Wikipedia 記事の言語間差異抽出手法の提案

原田 一慧[1]　木村 文則[2]　前田 亮[3]

[1]立命館大学情報理工学研究科　〒525-8577 滋賀県草津市野路東 1-1-1
[2]立命館大学衣笠総合研究機構　〒603-8577 京都府京都市北区等持院北町 56-1
[3]立命館大学情報理工学部　〒525-8577 滋賀県草津市野路東 1-1-1
E-mail: is0034hf@ed.ritsumei.ac.jp
fkimura@is.ritsumei.ac.jp, amaeda@media.ritsumei.ac.jp

あらまし　本論文では，Wikipedia の日本語の記事と英語の記事間において，記述の差異を自動で抽出する手法を提案する．Wikipedia はインターネット百科事典であり，世界中の人が自由に編集することが可能である．そのため，同じ対象について述べている記事であっても，書かれている言語によって，内容に差異が生じることがある．そこで，ある言語の記事に対し，別言語の記事にしか存在しない記述や，より詳細に述べられている部分を補完することで，付加情報を得ることが可能になる．そこで本研究では，日英の記事を対象とし，それらの同一記事間から差異部分を取り出す手法を提案する．既存手法と同じテスト記事を用いて比較実験を行った結果，既存手法に比べ，文単位で差異部分を抽出する精度の向上に成功した．
キーワード　Wikipedia，多言語，複数文書要約

1. はじめに

インターネットを簡単に利用する事ができる現代において，オンライン百科事典は貴重な情報源の一つであり，多くのユーザが物事を調べる際に利用している．知りたい事柄について，インターネット上で検索することで，多くの記事が表示され，そこから目的の情報を得ることが可能である．Wikipedia はそのようにインターネット上で公開されているオンライン百科事典の一つである．Wikipedia は誰でも編集が可能であるという特徴を持ち，また世界中の人たちによって記事が作成されている．そのように共同作業によって，記事が創りあげられることを目的としたプロジェクトであり，その成果が Wikipedia という百科事典本体となる．

ユーザは Wikipedia を用いることで多くの情報を得ることができるが，問題点が存在する．Wikipedia には多くの言語の記事が存在し，記事の総数は言語によって異なる．また同じ事柄について述べている記事であっても，言語が異なることでその記述内容に差異が生じる場合がある．それは Wikipedia がフリーなオンライン百科事典であるという特徴から生じる問題である．言語によって記事を編集している著者が異なる場合があり，その際記述に違いが生じる．一方の言語の記事では述べられている内容が，別の言語の記事では一切書かれていないことや，一方の記事よりさらに詳細な記述が別の言語の記事に存在する，といった記事が存在する．また，言語によって言及されている内容が正反対のことを述べている可能性も考えられる．Wikipedia はフリーなオンライン百科事典であるため，言語によって記述されている内容に違いがあるかどうかを事前に判断することはできない．ある言語の記事を基準にして，その内容を別の言語の記事でも記述する，といったルールが存在しないためである．この点から，ユーザがある事柄に関する知識を得ようとする際，ある一言語の記事を参照するだけでは目的の情報が得られない可能性が出てくる．しかし，多くのユーザにとって，自らの母国語以外の記事を読むことは非常に大きな負担となる．仮に母国語の記事を参照した後に，別の言語の記事から差異を記述した部分を得ようとした際，別言語の記事のどこに差異が記述されているか事前に判断できないため，別言語の記事を全て翻訳する必要がある．そこで，あらかじめ 2 つの言語の記事内容を比較し，別言語の記事に存在する差異を抽出することで，ユーザの情報収集における負担を軽減することが本研究の目的である．

本研究では複数文書要約の手法を応用し，記事間の文書長を考慮した差異抽出手法を提案する．抽出する差異情報は，ユーザの母国語記事と別言語の記事を比較した際に，別言語の記事に記述されている，

1. 別言語の記事にしか書かれていない内容
2. 別言語の記事でより詳細に述べられている内容

これら 2 つを指すこととする．また，差異情報は別言語の記事から文単位で抽出する．差異部分を文単位で抽出することで，母国語の記事と重複した内容を記述している部分をより有効に削除し，的確に差異情報をユーザに提示することを可能にする．

