Vol.2011-DBS-153 No.4
2011/11/3

多言語 Wikipedia の差異情報抽出手法の提案

藤原裕也[†1]　鈴木優[†2]　灘本明代[†3]

Wikipedia は，日本語版や英語版などの各言語版によって記述されている情報の量が異なる．そのため，一つの言語版だけでは情報量が不足する記事が多く存在する．特に文化に関する記事ではその傾向が顕著である．そこで本研究では多言語版 Wikipedia を用いて，日本語版における海外に関する記述の不足を，日本語版以外の記事から抽出し提示するシステムを提案する．具体的には，ユーザの入力した海外の事柄に関する英語版の記事と日本語版の記事を比較し，その差分情報を提示する．この時，言語間の情報の粒度が異なるため，一つの日本語版の記事が一つの英語版の記事に対応するとは限らない．そこで，比較対象となる複数の英語版の記事を，リンクグラフを用いて抽出し，抽出された複数の英語版の記事と日本語版の記事を比較対象とし，その差分情報を抽出する手法を提案する．

Extracting Difference Information from Multilingual Wikipedia

Yuya Fujiwara,[†1] Yu Suzuki[†2] and Akiyo Nadamoto[†3]

There are multilingual articles on the Wikipedia. The information between multilingual articles is different. Especially, the case of the articles which is written about culture is very different between languages. In this paper, we propose the system which extracts different information between Japan and other countries on the Wikipedia. Specifically, the system compare Japanese Wikipedia article which is written about foreign things with English Wikipedia. Then it extract different information which is written in English version. In this time, the granularity of information between two language is different, it means a Japanese article is not suitable for an English article. Then we propose how to extracts multiple English articles which is written about same a Japanese article based on link graph.

1. はじめに

Wikipedia[⋆1]は気軽に様々な事を調べることができるため，現在多くの人々が利用している．Wikipedia が通常の百科事典と大きく異なる点は，エンドユーザがコンテンツを作成している点，現在では 250 を超える言語版が存在している点などが挙げられる[⋆2]．各言語版はそれぞれ独立して管理されており，各々の言語版を閲覧しているユーザが記事の作成，追記，修正を行っている．そのため，ある事柄についての記事が特定の言語版のみに存在していたり，他の言語版と比較して多く作成されている言語版が存在していたりするなど，言語間で各々異なる成長を遂げている．例えば，日本人ユーザがイギリスの代表的な食べ物である「フィッシュ・アンド・チップス」を日本語版で調べてみると，少量の情報だけしか掲載されておらず，しかも目次の項目数も少ない．それに対し英語版では自国の言語圏内であるため目次の項目数が多く，また内容も詳細に記述されている (**図 1** 左参照)．また，この「フィッシュ・アンド・チップス」の日英両方の目次に出現する「歴史」という項目は，英語版にはイングランドでの歴史やスコットランドでのフィッシュ・アンド・チップスに関する歴史について記述されているが，日本語版にはそれらの情報は存在しない．一方，逆に日本の伝統的な建物である「平等院」を日本語版と英語版で比較すると，記述されている目次の項目やその数，記事の内容等が異なることがわかる．例えば日本語版の目次項目の一つである「鳳凰堂」や「文化財」といったものは日本語版だけに存在し，逆に英語版の目次項目の一つである「Gallery」は英語版だけに存在する (図 1 右参照)．

さらに，国や民族，文化によって人々の興味や関心が異なるため，Wikipedia に書かれている記事の種類がジャンルによって異なっている．日本語版の全閲覧数のうち 8 割が大衆文化についての記事に集中しているのに対し，英語版は 4 割である．逆に英語版は日本語版に比べると政治や地理などの記事の人気が高い[⋆3]．そのため，閲覧数に応じて記述の詳細度も

†1 甲南大学大学院 自然科学研究科
Konan University
†2 名古屋大学 情報基盤センター
Nagoya University
†3 甲南大学 知能情報学部
Konana University

⋆1 Wikipedia http://www.wikipedia.org/
⋆2 Wikipedia:全言語版の統計 http://ja.wikipedia.org/wiki/Wikipedia:全言語版の統計
⋆3 asahi.com 悩むウィキペディア　少ない管理人　芸能系ばかり人気
http://www.asahi.com/national/update/0303/TKY201003030157.html

