

異言語 Wikipedia を用いた補完情報の提示手法の提案

藤原 裕也[†] 鈴木 優^{††} 小西 幸男^{†††} 灘本 明代^{††††}

[†] 甲南大学大学院 自然科学研究科 〒658-8501 兵庫県神戸市東灘区岡本 8-9-1

^{††} 名古屋大学大学院 情報科学研究科 〒464-8601 愛知県名古屋市千種区不老町

^{†††} 甲南大学 国際交流センター 〒658-8501 兵庫県神戸市東灘区岡本 8-9-1

^{††††} 甲南大学 知能情報学部 〒658-8501 兵庫県神戸市東灘区岡本 8-9-1

E-mail: [†]mn124006@center.konan-u.ac.jp, ^{††}suzuki@db.ss.is.nagoya-u.ac.jp,

^{†††}ykonishi@center.konan-u.ac.jp, ^{††††}nadamoto@konan-u.ac.jp

あらまし Wikipedia は誰でも記事を作成・編集できる点から情報が不足の記事が多く存在する．さらに Wikipedia は多くの言語版が存在し、各々独自成長を遂げている．我々はその多言語性に注目し不足情報を異言語版から抽出し、それをユーザの閲覧している記事に補完する手法を提案してきた．今回、我々は新たに抽出したこの補完情報を提示する方法を提案する．具体的には、ユーザの閲覧している記事と抽出した補完情報との関係の深さを求め、その関係の深さに基づきユーザの閲覧している記事に補完情報のリンクを埋め込むことを行う．これにより、ユーザは閲覧している記事内のリンクをクリックすることにより、補完情報を容易に閲覧することが可能となる．

キーワード 情報補完, 情報抽出, Wikipedia, 多言語

Yuya FUJIWARA[†], Yu SUZUKI^{††}, Yukio KONISHI^{†††}, and Akiyo NADAMOTO^{††††}

[†] Graduate School of natural science graduate course, Konan University Graduate School 8-9-1 Okamoto Higashinada, Kobe, Hyogo, 6588501 Japan

^{††} Graduate School of Information Science, Nagoya University Furotyo, Chikusa, Nagoya, Aichi, 4640814 Japan

^{†††} Konan International Exchange Center, Konan University 8-9-1 Okamoto Higashinada, Kobe, Hyogo, 6588501 Japan

^{††††} Dept. of Intelligence and Informatics, Konan University 8-9-1 Okamoto Higashinada, Kobe, Hyogo, 6588501 Japan

E-mail: [†]mn124006@center.konan-u.ac.jp, ^{††}suzuki@db.ss.is.nagoya-u.ac.jp,

^{†††}ykonishi@center.konan-u.ac.jp, ^{††††}nadamoto@konan-u.ac.jp

Key words Complement Information, Information extraction, Wikipedia, Multilingual

1. はじめに

Wikipedia はインターネット上で使われているフリーの百科事典である．その特徴として記事を誰でも作成・編集することが可能であり、さらに 2013 年 1 月 現在 285 言語以上の多言語版が存在するといった点が挙げられる．しかし、このように誰でも記事を編集できることから編集される記事が偏り、情報が不足する記事が多く存在する場合がある．また、Wikipedia の多言語性に注目するとそれぞれの言語版は独立で管理されており、それぞれの言語を使用しているユーザたちにより記事を作成・編集されている．しかし、このことからある話題に対して情報の量が言語版によって異なる場合があり、同じタイトルの

記事であってもある言語版は情報の量が大きく充実しているが、他の言語版では情報の量が少なく不足している場合がある．例えば、日本の武道の一つである「剣道」の日本語版 Wikipedia の記事の場合は目次の数も多く、情報も詳細に記述されてある．それに対し英語版の記事は日本語版より目次の数も少なく、情報の量も少なくなっている．この理由としてそもそも英語版の「剣道」を編集するユーザが少ないといった点や英語版の記事を編集・作成しているユーザが英語圏内のユーザが主であり、「剣道」に対して十分な知識を持っていないため編集することが困難であると考えられる．このことからユーザが内容の充実していない記事に対して情報の補完をすることは困難である場合が存在する．そこで我々は Wikipedia の多言語性に注目し、2 つ

の言語版の記事を比較し補完情報を抽出し、それを情報の不足している言語版の記事に補完することを行う。これにより情報不足の記事をより豊富なコンテンツにすることを目的とする。

