陵，となりのトトロの舞台となったという情報が抽出できた

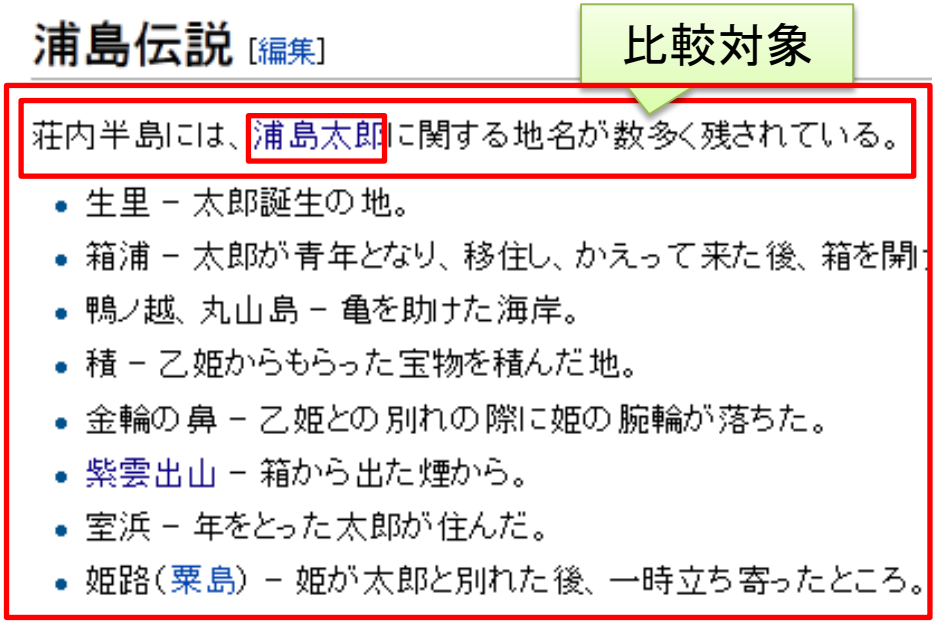
考察

- 部分一致記事において補完情報と成り得ない情報が抽出される場合が存在した
  - 例:ドラえもん
    - 部分一致記事として作者の藤子・F・不二雄が抽出
    - 藤子・F・不二雄の記事では多くのセクションでドラえもんのアンカー文字列が出現
    - ドラえもん以外に多くの漫画を描いており，ドラえもんの補完情報とならない情報が抽出された

考察

- 比較対象となる領域が正しく決定できない場合が存在
  - 例：浦島太郎
    - 部分一致記事として荘内半島が抽出

浦島伝説は浦島太郎の同義語



荘内半島

まとめと今後の課題

- まとめ
  - Wikipediaの多言語性に着目し，内容の充実していない記事に対し他言語のWikipediaを用いて，情報の補完を行う手法を提案した
  - 提案手法
    - 比較対象記事の決定
    - 比較対象領域の決定
      - **比較基準記事，包含関係記事，部分一致記事**
    - コンテンツの比較による補完情報抽出
- 今後の課題
  - 補完情報の分類(詳細な情報，新しい情報 etc…)
  - 補完情報の提示のインタフェース
  - 比較領域の決定の際に同義語への対応