複数文書要約手法では，比較対象となる 2 つの文章において，それぞれのどちらにも含まれている単語を抽出し，それらの単語をより多く含む文を，同じ内容を記述している文として判定する手法が利用されている．本研究では異なる言語の記事を対象とし，その精度を上げる

ため，抽出されたどちらの文章にも含まれる単語を重要度によってランク付けを行い，スコアを変動させることで既存手法では抽出できなかった文や，誤って抽出されていた文を省くことを可能にする．

本論文では，第 2 章で関連研究の紹介を行い，第 3 章で抽出する差異の定義について述べる．第 4 章では提案手法について述べる．そして，第 5 章では既存手法と提案手法の実験結果の比較について述べ，第 6 章では提案手法および実験結果についての考察を行う．第 7 章では今後の課題について述べる．

2. 関連研究

本章では，本研究と関連した研究についての紹介を行う．複数の文章を比較する研究は古くから行われているが，主に複数の文書に対するテキストマイニングを行っている研究について述べる．

2.1 異言語 Wikipedia を用いた補完情報の提示手法の提案

関連研究として，藤原らによる「異言語 Wikipedia を用いた補完情報の提示手法の提案」[1]を紹介する．この研究では，Wikipedia に存在する文化に関する記事に対して，2 つの言語の記事から差異を抽出する手法を提案している．閲覧者の母国語の記事に対し，比較対象となる別言語の記事を取得し，記事に存在する目次を用いて記事をセクションに分割する．分割されたセクションごとに内容を母国語記事と比較を行い，類似度を計算することで，差異と判定されたものを閲覧者に提示する．ここでは，本研究と関連する差異抽出の部分に注目し，類似度の計算部分について解説を行う．

この研究では，Wikipedia の記事をいくつかのセグメントに分割を行っている．記事の冒頭にある説明部分をサマリとし，それ以下の部分を目次に従ってセグメントとして分割する．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．この研究では，Wikipedia 記事を比較する際にセグメントに注目し，セグメントに基づいた差異情報抽出を行っている．まず，母国語記事と比較言語記事のどちらも形態素解析を行い，名詞のみを抽出する．次に辞書を用いて比較言語記事の名詞を母国語へと翻訳を行う．なお，この研究では翻訳時に単語の多義性など意味の曖昧性については考慮せず，今後の課題としている．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，コサイン類似度を用いてセグメント同士の類似度を求める．コサイン類似度を求める式は(1)を用いている．差異情報の決定については，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て類似度が閾値以下で合った場合に，そのセグメントを差異情報として抽出する手法で行っている．ここでの閾値は実験から 0.3 と決定されている．

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

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．xi 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，yi 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．

この研究では Wikipedia 記事をセグメントに分割し，それを 1 つの単位として差異情報抽出を行っている．そのため名詞の出現頻度を利用したコサイン類似度による類似度の計算を行い，それによって差異の判定をしているが，セグメント単位で抽出された情報が全て差異であるとは限らない．1 つのセグメントは，その種類によって文章の長さや，使われている単語の種類も大きく異なる．そのため，1 つのセグメントを差異情報と判断した場合でも，実際の差異情報がそのセグメント内の僅か 1,2 文程度である可能性も考えられる．そこで本研究では，抽出する差異情報の最小単位をセグメントではなく，1 文単位にすることで，より精度の高い差異抽出を目的としており，その点がこの研究とは異なる．

2.2 要約対象テキスト間の共通点と相違点の抽出

奥村らによるテキスト自動要約手法の論文[2]で，複数文書を自動要約する手法が紹介されている．複数文書要約は，いくつかの文章郡からその文章の内容を要約した文章を自動的に作成する研究であり，代表的な手法として，複数の文書で同じ内容を記述している部分を特定する手法がある．そのため，テキスト間で共通した内容を記述した箇所を同定することで，それ以外の部分が差異情報の候補であると判断できる．本研究でも同様に，2 つの言語の記事間において，共通した内容を記述した箇所を特定することで，差異を抽出している．奥村らは，テキスト間の共通点同定に，形態素レベルの比較手法があると述べられている．形態素レベルの比較では，まず比較する 2 文を形態素解析し，そこから自立語を抽出する．