これまで我々は情報が不足している記事に対して補完情報の抽出手法の提案を行った[1]が、その補完情報の提示方法の提案は行っていない。そこで今回我々は新たに抽出した補完情報を提示する方法を提案する。補完情報を提示する場合、ユーザが閲覧している記事と最も関連する場所にその補完情報を提示するのが好ましいと考えられる。そこで、本論文では補完情報と関連性の高いユーザが閲覧している記事の部分を抽出し、そこに補完情報のリンクを埋め込むことを行う。これにより、抽出した補完情報をユーザが閲覧している記事の適した場所に補完できるとともに、ユーザは記事内のリンクをクリックすることで容易に補完情報を閲覧することが可能となる。本論文ではユーザが閲覧したい言語を閲覧言語、補完情報を取得したい言語を比較対象言語と呼ぶ。さらにユーザの閲覧している記事閲覧記事、補完情報を抽出する比較対象言語の記事群を比較対象記事と呼ぶ。また、本論文の補完情報とは閲覧記事に存在しないかつ閲覧記事に関係のある情報のことを指す。

第2章で関連研究について、第3章で補完情報抽出手法について、第4章で補完情報提示手法について、第5章でプロトタイプシステムと評価実験、そして第6章でまとめと今後の課題について述べる。

2. 関連研究

Wikipediaにおける記事間の関連を抽出する研究は、数多く存在する。Michael [2] や Milne ら [3] 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。二つの概念が与えられたときに、カテゴリの中でその二つの概念がどれほど近いかを基準に関連度を算出している。これに対し、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異なる。高橋ら [4] は Wikipedia のリンク構造を用いて、歴史エンティティの重要度を算出している。具体的には、ある歴史エンティティから他の歴史エンティティへのリンクが張られている場合、前者は後者から影響を受けたと見なしている。また、Wikipedia の記事には時間や場所に関する情報が含まれることが多い。これらの情報に Brin ら [5] の PageRank に類似した反復計算アルゴリズムを用いて計算を行っている。Jaap ら [6] は Web のインリンクを重要と考え、Wikipedia と Web のリンク構造の違いについて研究を行っている。インリンクとは他サイトから自サイトへ貼られたリンクのことを表す。本研究では Wikipedia の記事同士の内容の違いを抽出しているため、ある話題に対して記事に掲載されていない情報や、ユーザが知らない情報を知ることができる点異なる。

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 [7] 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 [8] 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。具体的には Web サイトの階層構造からサブツリーと呼ばれる多数の木を形成し、木の中

での語の共起性を利用することで語の関連度を計測している。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を抽出する点異なる。

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 [9] 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。infobox とは Wikipedia の記事に含まれる、データが記述されている表のことである。しかし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異なる。Ching ら [10] は Wikipedia の多言語性に着目し記事をより豊かにするために Wikipedia の編集者を支援するためのフレームワークを提案している。具体的には比較する2つの Wikipedia の記事を文章ごとに分割し各文章における TF-IDF を求め、それを重みとしたコサイン類似度で類似度を測り、補完する記事と類似していない文章を補完情報として抽出し補完を行っている。しかし、彼らの研究は我々とは異なり Wikipedia の多言語間による情報の粒度の違いに着目していない。そのため、ある記事において複数にまたがっている記事が存在する場合はその情報を補完することはできない。我々は複数の比較対象記事の記事を抽出し、そこから補完情報を得ることができより完成された記事にする点異なる。吉野ら [11] は文化差の有無の判断は文化に関する十分な知識が必要となるため困難であると考え、文化差の有無を自動で判別する仕組みを提案している。その際に日本語版 Wikipedia と中国語版 Wikipedia を比較し文化差の検出を行っている。

