2C1-NFC-2-3

コンテンツの質を考慮した 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] 小西 幸男[*3] 灘本 明代[*4]
Yuya Fujiwara Yu Suzuki Yukio Konishi Akiyo Nadamoto

[*1]甲南大学大学院 自然科学研究科
Graduate School of natural science graduate course, Konan University

[*2]名古屋大学 情報基盤センター
Information Technology Center, Nagoya University

[*3]甲南大学 国際交流センター
Konan International Exchange Center, Konan University

[*4]甲南大学 知能情報学部
Dept.of Intelligene and Informatics, Konan University

We propose a system which extracts different information between two languages on the Wikipedia and presents it. When we compare two languages of Wikipedia articles, the granularity of information between them are different. Therefore, we propose a method for extracting multiple comparison articles using a link graph of Wikipedia. Then the system extracts different information that is included in comparison articles in Wikipedia by comparing one base article with the other comparison articles found using the link graph. Furthermore, we propose the calculation of credibility of the different information on the Wikipedia.

1. はじめに

本論文では Wikipedia において，ユーザの入力したキーワードに関するユーザの母国語版記事と比較言語版記事をそれぞれ抽出し，記事の内容を比較した上でその信頼度を算出し，信頼度の高い差分情報をユーザに提示する手法を提案する．

Wikipedia の記事は現在 285 言語で記述されているが，ユーザは母国語で記事を書く場合が多いため，一つの話題に対して各言語版の文書量や詳細度が異なる場合が多数存在する．例えば，イギリスの伝統的なスポーツである「ローンボウルズ」の英語版記事には，多岐にわたって詳細な記述がなされているが，同一項目に関する日本語版記事は記述量が少ない．なぜなら，日本語版の記事を作成しているユーザは主に日本人であり,「ローンボウルズ」のことについて十分な知識を持っているユーザが少ない点や，イギリス人と比較して日本人は比較的ローンボウルズに対して興味が薄いと点が原因であると考えられる．このように，Wikipedia の記事は言語版によって情報の量が異なり，母国語版だけでは得られる情報が不足する場合がある．

一般的にユーザは，母国語版の記事を読むことはあるが，母国語以外の言語版の記事を読むことは少ないと考えられる．なぜなら，ユーザにとって母国語で記述されている記事を読む労力と比較して他言語版の記事を読む労力は比較的大きいためである．ところが，他言語版の記事に含まれる情報をユーザが得ることは，ユーザにとって有用であると考えられる．そこで我々は，ユーザが閲覧している母国語版に不足している情報を他の言語版から抽出し追加することによって，閲覧している記事に関するユーザの理解度が上がると考えた．そこで本論文では，多言語 Wikipedia 間の差分情報抽出，提示システムを提案する．

ところが，Wikipedia は誰もが編集を行うことができるため，記述されている情報が常に正しいとは限らない．つまり，誤った記述が投稿された場合でも Wikipedia はその記述をそのまま記事に反映させるため，Wikipedia の記事には誤った情報が含まれる場合があり，記事の全部が必ずしも信頼できる情報だけで構成されているわけではない．既存研究において我々は，Wikipedia の記事編集履歴から信頼度を算出する方法について提案を行っていた．そこで本研究では，多言語 Wikipedia を比較し取得した差分情報の信頼度を求め，ある程度信頼性の高い情報だけを提示することを行う． 信頼度算出手法を差分抽出手法と組み合わせることによって，ユーザはより有益な情報を容易に取得することができると考えられる．

本論文では次の流れで Wikipedia の差分情報を抽出する．まずユーザは，興味がある情報に関するキーワードをクエリとして入力する．次にシステムはクエリをタイトルとする母国語版の記事を取得する．このとき，言語間リンクを利用して比較言語版で書かれた記事を取得する．本論文では比較言語版として英語を想定しているが，言語に依存した手法ではないため他の言語にも対応することができる．そして次に，比較言語版の記事と相互リンクしている記事群のうち内容が類似している記事群を比較言語版記事群として抽出する．これら比較言語版記事群と母国版記事をそれぞれ比較し，差分情報を抽出する．最後に，これら差分情報の質を求め，高い質の情報だけを利用者に提示する．