© 2011 Information Processing Society of Japan

言語版によって異なる．自国の言語版だけでは分からない情報が他の言語版では存在する場合があることが分かる．

複数の言語を理解することが困難なユーザが他国の文化や習慣を調べる時，言語の問題から，まず始めに母国語のサイトを調べる場合が多い．たとえ母国語以外の言語を読むことが可能であっても，母国語の記事を読んで理解するのと比較して，他国の言語の記事を読んで理解するのは時間がかかり労力を要する場合が多い．例えば，英語をある程度読むことができるが普段英語を使用していない日本人ユーザにとって，英語版「Fish and chips」の記事のように，情報量が比較的多い記事を全て読んで理解することは時間がかかるため困難な作業である．ここで我々は，日本語版で「フィッシュ・アンド・チップス」の記事を読み，ある程度情報を把握した上で，その日本語版の記事に不足している英語版の情報を挿入することによって，より必要な情報が多い記事となると考えた．つまり，閲覧しているユーザの母国語で記述されている Wikipedia の記事の中で不足している情報が他言語版にある場合，その情報を母国語版記事に挿入して提示すると，ユーザは翻訳されていない部分だけを他言語で閲覧することとなるため，母国語で読むことができる部分が多くなり，可読性が増加する．また，他言語だけで記述されている部分を読むことができるため，必要な情報が得られる割合も増加する．そこで本研究では， 上記のようなユーザを対象とした多言語版 Wikipedia 間の差分情報を抽出し，提示するシステムを提案する．

本論文では，はじめの一歩として，想定しているユーザを英語がある程度理解できる日本人であるとし，ユーザの入力した外国の事柄に関する日本語版と英語版 の記事をそれぞれ抽出し，記事の内容を比較し，その差分情報を提示する．この時，言語間の情報の粒度が異なるため，日本語版において 1 ページの記事が英語版では複数ページにまたがっている場合がある．そこで我々は，比較対象となる複数の英語版の記事を記事間のリンクグラフと我々の提案する関連度を用いて抽出する手法を提案する．さらに抽出された複数の英語版の記事と日本語版の記事を各々の記事の目次構造に基づいて比較し，その差分情報を提示する手法を提案する．本論文では日本語版 Wikipedia の記事を日本語版記事と呼び，英語版 Wikipedia の記事を英語版記事と呼ぶ．

図 2 に処理の流れを示す．

(1) ユーザは調べたい外国の事柄に関する用語をクエリとして入力する．
(2) 入力された用語に関する一つの日本語版記事と，一つの英語版記事をそれぞれ取得する．
(3) 取得した英語版記事と日本語版記事を，目次構造に基づき分割する．この最小単位をセグメント呼ぶ．
(4) 比較対象となる英語版記事群を (2) で取得した英語版記事のリンク構造を解析し，取得する．
(5) 英語版記事群に対してセグメントごとに形態素解析を行い，名詞を取得する．
(6) (5) で取得した名詞を翻訳する．
(7) 日本語版記事に対してセグメントごとに形態素解析を行い，名詞を取得する．
(8) (4) で取得した英語版の記事群の情報と，(2) で取得した日本語の記事の情報の二つの情報に対して，それぞれの目次構造を利用してコンテンツの比較を行い，差分情報を取得する．
(9) (8) で取得した差分情報をユーザに提示する．

図 1 多言語版 Wikipedia の例
Fig. 1 Example of multilingual Wikipedia pages

以下，2 章では関連研究を，3 章では比較対象 Wikipedia の記事の抽出手法について述べ，4 章では多言語記事の差分抽出について述べる． 5 章ではプロトタイプシステムを，6 章では評価実験について述べ，7 章ではまとめと今後の課題について述べる．