【例文 1】フーバー社は軽量の携帯電話を発売する
【例文 2】軽量の携帯電話が来月フーバー社から発売される

この 2 文について形態素レベルでの比較を行うとする．2 文を形態素解析すると，例文 1 からは「フーバー社，軽量，携帯電話，発売」，例文 2 からは「軽量，携帯電話，来月，フーバー社，発売」といった単語が抽出される．これらの単語の中でどちらの文にも一致して出現している単語の度合いを測ることで，共通箇所としての同定を行う．例文 1 では 4 語中 4 語が，例文 2 では 5 語中 4 語が互いに一致している．ここから，この 2 つの文は

互いに同じ単語を多く含むため，同様の内容について述べていると考えることができる．このようにして 2 文をテキスト間の共通箇所として同定する．

これに対し，構文解析技術を利用し，構文レベルで比較を行うという手法も存在する[3][4]．この手法では，あらかじめ比較対象の 2 文を構文解析しておき，その解析結果として得られた構文木をいくつかの規則を用いて変形することで，2 文が同内容であるかどうかの判定を行う．そしてその後その 2 文を統合し，1 つの文を生成するという手法である．上記の例文では，例文 2 は例文 1 を受動態にした文となっている．そのため能動態を受動態へと変形させる規則をあらかじめ生成しておくことで，例文 1 と例文 2 が同内容であると同定することができる．また，例文 2 には「来月」という例文 1 に含まれていない単語が存在する．しかし構文解析によって，「来月」という単語が「発売される」に係っていることが分かれば，例文 2 に能動態変形の規則を適用し，「来月」が「発売する」に係るよう変形した上で例文 1 との統合処理を行うことで，「フーバー社は軽量の携帯電話を来月発売する」といった文を生成することが可能である．しかしこの手法は，同じ言語の文同士を比較する上で利用できる手法であり，本研究の目的では適用が難しいと考えられる．

2.3 言語横断に関する Wikipedia の研究

Wikipedia を対象として，その記事の内容を別の記事と比較する研究は他にも多く行われている．Angela[5]らは Wikipedia の英語版の記事から，日本語の記事へリンクを作成するべき単語の抽出と，そのリンク先となる日本語記事を正しく抽出する研究を行っている．英語版の記事を読む際に，ある単語について別の言語の記事へのリンクが付与されていれば，情報をさらに広く取得することが可能となる．Angela らはある英語の記事と，それと同じ対象について述べている日本語の記事を自動的に発見し，ペアとして保存を行っている．同じ対象について述べている記事の発見には incoming リンク（同じ記事から張られているリンク）や，outgoing リンク（同じ記事へ張られているのリンク）を参考にしている．ある記事から 2 つの記事へリンクが繋がれている時，それらの記事は特に関係が強いと考えられる．また，2 つの別の言語記事が同じある記事へのリンクを持っていた場合，それらの記事の関連度も強いと思われる．このように，記事に張られているリンクをたどることによって，記事同士の関連度を計算し，同じ対象について述べている別の言語の記事の発見を行っている．本研究でも同様に Wikipedia の多言語の記事を扱っている．

Maofu[6]らは Wikipedia の記述の特徴を用いることで，中国語の記事と，同じ内容を持つ英語記事を発見する手法を提案している．Wikipedia では，ある言語の記事内で，別の言語の単語が文章内で出現する際，その用語の後ろに別言語での言い回しが書かれることがある．また注釈として記述されている場合も存在する．そのように注釈として記述されている単語を用いることで，文章中の単語が別の言語ではどのような名称であるかということを判断することができる．しかしこの手法は記事を編集した人の知識量に大きく依存すると考えられる．Wikipedia の記事は多くの人の手によって書かれ，編集が繰り返されるため，その情報の密度も記事によって異なる．そのため注釈を利用する手法では，全ての記事や単語について，正しく対応する別の言語の記事を抽出することは難しい．本研究では，編集者に依存せず，機械翻訳を行うことで単語の比較を行っている．