2. 関連研究

Wikipedia における記事間の関連抽出に関する研究は，多数存在する．例えば，Strube ら [Strube 06] や Milne ら [Milne 07] は，Wikipedia の記事が属するカテゴリ情報を利用して，概念間の関連度を算出している．我々の提案している関連度は記事間のリンクを利用している点が異なる． Nakayama ら [Nakayama 07] は Wikipedia のリンク構造を解析しシソーラス辞書を構築する手法を提案している．その際に，記事から記事へのパスの多さ，各パスの長さに着目し関連度を計測している．それに対し，本研究ではリンク数，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．

Wikipedia のリンク構造を用いてシソーラスを構築している研究も行われている．David[Milne 06] は Wikipedia のリンク構造を用いて，語の意味的関係を抽出している．Chen ら

連絡先: 藤原裕也，甲南大学大学院自然科学研究科，兵庫県神戸市東灘区岡本 8-9-1，mn124006@center.konan-u.ac.jp

[Chen 03] は Web ページ同士のリンク構造を解析し、Web シソーラス辞書を自動的に構築する手法を提案している．本研究との違いはリンク構造を用いて Wikipedia の記事同士の関連する度合い抽出する点で異なる．

多言語 Wikipedia を用いて言語版同士を比較している研究も存在する．Eytan ら [Adar 09] は Wikipedia の infobox の違いを抽出し他の言語版へ補完を行っている．我々は目次構造に着目し Wikipedia の記事を比較しているため，片方の言語版記事にない目次や記事の内容を知ることができる点が異なる．

Wikipedia 記事の質に関する研究も多く存在する．Wöhner ら [Wöhner 09] は記事の周期的に起こる変化に着目することによって，記事の質を測定している．Hu ら [Hu 07] は記事と編集者の相互依存関係に基づき，記事の質の高さによって順位付けを行っている．これに対して我々は，記事の残留度に注目して質を算出することによって，小さな計算量で十分な精度の質を算出することができる点が異なる．

3. 差分情報の抽出

3.1 比較対象記事群の抽出

多言語 Wikipedia は各言語版により情報の粒度が異なり，対応する記事が複数存在する場合がある．例えば「ローンボウルズ」の場合，日本語版ではローンボウルズの競技用の芝生であるボウリンググリーンに関して説明されている．英語版は同様に記事の中にボウリンググリーンの説明があるが，さらにそれを詳しく説明した「Bowling green」という別の記事が存在し，ローンボウルズに関する事項は複数の記事にまたがっている．そこで我々は Wikipedia のリンク構造を解析し，記事間の関連度を求め，この関連度により複数にまたがっている記事群を抽出する．以下に比較対象となる複数記事の抽出手法を示す．

1. ユーザの入力したクエリと同じタイトルを持つ比較言語記事から比較言語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しているノードは，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いるノード以外のノードを削除する．

3. 2. の基準リンクグラフにおいて，基準ノードとその他のノード間の関連度を求める．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比較言語記事群とする．

上記手順 3. で求める関連度とは，Wikipedia のある記事と双方向リンクされている記事との関わりがどれだけ深いかを調べるための尺度である．今回我々は新たに基準ノードの記事の部分的な情報と比較対象ノードの類似性、アンカー文字列の出現位置，アンカー文字列の出現回数に注目し関連度を求める手法を提案する．以下に関連度を抽出する手順を示す．

1. 基準ノードとなる Wikipedia の記事を分割する．この時，Wikipedia の一番最初の説明部分をサマリとしそれ以外のコンテンツをセグメントと呼ぶ．