© 2011 Information Processing Society of Japan

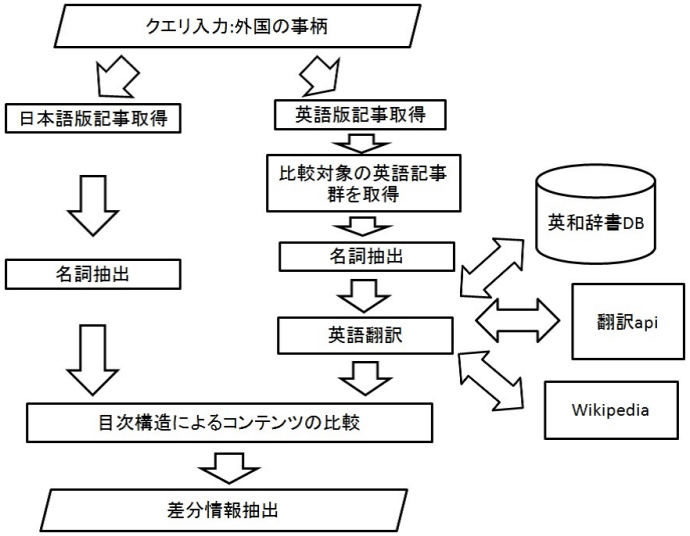

図 2 システムのフロー
Fig. 2 System Flow

2. 関 連 研 究

森ら[1)2)] は特定の話題に属する多言語 Wikipedia の記事の数の差異を抽出し，興味や関心の違いを抽出する手法を提案している．具体的には，Wikipedia のあるカテゴリに属する記事の件数を用いて言語間の差異の抽出を行っている．下位カテゴリがなくなるまでそのカテゴリと上位のカテゴリとの類似度を計算する．得られたカテゴリの数とそのカテゴリにある記事の件数をドイツ語版，フランス語版，日本語版，中国語版の Wikipedia と比較し，それぞれの言語版でどれだけそのカテゴリに興味があるのかを判定している．本研究では Wikipedia の記事の目次構造に基づきコンテンツを比較し，差分情報抽出する手法を提案している点が異なる．斎藤ら[3)] は日本語版，英語版，中国語版のニュース記事から国際的な感覚の違いを抽出ためその記事の収集方法を提案している．具体的には各言語版のニュース記事からそれぞれキーワードを抽出し，英語に翻訳し，索引化を行っている．その翻訳の際に Wikipedia の言語間リンクを用いて翻訳を行っている．本研究でも Wikipedia の言語間リンクを用いて翻訳を行っている．しかしながら，本研究では Wikipedia の記事同士を比較し，その記事の差分情報抽出する手法を提案している点が異なる．中崎ら[4)] は一つのトピックに対して日本語と英語のブログを比較し文化間の違いを抽出する手法を提案している．具体的には，あるトピックの日本語版，英語版記事の記事を取得し，その本文から太字と他の記事へのリンクを関連語として抽出する．そしてあるトピックの日本語，英語のブログサイト群から関連語を用いて順位付けを行い，分析している．本研究ではブログでの多言語比較ではなく Wikipedia の記事の比較になっている点が異なる．言語版により文化の記事は情報の量の違いが顕著であるため文化間の違いが抽出されやすい．立床ら[5)] は，Wikipedia 記事のうち時間情報や空間情報と結び付けられたコンテンツから知識ベースを抽出し，それを用いた検索手法を提案している．具体的には Wikipedia の記事同士の関連度，そして記事に関する時間情報，空間情報の関連度の計算を行い地球科学データに応用している．この際に Wikipedia の記事同士の関連度を計算するために中山ら[6)] の提案した pfibf を用いている．pfibf はある記事から別の記事へのパスの多さ，そしてある記事から別の記事への最短距離，その二つの要素を考慮している．それに対し，本研究ではある記事から別の記事へのリンク数，Wikipedia の目次構造におけるリンクの出現位置を考慮し，関連度の計算を行う手法を提案している．

3. 比較対象の記事の抽出手法

3.1 リンク構造解析

日本語版記事と英語版記事を比較する時，Wikipedia ではそれぞれの言語版で言語や文化の違いから情報の粒度が異なり，対応する記事が複数に存在する場合がある．例えば，イギリスに関する記事の場合，日本語版記事では 1 記事であるのに対し，英語版記事では詳細に書かれており，複数の記事になっていることが多い.「Cricket」の場合，日本語版記事ではクリケットのルールやバッティングの説明がクリケットという記事１つに書いてあるのに対し，英語版記事ではクリケットの記事だけでなくルール，バッティングの記事がそれぞれ存在し，複数ページにまたがっている．そこで，我々は英語版記事のリンク構造を解析する事により，比較対象の記事を抽出する．以下に抽出手順を示す．

(1) ユーザの入力したクエリと同じタイトルを持つ英語版記事から英語版記事をノードとするリンクグラフを作成する．このリンクグラフを基準リンクグラフと呼ぶ．そしてユーザの入力したクエリと同じタイトルを持つ記事のノードを基準ノードと呼ぶ．
(2) 基準ノードと双方向にリンクされているノードは，基準ノードの記事に深く関連すると考え，双方向リンクされており日本語版に存在しないノードを残し，その他のノー