情報補完に関する研究もいくつか行われている。Qiang ら [12] は TV ニュースで不足している情報を Web から取得し情報補完を行う手法を提案している。Qiang らは異メディア間の情報補完を提案しているのに対し、我々は Wikipedia の多言語性に注目して Wikipedia の情報補完を行う点異なる。また、Eklou ら [13] は Wikipedia の情報補完として、一般の Web ページから補完情報を抽出している。Wikipedia の情報補完を行う点は本研究と似ているが、本研究は Wikipedia の多言語性に注目している点異なる。また、Monz ら [14] はさまざまな言語で書かれた wiki ページ同士を比較し不足している情報を補完することでページを同期させる CoSyne というシステムを提案している。具体的にはある言語で書かれたある wiki ページと別の言語で書かれたある wiki ページを比較し不足情報を抽出し、適切な場所に補完することを行っている。本研究との違いは彼らはターゲットが wiki で作られたページ全てが対象である点、また単一の記事同士を比較し補完を行っているのに対し我々は単一の記事と情報の粒度の異なりから複数の記事と比較し情報補完を行っている点異なる。Kumaran ら [15] は Wikipedia 記事の翻訳を支援する WikiBABEL を提案している。具体的には一番記事が豊富である英語版 Wikipedia、その中のある記事をユーザが利用している言語に翻訳し新たなコンテンツ作成を支援することを行う。

3. 補完情報抽出手法

ここではこれまで我々が提案してきた補完情報抽出手法の流

れ[1]について簡単に述べる．

まずユーザは閲覧したい記事のタイトルと閲覧したい言語，比較対象とする言語を入力しそれぞれに対応した言語の記事を抽出する．

ここで，多言語 Wikipedia の情報の粒度の問題から，ある言語版の 1 記事に対する情報が他の言語版では複数の記事にまたがっている場合がある．その為本研究では，比較対象言語の記事のリンク構造の解析を行い，そして我々の提案する関連度を用いて複数の比較対象記事群を抽出する．

しかし，リンク構造と関連度に基づいて抽出した比較対象記事の中には閲覧記事と関係のない情報が含まれている記事が存在する場合がある．そこで次に我々は抽出された比較対象記事に対し比較対象となる領域の決定を行う．その際に我々は比較対象記事を (1) 閲覧記事と同じタイトルを持つ記事を比較基準記事，(2) 比較基準記事と包含関係になっている記事として包含関係記事，(3) 比較基準記事の一部が比較基準記事と関係する記事として部分一致記事の 3 つに分類する．

次に分類した各々の記事ごとに補完情報を抽出するために，閲覧記事と比較する領域の決定を以下の様に行う．

- 比較基準記事

比較基準記事はタイトルが閲覧記事と同じ為，この比較基準記事内の情報は閲覧記事と関係が強いと考え，記事全体を比較対象とする．

- 包含関係記事

包含関係記事は比較基準記事の部分集合であるため，比較基準記事の比較領域と同様に記事全体を比較対象とする．

- 部分一致記事

部分一致記事は，記事内の一部の情報が比較基準記事と関係が強いと考え，比較基準記事へのアンカー文字列のある位置に注目して以下の様に比較対象領域を決定する．

- セクションまたは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

- (1) セクションのタイトルにアンカー文字列がある場合
サブセクションも含めそのセクション全体を比較領域とする．

- (2) サブセクションのタイトルにアンカー文字列がある場合
親セクションとそのサブセクションを比較対象領域とする．ここでは，そのサブセクションの兄弟のサブ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．

- 本文中にのみ比較基準記事のアンカー文字列を含む場合
そのアンカー文字列が含まれている段落のみ比較基準記事と関係が強いと考え，その段落のみを比較対象領域とする．

次にこれらの比較対象の領域と閲覧記事を比較し補完情報の抽出を行う．まず閲覧記事と比較対象記事群をセグメントという単位に分割する．尚，セグメントとは Wikipedia の記事をセクション，サブセクションなどの最小の単位に分けたものである．各セグメントごとに名詞の抽出を行い，その出現頻度を求める．また，比較対象記事の名詞は閲覧言語に翻訳を行う．以下に比較対象記事のタイプ毎の補完情報抽出方法について述べる．

- 比較基準記事と包含関係記事の場合

閲覧記事と比較対象記事のセグメントごとの類似度を求める．

- 部分一致記事の場合

上記で求めた比較対象領域に当たるセグメントと閲覧記事の全てのセグメントとの類似度を求める．

ある比較対象記事のセグメントが閲覧記事の全てのセグメントに対し類似度がある閾値以下である場合，そのセグメントを補完情報として抽出する．

4. 補完情報提示手法

補完情報にも様々な種類が存在する．例えば，ある章に対して追加すべき情報や全く新しい情報等が存在する．そこで本研究では新たにある章に対して追加すべき情報を追加情報，全く新しい情報を新規情報と分類し，それぞれの補完情報に対し情報の提示手法を提案する．その際に本研究では補完情報抽出手法によって得られた補完情報と閲覧記事との関連性を求め，その関連性に基づきユーザの閲覧している記事の適した場所に補完情報を提示する．この時，Wikipedia 記事の章（以下，セクション）や節（以下，サブセクション）からなる目次構造と目次のタイトル，そして類似性に注目し関連性を求める．以下に提示する場所の決定手法を示す．