Jungi[7]や Petr[8]らは Wikipedia において，ある英語の記事と，それと同じ対象について述べている日本語の記事を自動的に発見するために，Wikipedia に存在する言語間リンクを利用した手法を用いている．Wikipedia に存在する言語間リンクとは，記事の編集者が作成した，ある記事と同じ対象について述べている別の言語の記事へリンクさせたリンク先一覧である．例えば日本語の「畳」という記事の言語間リンクは図 1 のようになっており，他の言語へのリンクが付与されていることがわかる．これらのリンクは 1 つの言語から複数の言語へと繋がれており，この研究では，その点を利用している．この研究では英語の記事に対して，同じ対象を述べた日本語記事を発見しているが，言語間リンクには直接英語と日本語が繋がれていない記事が存在する．そうした場合に，一度別の言語へとリンクを辿り，そこから日本語記事を発見するという手法を用いている．例えば英語の「La Fayette-class frigate」という記事には日本語への言語間リンクが存在しない．しかしロシア語へのリンクが存在するため，はじめにロシア語の記事を抽出し，その記事に付けられた言語間リンクから日本語記事を発見している．このように言語間リンクは編集者によって作成されているため，同じ対象について述べている記事同士が繋がれている可能性が高い．本研究においても言語間リンクを利用し，英語と日本語で対応した記事を抽出している．

図 1：日本語記事「畳」の言語間リンク

3. 差異情報の定義

ここでは，本研究で抽出する対象となる差異情報についての説明を行う．本研究では日本人のユーザを対象と

している．ユーザが Wikipedia の日本語記事を閲覧する際に，英語記事から差異情報を提示する，という前提で研究を行っている．ユーザが閲覧するのが日本語記事であり，差異情報を抽出する対象が英語記事となる．そこで本研究では差異情報を以下の 2 種類としている．

1. 日本語記事には書かれていなかった情報を含む文
2. 同じ事柄について述べているが，日本語記事よりも詳細に記述されている文

これらは日本語記事を閲覧するだけでは得ることの出来ない情報であり，英語記事から抽出することで，閲覧記事の補完をすることができる．また，記事をいくつかのセクションに分割するのではなく，文単位で情報抽出を行うのも本研究の特徴である．

2 つ目のパターンに該当する差異情報について例を述べる．以下の文 A は日本語の記事「ガントレット（刑罰）」で書かれている 1 文，文 B は同じ対象について述べている英語記事「Running the gauntlet」から抜粋された 1 文である．

A) また士官学校の訓練やしごきとして行われることもあった。
B) Also used in training, notably on military cadets, as in a scene in the movie Oberst Redl.

文 A と文 B はどちらもガントレットと呼ばれる刑罰について，士官学校の訓練やしごきで行われることがあるという内容の記述がされている．しかし，文 B には「as in a scene in the movie Oberst Redl」という記述が追加されており，ガントレットが「Oberst Redl」という映画のシーンで行われていることが述べられている．こうした文は同じ事柄について記述されているが，より詳細な情報が追加されている文であり，2 つ目のパターンに該当する文として分類できる．

差異情報として定義している文は 2 種類であると述べたが，どの程度差異が含まれていれば差異情報として認識するかは，記事を読むユーザによって判断が異なると考えられる．例えば，ある記事 A に「リンゴはバラ科の落葉樹の果実である」という記述があるとする．別の記事 B では「リンゴはバラ科リンゴ属の落葉樹の果実である」と記述されていた場合，記事 A に欠落している情報は「リンゴ属」という一部分のみである．記事 A を読んだユーザが記事 B から差異情報を得たいと考えた時に，「リンゴはバラ科リンゴ属の落葉樹の果実である」という文を差異であると判断するかどうかは，ユーザによって異なるということである．