© 2011 Information Processing Society of Japan

Vol.2011-DBS-153 No.4
2011/11/3

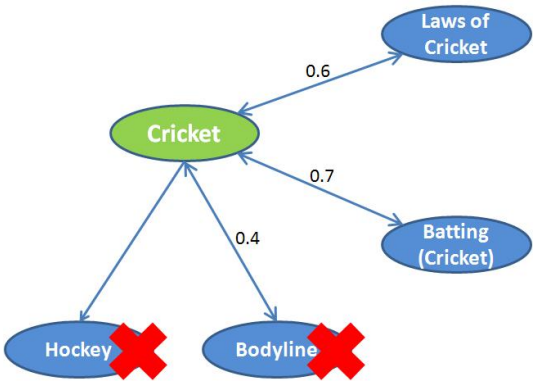

図 3 リンク構造の解析
Fig. 3 Analysis of the link structure

ドをリンクグラフから削除する．(**図 3**)

(3) 基準ノードとその他のノード間の関連度を求める．

(4) その関連度が事前に設定した閾値以下であるものを削除し，残ったノードの記事を比較対象の英語版記事群とする．

3.2 関 連 度

これまで我々の研究[7] では，リンクグラフのノード間をコサイン類似度により記事同士の類似度を測定し，比較対象の英語記事群を取得した．しかし，それだけでは比較対象英語記事を完全に抽出することができない問題が生じた．例えば，上記の 3.1 節でも挙げた「Cricket」という記事とリンクしかつ関連性の高い「Batting」の記事との類似度は低いため,「Batting」の記事を比較対象として抽出できない．そこで本論文ではリンクのアンカー文字列の出現位置に注目し関連度の計算を行うことによって，基準ノードに関連する複数ページ群を抽出する．ここで，アンカー文字列の出現位置はその記事の目次構造によって決定する．我々の提案する関連度とは，Wikipedia のある記事と双方向リンクされている記事がどのぐらい関わりが深いかを測定するための尺度である．

以下，関連度を求める手順を示す．

(1) 基準ノードの記事の目次構造に従って記事内の情報を分割する (**図 4**)．ここで分割された最小単位をセグメントと呼ぶ．

(2) 基準ノードの記事のタイトルおよびサマリーをルートノードとし，セグメントを子ノードとした木構造を作成する (**図 5**)．このときルートノードを深さ 1 とし，目次の章のセグメントを深さ 2 のノード，その下の節のセグメントを深さ 3 のノードとする．深さごとに目次番号の若い子ノードが左側に来るような木構造にする．この木構造をセグメントツリーと呼ぶ．

(3) 基準リンクグラフ内の各ノードが，基準ノードのセグメントツリーのどのノードからリンクされているかを抽出する．例えば,「Cricket」の記事のセグメントツリーにおいて「Batting」のアンカー文字列がどのノードからリンクが張られているのかを抽出する．

(4) 基準ノードとリンクが張られている回数の多い記事，図 5 より，リンクの貼られいてる位置がセグメントツリーの深さの浅い所にあり，尚且つセグメントツリーの左側すなわち目次番号が早いほど基準ノードにとって関連性が高いと考え，以下の式 (1) を用いて関連度 W_{kl} を計算をする．

$$W_{\mathrm{kl}} = af * cos(k,l) + \sum_{i=1}^{af} (\frac{1}{d_{\mathrm{i}}})^{n_{\mathrm{i}}} * (n_{\mathrm{i}} - o_{\mathrm{i}} + 1) \tag{1}$$

ここで af は基準ノードの記事と双方向リンクされているある記事に対するアンカー文字列の個数を指す．d_i は基準ノードの記事と双方向リンクされているノード i の深さを，n_i はノード i との兄弟ノードの数を，o_i はノード i の深さにおける左からの順番を示している．$\cos(k,l)$ は基準ノードと比較対象ノードの出現頻度ベクトルにおけるコサイン相関値であり， k は基準ノードの記事の名詞の出現頻度ベクトルであり，l は基準リンクグラフ内のその他のノードの記事の名詞の出現頻度ベクトルを示している．

例えば，上記の「Cricket」の記事と双方向リンクされている「Batting」の記事の場合,「Batting」のアンカー文字列が「Cricket」のセグメントツリーの ”Bat and ball ”というセグメントと ”Pitch, wickets and creases ”というセグメントに存在することから af=2 となる．”Bat and ball ”はセグメントツリーにおいて深さ 3 のところに存在するため d_1=3, 深さ 3 のノードの数が 25 あるため n_1=25 となり，深さ 3 において ”Bat and ball ”は左から 11 番目に存在するため o_1=11 となる．同様に ”Pitch, wickets and crease ”における各値は d_2=3, n_2=25, o_2=3．これらを上記の (1) 式に代入し関連度の計算を行う．このようにして Batting の関連度は 0.71 となった．関連度が事前に設定した閾値以上の記事を，比較対象の記事とする．閾値が 0.6 である場合,「Batting」は閾値以上の関連度があるため,