- 追加情報

以下 3 つ手順のいずれかで補完できる情報は追加情報として補完を行い提示をする．

- (1) 補完情報が比較対象記事のセクションであり，且つそのセクションのタイトルと同じタイトルを持つセクションが閲覧記事に存在する場合

同じタイトルであるので関係性が高いと考え，閲覧記事のそのセクション内に補完情報のアンカー文字列を提示する．例えば，閲覧記事を英語版 Wikipedia の「Kendo」とし，比較対象記事を日本語版 Wikipedia の「剣道」とした場合，抽出された補完情報が「歴史」というセクションの時，同じタイトルのセクションである英語版 Kendo の History のセクション内に「歴史」に対するアンカー文字列を挿入しユーザに提示する．

- (2) (1) で補完できないかつ補完情報が比較対象記事のサブセクションの場合

サブセクションの情報は親セクションを詳細に記載していると考え，そのサブセクションの親セクションを確認し，もしその親セクションと同じタイトルをもつセクションまたはサブセクションが閲覧記事に存在する場合，その親セクションまたはサブセクション内に補完情報のアンカー文字列を提示する．例えば，英語版 Wikipedia と日本語版 Wikipedia の「剣道」を比較した場合では「歴史」のサブセクションである「江戸時代」が補完情報として抽出された場合，その親セクション「歴史」と同じタイトルを持つ「History」のセクション内に補完情報である「江戸時代」のアンカー文字列を提示する．

- (3) それ以外の場合

補完情報がセクションやサブセクションの一部分である場合は，補完情報と閲覧記事との類似している場所にリンクを張るのが良いと考え，補完情報と閲覧記事との類似度を測る．ここで閲覧記事をセグメントに分割し，閲覧記事のセグメントと抽出された補完情報との類似度を以下の式を用いて求める．

$$\cos(x, y) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}} \quad (1)$$

上記の式 x は閲覧記事における 1 つのセグメントの名詞の出現頻度ベクトルであり、 y は抽出された補完情報の名詞の出現頻度ベクトルを指す。 x_i は閲覧記事における 1 つのセグメント内に存在する名詞 i の出現頻度であり、 y_i は補完情報を閲覧言語に翻訳した名詞 i の出現頻度である。

全てのセグメントに対し類似度がある閾値 δ 以上の場合、その中で最も類似度が高いセグメントに情報を補完し提示を行う。

● 新規情報

上記の (3) の手順の際に類似度がある閾値 δ 未満である場合、どれにも類似している情報がないと考え、閲覧記事に存在しない情報であると判断し新規情報として閲覧記事の関連項目に当たる目次の手前にその情報を補完し提示を行う。

5. プロトタイプと評価実験

5.1 プロトタイプシステム

提案手法を用いてプロトタイプシステムを作成した。作成したプロトタイプシステムは閲覧言語を英語版、比較対象言語を日本語版としている。プロトタイプシステムでは日本語の形態素解析として MeCab を、英語の Tagger として TreeTagger を使用し、翻訳 API として Microsoft Translate API を、データベースに MySQL を、開発言語に PHP を用いて作成した。以下にシステムフローを示す。