本研究で実験を行うにあたって，あらかじめ正解となる差異情報を決定する必要がある．そこで本研究の実験における正解は，著者の一人が実際に記事を読み比べ，2 種類の差異情報となる文を抽出した文集合を用いている．その際，上記の例のように，一部分のみ差異情報が含まれている文であっても，差異情報として判断を行った．

4 提案手法

本研究では閲覧記事と比較記事から同内容を記述した文を特定することで，差異情報を抽出する手法を提案する．提案手法の概要図を図 2 に示す．入力として，ユーザが閲覧している日本語記事を用いる．入力された日本語記事を「閲覧記事」，閲覧記事と言語間リンクで繋がれている英語記事を「比較記事」とする．言語間リンクとは Wikipedia に存在するリンクであり，ある言語の記事について，それと同じ対象について述べている他の言語の記事へのリンク一覧である．

本手法は，共通重要語リスト作成と，スコア計算の 2 つのセクションに分かれる．共通重要語リスト作成セクションでは，入力された閲覧記事と，比較記事の双方の記事を形態素解析し，共通重要語リストを作成する．スコア計算セクションでは，共通重要語リスト作成セクションで作成された，閲覧記事の名詞リストと共通重要語リストを使用して，比較記事の 1 文ごとにスコア付けを行う．全ての文のスコア付けを行ったのち，あらかじめ設定した閾値以下のスコアを持つ文を，差異情報として抽出し，ユーザへと提示を行う．各セクションについての詳細な説明を以下の項で行う．

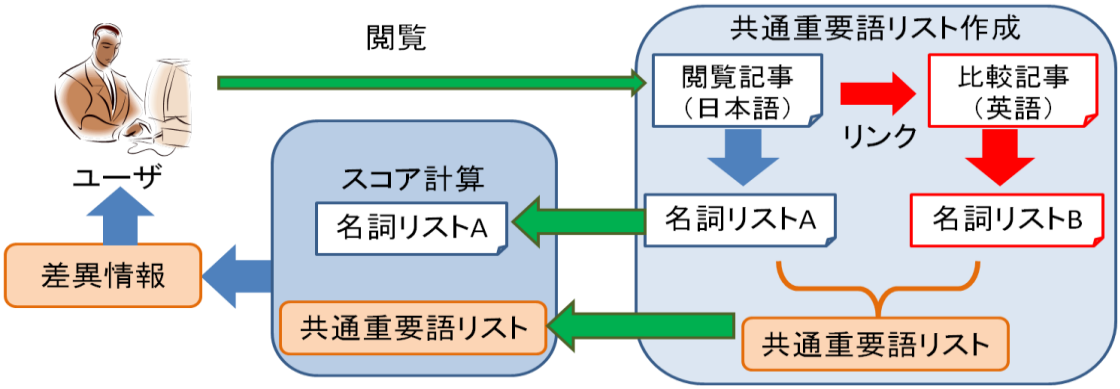

図 2：提案手法の概要図

4.1 共通重要語リスト作成

このセクションでは，閲覧記事と比較記事のどちらにも出現している共通重要語を以下の手順で抽出する．

1. 閲覧記事（日本語記事）を形態素解析器 MeCab[9]を用いて形態素解析し，名詞のみを抽出
2. 閲覧記事から抽出された名詞を機械翻訳器の Google 翻訳を用いて英語へと翻訳
3. 比較記事（英語記事）を多言語用形態素解析器 TreeTagger[10]を用いて品詞解析を行い，名詞のみを抽出
4. 手順 1 および 3 で抽出された名詞を TF-IDF 法を用いてランク付けを行う
5. ランク付けされた名詞の上位の単語を，閲覧記事，比較記事それぞれで抽出する．この名詞を「重要語リスト」と呼ぶ
6. 閲覧記事と比較記事それぞれの重要語リストを比較し，共通する単語を抽出．これを「共通重要語」と呼ぶ

このセクションでははじめに記事を形態素解析し，名詞を抽出するが，2 つの記事でどちらでも重要度の高い単語の抽出を行う．TF-IDF による重要語のランク付けは，閲覧記事，比較記事それぞれで行う．IDF の計算に用いる記事群は，閲覧記事であれば Wikipedia の日本語記事全て，比較記事であれば Wikipedia の英語記事全てである．