© 2011 Information Processing Society of Japan

Vol.2011-DBS-153 No.4
2011/11/3

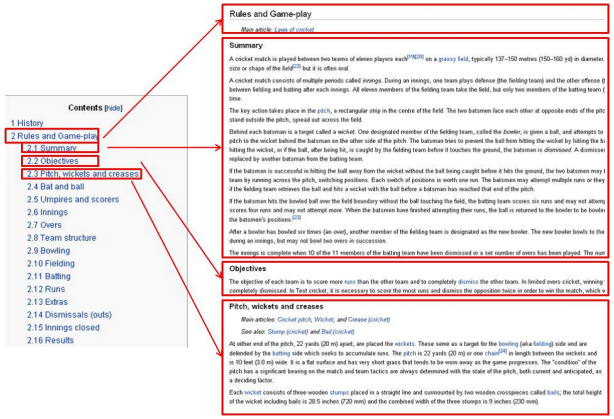

図 4 Wikipedia の目次構造の分割
Fig. 4 segment contents structure of Wikipedia

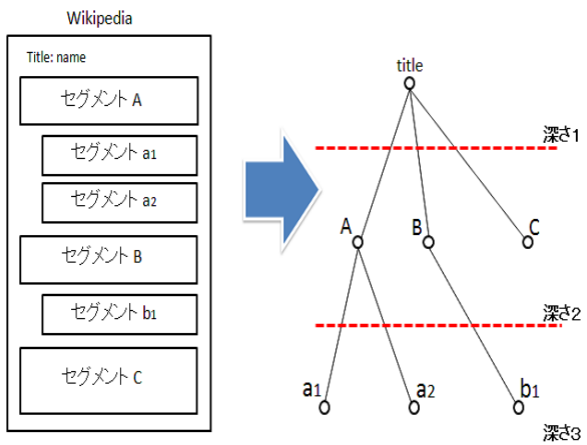

図 5 Wikipedia の記事を木構造化
Fig. 5 Tree posture Creator of the article of Wikipedia

比較対象の記事となる．

4. 多言語記事の差分抽出

4.1 多言語 Wikipedia の比較

言語にかかわらず，Wikipedia の記事は目次構造に基づいて段落に分かれていることが多い．このとき，Wikipedia の段落は意味的に分かれている可能性が高いと考えられる．そこで，我々は多言語 Wikipedia を比較する際に目次構造に注目し，目次構造に基づくコンテンツの比較を行い (**図 6** 参照)，類似している段落の中から，その差分情報を抽出する．

ここでは，日本語版記事，英語版記事の段落毎にテキストの形態素解析を行い，名詞だけを抽出する．そして比較を行うため辞書を用いて英語の名詞を日本語に翻訳する．本研究では翻訳のために GENE95 辞書[⋆1]を使用する．また，GENE95 辞書に掲載されていない単語は Google Ajax api[⋆2] と Microsoft Translator api[⋆3] の翻訳を使用する．しかし，「Bowls」や「World Cricket League」といったような固有名詞や「ヘンリー 8 世」などの人名は翻訳することができない．そこで，Wikipedia の言語間リンクも用いて翻訳を行う．ここでは，ある単語の翻訳語を探したいとき，その単語に関する記事を検索し，その記事に含まれる言語間リンクを用いて翻訳語の言語版記事を抽出し，その記事のタイトルを翻訳語とする．なお，翻訳時に単語の多義性など意味の曖昧性が問題になるが，今回はこの問題は考慮せず，今後の課題とする．

次に日本語版記事と英語版記事に含まれる名詞の出現回数を求める．そして，以下の (2) 式のコサイン類似度を用いて，各々の段落における日本語版の目次と英語版のセグメントの類似度を求め，ある閾値以下の段落を差分情報として抽出する．なおここでの閾値は実験より 0.3 とする．

$$\cos(x, y) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}} \tag{2}$$

ここでの x は日本語版記事の 1 つのセグメントであり，y は英語版記事の 1 つのセグメントである．x_i は日本語版記事の 1 つのセグメント内に存在する名詞 i の出現回数．y_i は英語版記事の 1 つのセグメント内に存在する日本語に翻訳した名詞 i の出現回数を表す．

⋆1 GENE95 辞書 http://www.namazu.org/ tsuchiya/sdic/data/gene.html
⋆2 Google Ajax api http://code.google.com/apis/language/
⋆3 Microsoft Translator api http://www.microsofttranslator.com/dev/