(1) ユーザは閲覧したい記事のタイトルと、閲覧言語、比較対象言語を入力する。

(2) システムは閲覧言語の Wikipedia から入力されたタイトルの記事を取得する。

(3) (1) で入力したタイトルを持つ記事を比較対象言語の Wikipedia の記事から取得する。

(4) (3) で取得した記事のリンク構造を解析し、比較対象の記事群を比較記事の Wikipedia から抽出する。

(5) (4) で取得した比較対象記事に対し記事の分類を行い、その分類ごとに比較対象となる領域を決定する。

(6) (2) で取得した閲覧記事と (5) で選択した比較対象記事の各比較対象領域を比較しその差分情報を抽出する。抽出した差分情報を補完情報とする。

(7) (6) で取得した補完情報と閲覧記事を比較し、閲覧記事の適した場所に補完情報へのリンクを張りユーザに提示を行う。

プロトタイプシステムのアウトプット画面を図 1、図 2 に示す。図 1 の画面左側が閲覧記事、右側が比較対象記事である。比較対象記事の右上に存在するタブは比較対象記事であり、ユーザがタブを選択することで他の比較対象記事の補完情報を表示できる。また閲覧記事のセクションまたはサブセクションの右上に存在するリンクは補完情報へのリンクである。このリンクをクリックすることで画面右側の比較対象記事が自動でそのセクションまたはサブセクションに対応した補完情報を提示する。比較対象記事において対応した補完情報がある場合が赤い文字で表示し、全く新しい情報であれば緑の文字で表示をする (図

表 1 実験

No.	クエリ	データセット		比較対象領域		
		A ^(注1)	B ^(注2)	適合率	再現率	F 値
1	My Neighbor Totoro	28	1	1	1	1
2	Doraemon	34	2	0.38	1	0.55
3	Iaido	7	2	1	1	1
4	Manzai	24	2	0.5	0.5	0.5
5	Yukata	6	4	0.7	0.8	0.75
6	Urashima Taro	10	5	1	0.67	0.8
7	Pikachu	12	3	0.63	1	0.77
8	Kinkaku-ji	18	4	0.71	0.83	0.76
9	Hello Kitty	30	8	0.68	0.81	0.74
10	Kyudo	24	4	0.71	0.5	0.59
平均	-	-	-	0.73	0.81	0.74

2)。

5.2 比較対象領域の実験

部分一致記事の比較対象領域の決定手法の有用性を示す為、実験を行った。本実験では、ある比較基準記事に双方向リンクされている記事すべてを対象としその記事が比較基準記事との部分一致記事かどうかの判定を行った。使用したデータセットとその結果を表 1 に示す。結果の良かったものとして居合道や弓道などの武道が挙げられる。比較対象領域として書く武道における称号の範士や教士が比較対象領域とすることができた。また、結果の悪かったものとして閲覧記事の比較対象と成り得ない領域を比較対象領域と認識する場合が存在した。例えば Doraemon (ドラえもん) の場合、比較対象記事として作者である藤子・F・不二雄が部分一致記事として抽出された。そして藤子・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した。しかし、彼はドラえもん以外にも多くの漫画を描いており、いくつかのセクションはドラえもんだけでなく他の漫画の話を含む情報が存在し、それが比較対象の領域となる場合があった。これは今後の課題である。

5.3 情報補完の実験

ここでは上記の閾値 δ の設定と情報補完の精度を測るための実験について述べる。3. 章で述べた補完情報抽出手法により抽出された補完情報を用いて最適な閾値 δ を求めた。今回、閲覧言語に英語版を、比較対象言語に日本語版を用いて実験を行った。その際に表 2 に示す 12 個のクエリを用い、閾値を 0 から 0.2 の範囲を 0.05 刻みにし、各々の平均適合率を比較した。結果を図 3 に示す。図 3 からわかるように $\delta=0.07$ の時に最も高い適合率を得ることができた。このことから本研究では情報補完する際の閾値 δ を 0.07 とする。また、その時の各クエリの適合率を表 2 に示す。

結果と考察

結果より平均適合率が 0.65 を得られた。結果の良かった例として Kyudo (弓道) があげられる。Kyudo の補完情報として「弓道場」が Kyudo の道場について説明している場所に補完する

(注1): 双方向リンクされている記事の総数

(注2): 部分一致記事の数

Kendo

Kendo

Kendo, meaning *Way of The Sword*, is a modern [Japanese](#) martial art of sword-fighting based on traditional [swordsmanhip \(kenjutsu\)](#) which originated with the [samurai](#) class of feudal Japan.^{[1][2]}

Kendo is a physically and mentally challenging activity that combines [martial arts](#) practices and values with sport-like strenuous physical activity.^[3]

History

[Takasugi Shinsaku Late Edo period](#)
[Kendo practitioner](#)
[Takasugi Shinsaku Late Edo period](#)
Kendo practitioner.

Since the earliest [samurai](#) government in Japan, during the [Kamakura period](#) (1185–1333), sword fighting together with [horse riding](#) and [archery](#), were the main martial pursuits of the [military](#) clans^[Citation needed]. In this period kendo developed under the strong influence of [Zen Buddhism](#)^[Citation needed]. The samurai could equate the disregard for his own life in the heat of battle, which was considered necessary for victory in individual combat, to the Buddhist concept of the illusory nature of the distinction between life and death^[Citation needed].