2. 基準リンクグラフ内の各ノードが，基準ノードのサマリ，セグメントのどの部分からリンクを張っているかを全て抽出する．

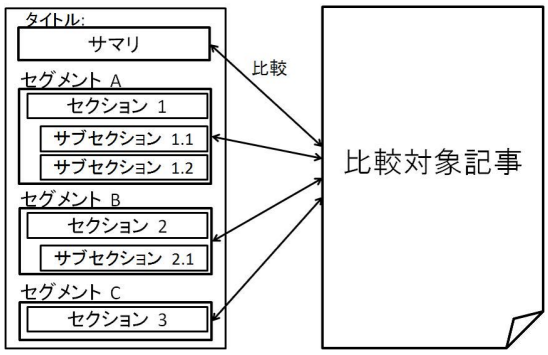

図 1: 関連度

3. そのサマリまたはセグメントと比較対象ノードとを対象にし，以下の式 (1) (2) を用いて関連度を求める (図 1)．

$$R_{kl} = \alpha \cdot (tf_{sum} \cdot S_{sum}) + \sum_{i=1}^{n} (tf_i \cdot S_i) \quad (1)$$

$$W_{kl} = R_{kl} / \max(R_{\mathrm{km}}) \quad (2)$$

ここでの tf_{sum} は基準ノード k のサマリに出現するアンカー文字列の出現回数を指す．S_{sum} は基準ノード k のサマリと比較対象ノード l の名詞の出現頻度を重みとしたコサイン類似度である．tf_i はそれ以外のあるノードすなわちある部分グラフに出現するアンカー文字列の出現回数を，そして S_i は基準ノード k におけるその部分グラフと比較対象ノード l の名詞の出現頻度を重みとしたコサイン類似度である．なお，我々は基準ノードのサマリにリンクを張っている比較対象の記事は関連性が高いと考え α を掛けている．なぜならサマリの情報はその記事全体の大まかな概要を示してあり，そこに出現するアンカー文字列はこの記事を知る上で必要不可欠な情報であると判断したためである．なお，ここでの関連度の式 (1) の重み α と閾値 β を，予備実験の結果から α=3，β=0.2 と設定した．

3.2 コンテンツの比較

言語にかかわらず，Wikipedia の記事は目次構造に基づきセグメントに分かれていることが多い．このとき，Wikipedia のセグメント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 を比較する際にセグメントに注目し，セグメントに基づくコンテンツの比較を行い，類似していないセグメントを抽出し差分情報とする．まず初めに母国語記事と比較言語記事をそれぞれ形態素解析を行い，名詞のみ抽出する．次に辞書を用いて比較言語記事の名詞を母国語に翻訳する．また，辞書に掲載されていない単語は Google Ajax API や Microsoft Translate API を用いて翻訳を行う．この時，固有名詞や人名などは翻訳することができない場合が多い．そこで Wikipedia の言語間リンクを用いて翻訳を行う．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 Wikipedia の記事を検索し，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することにより，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となるが，本論文ではこの問題は考慮せず，今後の課題とする．次に母国語記事と比較言語記事のセグメントごとに名詞の出現回数を求める．以下の式 (3) のコサイン類似度を用いて，母国語記事と比較言語記事のセグメントごとの類似度を求める．ある比較言語記事のセグメントが母国語記事の全てのセグメントに対し類似度が閾値 γ 以下であっ

た場合，その比較言語記事のセグメントを差分情報として抽出する．

$$\cos(x,y) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}} \tag{3}$$

ここでの x は母国語記事の 1 つのセグメントであり，y は比較言語記事の 1 つのセグメントである．x_i は母国語記事の 1 つのセグメント内に存在する名詞 i の出現回数のベクトルであり，y_i は比較言語記事の 1 つのセグメント内に存在する母国語に翻訳した名詞 i の出現回数ベクトルを表す．

3.3 コンテンツの質の計算

コンテンツの比較によって得られた差分情報は，必ずしも有益な情報であるとは限らない．Wikipedia の記事には質の低い情報が混在しているため，得られた差分情報の中には質の低い情報が存在することがある．そこで我々は以前，文献 [鈴木 10] によって提案された，記事の残存率に基づく質の算出手法を用いて記事の情報の質を測定する．