© 2011 Information Processing Society of Japan

Vol.2011-DBS-153 No.4
2011/11/3

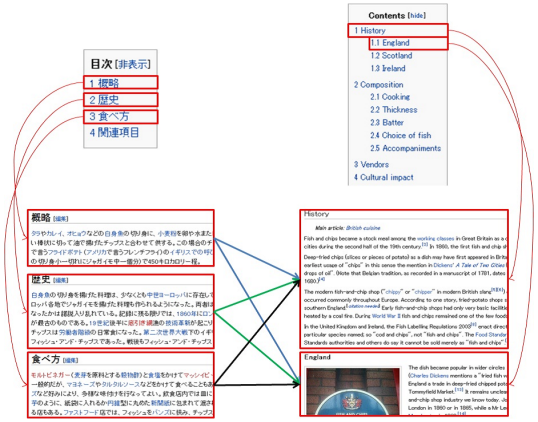

図 6 目次構造とコンテンツ
Fig. 6 Table of contents structure and contents

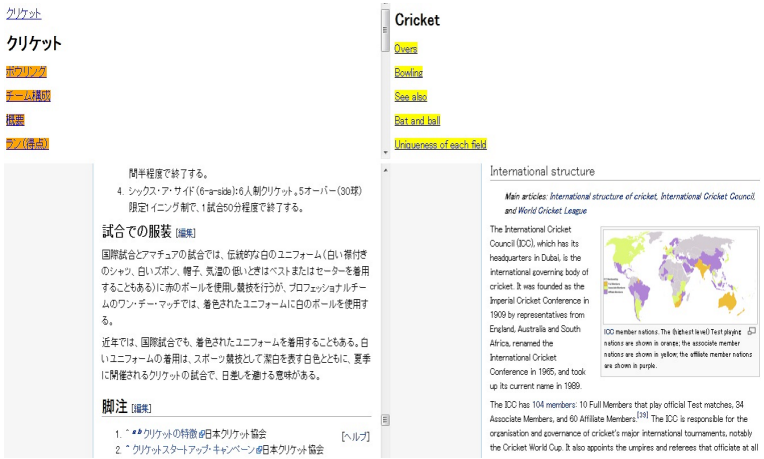

図 7 プロトタイプシステムの出力画面
Fig. 7 output of prototypesystem

5. プロトタイプシステム

我々が作成したプロトタイプシステムでは，開発言語に Ruby[⋆1]，日本語の形態素解析器に Mecab[⋆2]，英語の形態素解析器に Tree Tagger[⋆3]，データベースに MySQL[⋆4]を用いた．プロトタイプシステムではユーザがキーワードを入力すると，システムは出力画面**図 7** に示すように，日本語と英語の差分情報を提示する．図 7 に示す出力画面では左側に日本語版記事の目次コンテンツを，右側に英語版記事の目次コンテンツを示している．英語版記事が複数にまたがる場合は左上に表示されてあるタブを押すことでその記事の目次に切り替えが可能である．なお，得られた差分情報は日本語版にしか存在しない目次コンテンツをオレンジ色のマーカーで，英語版にしかない目次コンテンツを黄色のマーカーで示してある．目次をクリックすると Wikipedia 上のその記事の目次が下のフレームに表示され，コンテンツの内容を確認できるようになっている．

⋆1 Ruby http://www.ruby-lang.org/ja/
⋆2 Mecab http://mecab.sourceforge.net/
⋆3 Tree Tagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
⋆4 Mysql http://www-jp.mysql.com/

6. 実　　験

本論文で提案した二つの手法が有用であることを示すために，評価実験を行った．まず，ある言語版で書かれた一つの記事に対応する他の言語版記事を正しく抽出することができるかどうかを，評価実験 1 で確かめた．次に評価実験 2 において，比較対象記事から適切な差分を抽出することができているかどうかを確かめた．

6.1 実験 1：比較対象記事の抽出手法の比較実験

本論文にて提案した比較対象記事の取得方法の適合率を確認するため，以前の我々の提案手法[8] と本論文の提案手法の適合率，再現率，F 値を比較した．なお，実験データとしてイギリスの文化である英語版の Cricket，Warwick Castle，Snooker，Fish and chips，Goodwood Festival of Speed，Bowls，Polo，Association football の 8 つの単語を用い，関連度の閾値を 0.6，0.7 と比較し，平均適合率が高かった 0.6 を採用した．再現率の正解データは人手により抽出した．正解データの数は Criket では 56 個，Warwick Castle では 2 個，Snooker では 15 個，Fish and Chips では 5 個，Goodwood Festival of Speed では 2 個，Bowls では 1 個，Polo では 4 個，Association football では 9 個である．