Those swordsmen established schools of *kenjutsu* (the ancestor of kendo) which continued for centuries and which form the basis of kendo practice today.^[2]

閲覧記事(英語版)

剣道

タブ

剣道 範士 教士 撓競技 日本剣道形

剣道(けんどう)は、[剣術](#)の[竹刀稽古](#)([撃剣](#))を競技化した[武道](#)。

歴史

江戸時代

剣道の直接の起源は防具と竹刀を使用する打ち込み稽古である。[江戸時代](#)中期の正徳年間(1711年 - 1715年)に直心影流の長沼国郷が面・小手を製作し、竹刀による打ち込み稽古法を確立した。宝暦年間(1751年 - 1763年)に中西派一刀流の中西子武が鉄面・具足式に改良した。防具の発達にともない袋竹刀より強固な割竹刀が作られるようになった。

江戸時代後期から末期には、竹刀打ち中心の道場が興隆し、流派を超えて試合が行われた。幕末の江戸三大道場といわれる鏡新明智流士学館、北辰一刀流玄武館、神道無念流練兵館や、幕府の設立した講武所が有名である。北辰一刀流の創始者千葉周作は剣術の技を六十八手に分類し、講武所頭取並の男谷信友は竹刀の全長を3尺8寸と定めた。防具を装着して竹刀で打ち合う稽古は「撃剣」ともいわれた。

この時代の試合は審判規則や競技大会はなく、個人や道場によるものであった。また、10本勝負が通例とされていた。

比較対象記事(日本語版)

図 1 プロトタイプシステムの出力画面 1

Competition

[complement:1]

対応した補完情報提示

A scorable [[nihongo]] in a *kendo* competition (*tai-kai*) is defined as an accurate strike or thrust made onto a *datotsu-bui* of the opponent's *kendo-gu* with the *shinai* making contact at its *datotsu-bu*, the competitor displaying high spirits, correct posture and followed by *zanshin*.^[24]

Datotsu-bui or point scoring targets in *kendo* are defined as:^[25]

- *Men-bu*, the top or sides of the head protector (*sho-men* and *sayu-men*).
- *Kote-bu*, a padded area of the right or left wrist protector (*migi-kote* and *hidari-kote*).
- *Do-bu*, an area of the right or left side of the armour that protects the torso (*migi-do* and *hidari-do*).
- *Tsuki-bu*, an area of the head protector in front of the throat (*tsuki-dare*).

Datotsu-bu of the 'shinai' is the forward, or blade side (*jin-bu*) of the top third (*monouchi*) of the *shinai*.^[25]

[[nihongo]], or continuation of awareness, must be present and shown throughout the execution of the strike, and the *kendoka* must be mentally and physically ready to attack again.

In competition, there are usually three [[nihongo]]. Each referee holds a red flag and a white flag in opposing hands. To award a point, a referee raises the flag corresponding to the colour of the ribbon worn by the scoring competitor. Usually at least two referees must agree for a point to be awarded. The match continues until a announcement of the point that has

閲覧記事(英語版)

異種試合

異種試合とは、異なる武道との試合のことである。[高野佐三郎](#)が1920年(大正9年)に著した『[日本剣道教範](#)』には銃剣、槍、鎖鎌との戦い方が解説されている。[昭和天皇](#)試合では銃剣術との試合が行われた。

現在は全日本剣道演武大会など特別な大会で、なぎなたとの試合がエキシビジョン的に行われる程度である。

主な大会

- [全日本剣道選手権大会](#)
- [全日本女子剣道選手権大会](#)
- [世界剣道選手権大会](#)
- [全国警察剣道選手権大会](#)
- [全国警察剣道大会](#)
- [全国教職員剣道大会](#)
- [全日本選抜剣道八段優勝大会](#)(旧 明治村剣道大会)
- [全日本都道府県対抗剣道優勝大会](#)
- [全日本東西対抗剣道大会](#)
- [全日本剣道演武大会](#)(京都大会)
- [国民体育大会剣道競技](#)(国体)
- [全国家庭婦人剣道大会](#)
- [全日本高齢者武道大会](#)

新しい情報

比較対象記事(日本語版)