具体的な質の算出手法について述べる．Wikipedia には記事 $i = 1, 2, \cdots, M$ が存在し，それぞれの記事には $v_{i,j} \in V$ のバージョンが存在する．V は全ての記事の全てのバージョンの集合である．ここで $j = 0, 1, \cdots, N_i$ は記事のバージョン番号を表す．$j = 1$ のとき，つまり $v_{i,1}$ は記事 i が最初に作成されたバージョンを表す．$j = 0$ のとき，つまり $v_{i,0}$ は記事 i が作成されているが内容が無い状態を表す．Wikipedia の記事を記述した著者 $e = 1, 2, \cdots, K$ は，一つ以上の記事のバージョンを作成している．著者 e が作成した記事は $A_e = \{v_{i,j} | v_{i,j} \in V \text{ and } v_{i,j} \text{ is written by } e\}$ であり，一つのバージョンを作成した著者は一人である．ただし，$j = 0$ のバージョンの著者は存在しないと仮定する．

まず，j 回目の編集においてどの部分を追加・削除したかを特定するために，$v_{i,j-1}$ と $v_{i,j}$ との増加部分 $add_{i,j}$ および削除部分 $del_{i,j}$ を求める．

次に，p $(p = 0, 1, \cdots, N_i - j)$ 回後に編集されたバージョン $v_{i,j+p}$ において，$add_{i,j}$ と $del_{i,j}$ が残存している割合を算出する．ここで，$p = 0$ のときは $\delta(add_{i,j}, 0) = add_{i,j}$ ，$\delta(del_{i,j}, 0) = del_{i,j}$ とする．まず，$v_{i,j+p}$ の中から $add_{i,j}$ ，$del_{i,j}$ に相当する部分 $\delta(add_{i,j}, p), \delta(del_{i,j}, p)$ を抽出する．次に，追加部分，削除部分の残存率である追加残存率，削除残存率 $R^{add}(i,j,p), R^{del}(i,j,p)$ を (4) ，(5) 式によって求める．

$$R^{add}(i,j,p) = \frac{|\delta(add_{i,j}, p)|}{|add_{i,j}|} \tag{4}$$

$$R^{del}(i,j,p) = \frac{|\delta(del_{i,j}, p)|}{|del_{i,j}|} \tag{5}$$

ここで，$|\delta(add_{i,j}, p)|, |\delta(del_{i,j}, p)|, |add_{i,j}|, |del_{i,j}|$ はそれぞれ，$\delta(add_{i,j}, p), \delta(del_{i,j}, p), add_{i,j}, del_{i,j}$ に含まれる文字数である．

次に，標準残存率を利用して残存率を正規化する．標準残存率とは p 回後に追加，削除された時の残存率の平均値である．標準残存率を利用して正規化を行う理由として，事前に行った予備実験において，編集回数が増加するごとに残存率が低下することが分かったことが挙げられる．正規化された残存率 $\overline{R^{add}(i,j,p)}$，$\overline{R^{del}(i,j,p)}$ を (6), (7) 式によって求める．

$$\overline{R^{add}(i,j,p)} = \frac{R^{add}(i,j,p)}{S^{add}(p)} \tag{6}$$

$$\overline{R^{del}(i,j,p)} = \frac{R^{del}(i,j,p)}{S^{del}(p)} \tag{7}$$

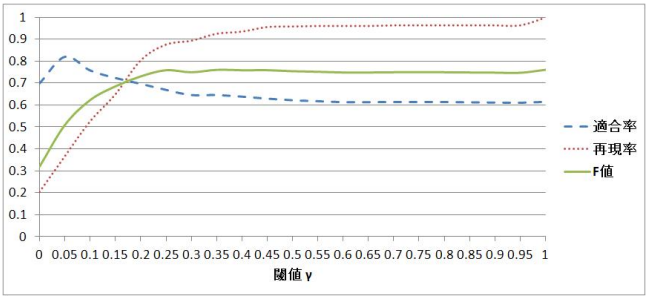

図 2: 閾値 γ と適合率，再現率，F 値の関係

ここで $S^{add}(p)$, $S^{del}(p)$ はそれぞれ p 回後の編集における残存率の平均値である．

そして，追加残存率と削除残存率を組み合わせて，記事変更信頼度 $\tau(v_{i,j})$ を求める．記事変更信頼度は，追加残存率と削除残存率の総和であり，(8) 式で求める．

$$\tau(v_{i,j}) = \sum_{q=1}^{N_i - j} R^{add}(i,j,p) + \sum_{q=1}^{N_i - j} R^{del}(i,j,p) \tag{8}$$