© 2011 Information Processing Society of Japan

Vol.2011-DBS-153 No.4
2011/11/3

表 1 実験 1 の結果
Table 1 Result of experiment1

クエリ	過去の手法			提案手法		
	適合率	再現率	F 値	適合率	再現率	F 値
Cricket	100 %	40 %	50 %	97 %	55 %	70 %
Warwick Castle	0 %	0 %	0 %	50 %	33 %	40 %
Snooker	100 %	7 %	13 %	63 %	33 %	43 %
Fish and chips	67 %	20 %	31 %	50 %	60 %	55 %
Goodwood Festival of Speed	0 %	0 %	0 %	100 %	100 %	100 %
Bowls	0 %	0 %	0 %	0 %	0 %	0 %
Polo	0 %	0 %	0 %	100 %	25 %	40 %
Association football	100 %	33 %	50 %	100 %	63 %	77 %
平均	46 %	13 %	19 %	70 %	46 %	53 %

結果と考察

結果を**表 1** に示す．抽出された記事の例として Cricket であれば Cricket と野球との違いや Cricket の歴史．Warwick Castle であれば Warwick Castle の歴代の所有者の一覧の記事などを抽出することができた．

表 1 からわかるように以前の手法よりも本論文の提案手法の方が適合率が 46% から 70% ，再現率が 13% から 46% ，F 値が 19% から 53% と向上し，提案手法の有用性を示すことができた．これは過去の手法では，比較対象の複数の英語版記事を抽出するときに，ある記事と記事との類似性に注目しているため，適合率が低下していたためである．例えば Cricket において，List of international Cricket Council members と言った記事は Cricket の様に競技の説明をしているわけではないので類似性が低くなり結果が悪くなっていた．それに対し，本提案手法では，記事間のコンテンツの類似度ではなくアンカー文字列の位置と出現回数を考慮した関連度を用いているため，List of international Cricket Council members の記事は Cricket の記事の上位にアンカー文字列が多く，しかもその出現回数が多いため関連度が高くなり，結果，適合率，再現率，F 値が高くなったことがわかる．また，Cricket や Association football など，複数ページにまたがっている記事は適合率，再現率，F 値が高くなり，良い結果となった．理由としては記事の量が多いためアンカー文字列の出現回数が多くなり，関連度が大きくなったためであると考えられる．

結果が悪い例として，Snooker の記事において Snooker のルールや，リーグなどが抽出できなかった点が挙げられる．原因は，ある深さにおける目次項目が多いため n_i の値が大きくなり，目次の番号が若くても上記 (1) の式で加える値が小さくなってしまったからであると考えられる．また Bowls の結果は以前の提案手法と同様に関連記事を抽出することができなかった．この原因は正解データである World Bowls Events がセグメントツリーにおいて右下に存在しルートノードに遠かったためである．本論文の提案手法ではセグメントツリーにおいてリンクの貼られいてる位置が浅く，尚且つ左側すなわち目次番号が早いほど基準ノードにとって関連性が高いと考えいため World Bowls Events は値が低くなった．これは今後の課題である．

表 2 実験 2 の結果
Table 2 Result of experiment2

クエリ	適合率	再現率	F 値
ローンボウルズ	56 %	83 %	69 %
クリケット	79 %	88 %	83 %
ハッシュドビーフ	100 %	75 %	86 %
ジョン・ネイピア	83 %	63 %	71 %
スカッシュ	88 %	74 %	80 %
平均	81 %	80 %	80 %

6.2 実験 2：差分抽出手法の評価実験

次に，作成したプロトタイプシステムを用いて提案の差分抽出手法の有用性を示すために，評価実験を行った．入力クエリとしてイギリスの文化に関する 5 種類のクエリ（ローンボウルズ，クリケット，ハッシュドビーフ，ジョン・ネイピア，スカッシュ）をそれぞれ用いた．尚，今回は差分情報抽出手法の有用性を示すだけであるので本実験における比較対象のページは各クエリ 1 ページとした．プロトタイプシステムで得られた差分情報の適合率，再現率，F 値を求める．ここでの再現率は人手により抽出した英語だけにしかない目次コンテンツを正解データとした．正解データはローンボウルズでは 6 個，クリケットでは 25 個，ハッシュドビーフでは 8 個，ジョン・ネイピアでは 8 個，スカッシュでは 19 個である．