図 2 プロトタイプシステムの出力画面 2

ことができた。さらに Kyudo には記載されていない情報である「礼法」は Kyudo のすべてのセグメントに対し類似度が0ま

たは0に限りなく近い値となり、新しい情報として See also(関連項目)の手前に補完することができた。また日本で生まれた

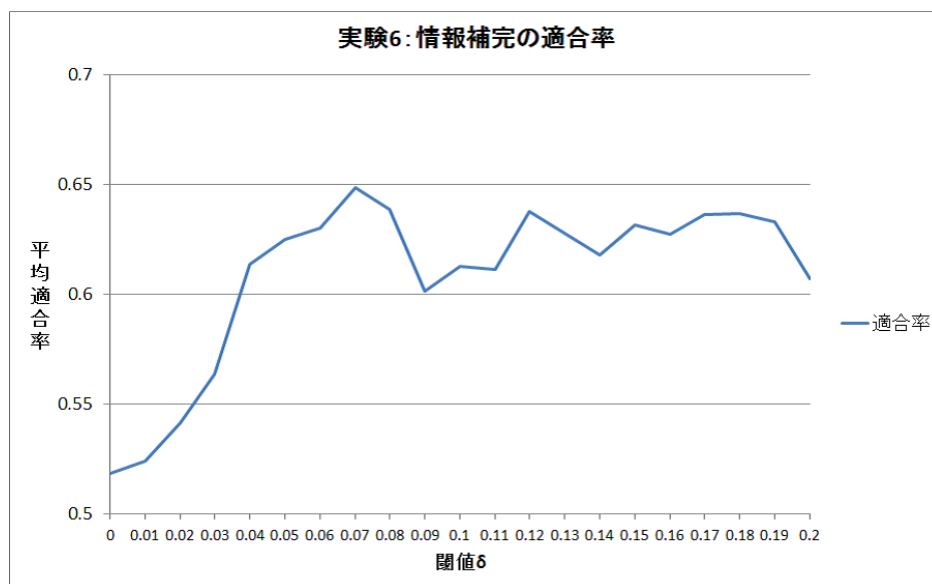


図 3 閾値 δ の設定

キャラクターである Hello Kitty (ハローキティ) の場合ではハローキティが登場するゲームの情報が英語版の「Video games」の項目にすべて補完することができた。

結果の悪いものの一つの補完情報が本来複数の場所に補完されるべきである場合が存在した。例えば、Ninja(忍者)であれば、忍者が使用していたとされる道具についての情報がNinjaの「Weaponry」に補完された。しかし、この情報はWeaponryだけでなくNinjaの「Tool」にも補完されることが望ましい情報でもある。今回、提案した手法では1つの補完情報は1つの補完場所しか持たない。今後はこのような1つの補完情報が複数の補完場所を持つ場合の補完方法とその提示方法について考えていかなければならない。また、おにぎり(Onigiri)であれば比較対象記事としておにぎりの種類の一つであるツナマヨネーズの記事から「概要」が補完情報として抽出され、英語版のOnigiriの概要に当たる「Overview」に補完された。しかし、この概要はおにぎりの概要を説明している部分であるため補完場所としては不適切であると考えられる。このように関連する記事として抽出した比較対象記事の補完情報の場合には別の補完方法を考えるべきであると考えられる。日本の映画の一つであるMy Neighbor Totoro(となりのトトロ)であれば「サウンドトラック」などの音楽の情報が正しく補完できなかった。この情報は本来英語版の「Media」に補完されることが望ましい。しかし「Media」と「音楽」は概念が似ているため今後は単語の概念距離を考慮して情報補完することが望ましいと考えられる。これらは今後の課題である。

6. まとめと今後の課題

本論文ではWikipediaの多言語性に注目し、情報が不足している記事に対し他言語のWikipediaと比較し情報の補完を行う手法を提案した。その際に我々は新たにユーザが閲覧しているコンテンツと抽出した補完情報の関連性に注目し、その関係に基づきユーザの閲覧している記事の適した場所に補完、提示す

表 2 情報補完の実験結果

クエリ	適合率
My Neighbor Totoro (となりのトトロ)	0.53
Ninja (忍者)	0.67
Urashima Taro (浦島太郎)	0.71
Yukata (浴衣)	0.5
Doraemon (ドラえもん)	0.65
Hello Kitty (ハロー・キティ)	0.65
Iaido (居合道)	0.71
Kendo (剣道)	0.83
Kyudo (弓道)	0.75
Manzai (漫才)	0.67
Tempura (天ぷら)	0.61
Onigiri (おにぎり)	0.5
平均	0.65