記事変更信頼度を算出するとき，編集回数による正規化を行わない．なぜならば，編集回数が多いとき編集の記事変更質は高くなるべきであると考えたためである．

最後に，記事変更信頼度の平均を 0 とする．なぜならば，Wikipedia における記事変更の大半は小さな変更であり，記事自体の質は変化しないと考えられる．それらの記事変更信頼度の変化よりも低い記事変更信頼度があったとき，その変更は記事の質を低下させていると考えられるためである．最終的な記事変更信頼度 $\overline{\tau(v_{i,j})}$ は (9) 式で求める．

$$\overline{\tau(v_{i,j})} = \tau(v_{i,j}) - \frac{\sum_{i=1}^{M}\sum_{j=1}^{N_i} \tau(v_{i,j})}{\sum_{i=1}^{M} N_i} \tag{9}$$

この $\overline{\tau(v_{i,j})}$ が閾値 σ 以上である節だけをユーザに提示する．

4. 評価実験

差分抽出手法の有用性を示すための評価実験を行った．予備実験より得られた関連度を求める際の重み α を 3.0 とし，関連度の閾値 β を 0.2 とした．そして類似度の閾値 γ を 0.0 から 1.0 で 0.05 間隔に設定し，各閾値における差分情報の適合率，再現率，F 値を求めた．なお，正解データは人手により抽出を行った．また，ここでの差分情報は母国語版にない情報とする．

図 2 に，γ を変化させたときの再現率，適合率，および F 値を示す．この図から，適合率と再現率が共に高い値を示している γ は 0.2 であることから，γ の値を 0.2 と設定した．このときの各クエリごとの適合率，再現率，F 値の結果を表 1 に示す．なお，Q1 から Q13 はそれぞれバノック，ウォリック城，ブラックドッグ (亡霊)，フィッシュ・アンド・チップス，グッドウッド・フェスティバル・オブ・スピード，ローンボウルズ，ブルー・プラーク，バーレスク，スコットランドの国旗，ゲーリックハンドボール，キッパー (魚料理) ，スコットランド国立美術館，リプトンを示している．この表より，適合率の平均が 70%，再現率の平均 81%，F 値の平均が 73% と高い値であるため，提案手法が有用であることを示すことができたと考えている．

表 1: 各クエリの適合率，再現率，F 値

クエリ	正解数	適合率 (%)	再現率 (%)	F 値
Q1	2	33	50	40
Q2	12	79	92	85
Q3	32	89	78	83
Q4	11	45	82	58
Q5	10	60	60	60
Q6	9	50	100	67
Q7	28	76	89	92
Q8	22	71	45	56
Q9	56	98	88	92
Q10	16	68	94	79
Q11	16	88	94	91
Q12	4	57	100	72
Q13	8	71	63	67
平均	-	70	81	73

高精度に抽出された差分情報の例として，ブラックドッグ (亡霊) であればイングランドやスコットランドでのブラックドッグの伝承のされ方や種類などについて書かれている情報が抽出することができた点が挙げられる．また，イギリスの伝統的なスポーツであるゲーリックハンドボールであればゲーリックハンドボールの種類などが差分情報として抽出することができた．実験の結果から，精度が良いときには，母国語版である日本語版と比較言語である英語版とのコンテンツの量が大きく異なっており，差分情報が多く存在したため，F 値が高くなったと考えられる．例えばブラックドッグ (亡霊) の場合は，日本語版はブラックドッグ (亡霊)，概要，関連項目，参考資料といった 4 個のコンテンツに分かれているのに対して，英語版は 15 個のコンテンツに分かれており，コンテンツの量が大きく異なる．そのためコンテンツの比較の際に比較する回数が少なくなり，比較的差分情報が抽出されやすくなったと考えられる．