結果と考察

実験結果を**表 2** に示す．表 2 から分かるように，F 値の平均が 80%と良い結果を得ることができた．差分情報の例として「クリケット」という記事であれば日本語版には記載されていない目次コンテンツとして「History」，といったクリケットにおける歴史や「International structure」といったクリケットにおける国際構造についての情報が抽出された．また，ハッシュドビーフやスカッシュは日本語版と英語版との目次の数の差や情報の量が大きく異なっており，適合率が良くなったと考えられる．それに対し，ローンボウルズは悪い結果となった．原因として，ローンボウルズの記事に頻出するインドア・ローンボウルズやクラウン・

© 2011 Information Processing Society of Japan

ローンボウルズなどの固有名詞が翻訳できなかったためである．また，strike がストライクではなくストライキに，head が頭ではなくヘッドのような表記ゆれや，grass が芝生ではなく種をまくと言った単語の多義性により正しく翻訳されない場合があった．今後はこの多義性を緩和することを課題し適合率の向上を図る．

7. まとめと今後の課題

本論文では, ユーザの入力したキーワードに対応する一つの言語版とそれ以外の言語版との 2 つの言語版の Wikipedia 上での差分情報を取得する手法の提案を行った．評価実験では 1 つの言語版として日本語版，それ以外の言語版として英語版を利用した．差分情報を抽出する際，日本語版記事と英語版記事の情報の粒度の違いを考慮し，我々の提案した関連度を用いて比較する複数の英語版記事の抽出手法を提案した．評価実験の結果，適合率の平均が 72% 再現率が 49% F 値が 57% となり，以前の提案手法より適合率が向上した．また，記事の目次構造に基づいた比較手法の提案を行った．実験の結果，適合率の平均が 81%，再現率の平均 80%，そして F 値の平均が 80%となり，良い結果を得ることができた．今後の課題は以下の通りである．

- 重要度，信頼度の計算
  本論文では差分情報の取得方法に関する提案を行ったが，ある言語版に存在しない情報が他方の言語版に存在した場合であっても，その情報を提示するべきでない場合もある．なぜならば，提示された情報がユーザにとって不要であれば，提案システムの有用性が損なわれるシステムとなってしまうためである．そこで今後は，得られた差分情報の重要度，信頼度を測定し他方の言語版に加えるために，重要度，信頼度の算出手法を提案し，Wikipedia の情報の網羅性，信頼性の向上を図る．
- 単語の多義性
  本研究では，多言語間の差分情報抽出手法を提案した．しかし，今回は単語の多義性は考慮していない．そのため今後は，単語の多義性を解消させることで差分情報抽出手法の適合率を向上させることに取り組む．
- 差分情報の提示方法
  本研究では，差分情報の抽出方法についての提案を行ったが，その提示方法については言及していない．今後は，差分情報を他方の言語版に加え提示する予定である．そのため，差分情報の提示方法について考案する必要がある．
- 英語以外の言語版への適応
  本評価実験では日本語版記事と英語版記事を比較した．今後は比較対象を英語版のみならず中国語や韓国語等他の言語への対応を行いたい．また，二言語間の比較から三言語間の比較へと新たに取り組む．

参 考 文 献

1) 森　竜也，増田英孝，中川裕志，清田陽司：Wikipedia における言語間の差異マイニング，情報処理学会創立 50 周年記念 (第 72 回) 全国大会， No.5, pp.181–182 (2010).
2) 森　竜也，増田英孝，中川裕志，清田陽司：多言語に展開する Wikipedia の特徴の比較調査，第 21 回セマンティックウェブとオントロジー研究会 (2009).
3) 斎藤雄介，山田剛一，絹川博之，中川裕志：日中英ニュース記事比較のための収集と検索，情報処理学会 第 71 回 全国大会， No.2, pp.269–270 (2009).
4) 中崎寛之，川場真理子，山崎小有里，宇津呂武仁，福原知宏：同一トピックの日英ブログにおける文化間差異の発見支援，*DEIM Forum 2009*.
5) 立床雅司，高橋　慧，斎藤昭則，吉川正俊：Wikipedia のリンク構造に基づく関連度を利用したコンテンツ検索手法と地球科学データへの応用，*DEIM Forum 2009*, pp. 1–8.
6) K. Nakayama, T. Hara, S. Nishio：Wikipedia Mining for An Association Web Thesaurus Construction, *WISE 2007*, pp.1–11.
7) 藤原裕也，灘本明代：言語 Wikipedia を用いた伝統文化の差異情報抽出の提案，情報処理学会 第 73 回 全国大会， No.1, pp.1.575–1.576.
8) 藤原裕也，灘本明代：Wikipedia の言語間比較による差異情報抽出手法の提案，情報処理学会研究報告， データベース・システム研究会報告， No.2011.

© 2011 Information Processing Society of Japan