る方法について提案した。その際に補完情報と閲覧記事のコンテンツとの関連性をもとめ、それにより補完する場所を決定し提示することを行った。

今後の課題を以下に示す。

- 補完情報の分類

補完情報は閲覧記事に存在しない新しい情報や、より詳細な情報。そして視点の違う情報等が考えられる。本論文ではこれらをまとめて補完情報として抽出を行ったが、今後は抽出できた補完情報を分類し補完を行う必要がある。

- 多言語を対象とした実験

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。

謝辞

本研究の一部は甲南大学総合研究所及び JSPS 科研費 23700113, 24500134 の助成によるものである．ここに記して謝意を表す．

文 献

- [1] 藤原裕也, 鈴木優, 小西幸男, 瀬本明代: 言語間比較による Wikipedia の補完情報抽出手法の提案, 第 5 回 Web とデータベースに関するフォーラム (WebDB Forum '12) (2012).
- [2] Strube, M. and Ponzetto, S. P.: WikiRelate! computing semantic relatedness using wikipedia, *Proceedings of the 21st International conference on Artificial intelligence (AAAI'06)*, pp. 1419–1424 (2006).
- [3] Milne, D.: Computing Semantic Relatedness using Wikipedia Link Structure, *Proc. of New Zealand Computer Science Research Student Conference (NZCSRSC'07)*, p. CDROM (2007).
- [4] Takahashi, Y., Ohshima, H., Yamamoto, M., Iwasaki, H., Oyama, S. and Tanaka, K.: Evaluating significance of historical entities based on tempo-spatial impacts analysis using Wikipedia link structure, *Proceedings of the 22nd ACM conference on Hypertext and hypermedia, HT '11*, New York, NY, USA, ACM, pp. 83–92 (2011).
- [5] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *WWW7: Proceedings of the seventh international conference on World Wide Web 7, WWW7*, Vol. 30, Elsevier Science Publishers B. V., pp. 107–117 (1998).
- [6] Kamps, J. and Koolen, M.: Is Wikipedia link structure different?, *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pp. 232–241 (2009).
- [7] Milne, D., Medelyan, O. and Witten, I. H.: Mining Domain-Specific Thesauri from Wikipedia: A Case Study, *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 442–448 (2006).
- [8] Chen, Z., Liu, S., Wenyan, L., Pu, G. and Ma, W.-Y.: Building a web thesaurus from web link structure, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 48–55 (2003).
- [9] Adar, E., Skinner, M. and Weld, D. S.: Information arbitrage across multi-lingual Wikipedia, *Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09*, New York, NY, USA, ACM, pp. 94–103 (2009).
- [10] Ching-man, A. Y., Duh, K. and Nagata, M.: Assisting Wikipedia Users in Cross-Lingual Editing, *WebDB Forum 2010*, pp. 11–12 (2010).
- [11] 吉野孝, 宮部真衣: 日本語版中国語版 Wikipedia における相互言及の有無を用いた文化差検出方法の提案, 第 5 回 Web とデータベースに関するフォーラム (WebDB Forum '12) (2012).
- [12] Qiang, M., Nadamoto, A. and Tanaka, K.: Complementary information retrieval for cross-media news content, *Elsevier ARTICLE Information Systems Volume 31, Issue 7, November 2006*, pp. 659–678 (2006).
- [13] Eklou, D., Asano, Y. and Yoshikawa, M.: How the web can help Wikipedia: a study on information complementation of Wikipedia by the web, *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2012)* (2012).
- [14] Monz, C., Nastase, V., Negri, M., Fahrni, A., Mehdad, Y. and Strube, M.: CoSyne: a framework for multilingual content synchronization of wikis, *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym '11)*, pp. 217–218 (2011).
- [15] Kumaran, A., Ashok, B., Saravanan, K., Ande, A., Sharma, A., Vedantham, S., Natampally, V., Dendi, V. and Maurice,

S.: WikiBABEL: A System for Multilingual Wikipedia Content, *Proceedings of the 'Collaborative Translation: technology, crowdsourcing, and the translator perspective' Workshop (co-located with AMTA 2010 Conference)* (2010).