精度が悪かったクエリの例として，バノックやフィッシュ・アンド・チップスなどが挙げられる．これは形態素解析する際に固有名詞を一般名詞レベルまで分割したことが原因であると考えられる．例えば英語版の記事である Fish and chips であれば “Fish and chips” という出現頻度の高い固有名詞を “Fish” と “chips” という二つの一般名詞に分け，コンテンツの比較を行う際に類似している内容でも類似度が低くなり適合率が悪くなったと考えられる．これは今後の課題である．

また提案手法は，少量の情報の基準ノード記事と大量の情報の比較対象記事の場合には有効ではなかった．例えば，スコットランドの旗の種類について書いてある “Flag of Scotland” の記事と，イギリスの旗の種類について書いてある “Royal Standard of the United Kingdom” の記事が関連するかどうかを計算する場合を考える．このとき，“Flag of Scot land” の一部分の情報と “Royal Standard of the United Kingdom” の一部分の情報は似ているが，我々の提案している関連度が低くなり，関連するページとして抽出されなかった．これは “Flag of Scotland” の一部分の情報と “Royal Standard of the United Kingdom” 全体の情報を比較していたためである．

5. おわりに

本論文では Wikipedia において，ユーザの入力したキーワードに関するユーザの母国語版記事と比較言語版記事をそれぞれ抽出し，記事の内容を比較した上でその信頼度を算出し，信頼度の高い差分情報をユーザに提示する手法を提案した．差分情報を抽出する際，母国語版記事と比較言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言語版記事を抽出手法を提案した．そして母国語記事と抽出された複数の比較言語記事を比較し，その差分情報を抽出した．さらに抽出された比較言語記事の差分情報の情報の質を測定する手法の提案を行った．

今後は，情報の質を用いた精度の実験や，ユーザ実験などによる提案手法の有用性を示す実験を行う予定である．また，単語の翻訳を行う際に，本研究では単語の多義性は考慮しなかった．そのため今後は，単語の多義性を解消することによって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ことを考えている．

参考文献

[Adar 09] Adar, E., Skinner, M., and Weld, D. S.: Information arbitrage across multi-lingual wikipedia, *2nd ACM International Conference on Web Search and Data Mining*, pp. 94–103 (2009)

[Chen 03] Chen, Z., Liu, S., Wenyin, L., Pu, G., and Ma, W.-Y.: Building a web thesaurus from web link structure, *26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 48–55 (2003)

[Hu 07] Hu, M.: Measuring article quality in wikipedia: Models and Evaluation, *16th ACM conference on Conference on information and knowledge management*, pp. 243–252 (2007)

[Milne 06] Milne, D., Medelyan, O., and Witten, I. H.: Mining Domain-Specific Thesauri from Wikipedia: A case study, *The 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 442–448 (2006)

[Milne 07] Milne, D.: Computing Semantic Relatedness using Wikipedia Link Structure, *New Zealand Computer Science Research Student Conference*, Vol. 7, p. 8 (2007)

[Nakayama 07] Nakayama, K., Hara, T., and Nishio, S.: Wikipedia Mining for An Association Web Thesaurus Construction, *Web Information Systems Engineering*, Vol. 4831, pp. 322–334 (2007)

[Strube 06] Strube, M. and Ponzetto, S. P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia, *American Association for Artificial Intelligence*, Vol. 2, pp. 1419–1424 (2006)

[Wöhner 09] Wöhner, T. and Peters, R.: Assessing the Quality of Wikipedia Articles With Lifecycle Based Metrics, *5th International Symposium on Wikis and Open Colaboration*, pp. 1–10 (2009)

[鈴木 10] 鈴木 優, 吉川 正俊：Wikipedia におけるキーパーソン抽出による信頼度算出精度及び速度の改善, 情報処理学会論文誌:データベース (TOD47), 第 3 巻, pp. 20–32 (2010)