閲覧記事から得られた重要語リストは，閲覧記事から抽出された名詞の中で TF-IDF 値の高かった上位の単語である．比較記事から得られた重要語リストも同様に，比較記事から得られた名詞の TF-IDF 値が高かった上位の単語リストである．本手法では上位の単語を決まった割合で抽出し，重要語リストとしている．割合とは，閲覧記事および比較記事それぞれの名詞で，TF-IDF 値によるランク付けを行い，その順位で上位の数%の単語を選ぶ，ということである．この割合については，最適な数値を実験の項で決定している．

閲覧記事と比較記事のそれぞれから得られた重要語リストの中で，共通している単語を抽出したものが共通重要語リストとなる．次のスコア計算セクションでは，本セクションで作成した，閲覧記事の重要語リストと共通重要語リストを使用する．

4.2 スコア計算

ここでは比較記事から差異を抽出するためのスコア計算を行う．本研究では文単位で差異抽出を行うため，英語記事の文ごとにスコアの計算を行う．文ごとのスコア計算は以下の手順で行う．

1. 共通重要語の単語がその文に含まれていれば，その個数に応じて文に重み A を加算
2. 閲覧記事の重要語リストの単語がその文に含まれていれば，その個数に応じて文に重み B を加算
3. 加算された重みをその文の長さ（単語数）で除算する．この値をその文のスコアとする
4. スコアがあらかじめ設定した閾値を下回っていた場合，その文を差異情報として抽出する

共通重要語リストに含まれる単語はどちらの記事でも重要であり，かつ共通している単語である．共通重要語を含む文は，どちらの記事でも同じ内容を記述している可能性が高いと仮定できるため重みを大きく加算する．ここで加算する大きい重みを「重み A」とする．また，日本語記事から得られた重要語リストの単語を含む文であれば，同内容を記述している可能性が高いと考えられる．しかし，重要度が高くない単語であるため，重みの加算は小さくする．ここで加算する小さい重みを「重み B」とする．重み A，重み B の数値に関しては，5 章の実験でいくつかのパターンで検証を行って決定している．

3 番目のステップでは文の長さによって正規化している．長い文であれば，使用されている単語の数も多くなるため，重みが過剰に加算されるおそれがある．そのため全ての文で偏りのないスコア計算を行えるように，文の単語数で重みを除算している．これにより，長さの大きく異なる文であっても同様にスコア付けを行うことができる．

スコアが高い文は閲覧記事の内容と同内容を記述している可能性が高いと考えられる．そのため閾値を下回るスコアを持つ文を差異として判定を行う．閾値については 5 章で述べる．

5. 実験

提案手法の有効性を確かめるため，Wikipedia から抽出したテスト記事 10 記事（表 1）を用いて実験を行った．3 章の最後で述べたように，あらかじめ日本語版のテスト記事とそれぞれの記事と対応する英語記事を用意し正解を作成した．

表 1：実験で用いたテスト記事一覧

日本語記事	対応する英語記事
オーデル川	Oder
オガワコマッコウ	Dwarf sperm whale
ヨルムンガンド	Jörmungandr
赤外線天文学	Infrared astronomy
ニッポノサウルス	Nipponosaurus
畳	Tatami
ガントレット（刑罰）	Running the gauntlet
ショーテル	Shotel
ヘンゼルとグレーテル	Hansel and Gretel
ラビオリ	Ravioli

6.1 ベースライン

今回実験を行って，提案手法の有効性を確かめるために，比較対象としてベースラインとなる手法による実験を行った．

既存手法として，文章の内容がどれほど似ているかを類似度として計算する手法や，複数文書要約と呼ばれる，複数の文書を自動的に要約するために，同じ内容を記述した部分を抽出するといった手法がある．類似度計算は，ベクトル空間モデルを用いて，文書同士を比較する手法が一般的に用いられる．しかし，本研究の対象は記事と文の比較であるため，2 つの文書で使用されている単語数や，文章長が極端に異なる．そのため，ベクトル空間モデルを用いた計算では数値が適切な値が得られない可能性がある．また，2.2 節で述べたような複数文書要約手法を利用することで，2 つの異なる言語の記事から差異を抽出するといった研究はこれまで行われていない．

ベースラインとしては，2.2 節の文献[2]で述べられている複数文書要約に用いられる手法で実験を行う． ベースライン手法は以下の手順で行う．

1. 日本語版のテスト記事それぞれを形態素解析器 MeCab を用いて形態素解析を行い，名詞を抽出
2. 抽出された名詞を機械翻訳器である Google 翻訳 を用いて英語へと翻訳
3. 英語版のテスト記事をそれぞれ多言語用品詞解析器である TreeTagger を用いて品詞解析を行い，名詞を抽出
4. 日本語記事から抽出された名詞群と英語記事から抽出された名詞群を比較し，共通している名詞を「共通出現語 」として抽出
5. 英語記事の文ごとに，共通出現語を含む割合を計算し，全ての文の値の平均値で，各文の値を割った数値をその文のスコアとする
6. スコアがあらかじめ設定した閾値以下の値をもつ文を差異として抽出

この手法では，日本語記事と英語記事で同じ内容を記述している部分を特定することで，差異を見つけ出すことを目的としている．ベースラインでは日本語，英語記事どちらにも出現している単語全てを用いて，文ごとにスコア付けを行う．どちらの記事でも使用されている単語を多く含む文であれば，それらは 2 つの記事で同様の内容を記述している可能性が高いと仮定できる．逆に，共通出現語を含む割合の低い文であれば，その内容に共通点が無いと考えられるため，差異候補となる．

6.2 実験条件

提案手法では，3 つの値を決定する必要がある．1 つ目は 4.1 節で述べたように，閲覧記事と比較記事双方で TF-IDF 値によってランク付けされた名詞リストの上位何%を重要語とするかである．今回は上位 10%，20%，30%の 3 パターンで比較を行う．

2 つ目は 4.2 節で述べた文のスコア計算セクションにおいて，各文に加算する「重み A」と「重み B」の値である．本研究では重み B を 0.1 と固定し，重み A を 0.1，0.3，0.5，0.7 の 4 パターンに値を変更し，結果の比較を行う．

3 つ目は，スコア計算を終えた段階で，どの文が差異情報であるか判定するための閾値である．閾値は 0.02 から 0.01 刻みで 0.09 まで変化させて実験を行った．

6.3 実験結果

実験結果をグラフにして比較を行った．ベースラインの結果が図 2，提案手法での結果が図 3 から図 6 である．

10 個のテスト記事はあらかじめ正解として抽出する文を用意してあるため，それらをどれだけ正確に抽出できたかを測ることができる．手法によって抽出された文と正解を比較し，適合率と再現率から F 値を計算する．図 3 ではベースラインにおける F 値を，スコアの閾値を変化させて，グラフ化している．また，提案手法に関しては，重み A，B の値の組み合わせごとに閾値を変更して実験した．重要語を選定する割合に関しては，図 4 が上位 10%，図 5 が上位 20%，図 6 が上位 30%の割合で行った結果を表している．

ベースラインでは閾値 1.2 での F 値 51.0%が最大であったのに対し，提案手法では上位 20%の名詞を重要語と決定した際の，重み A を 0.7，閾値 0.06 と設定した場合の F 値 55.6%が最大である．比較実験を行った結果，F 値を約 5.6%向上させることに成功した．

6.考察

テスト記事を用いた実験で，ベースラインに比べ，提案手法では F 値の平均が向上した．

ベースラインについて，共通出現語が多くの文に分散して含まれているため，各文ごとのスコアにあまり大きな違いが得られなかったのではないかと考えられる．一般的な単語であっても割合の計算に用いられるため，本研究の目的である差異情報を記述した文を特定できなかったと考えられる．

例えば，ベースラインではテスト記事の 1 つである「赤外線天文学」という記事において，スコア計算するために用いた単語の中に「time」や「case」，「way」などが含まれていた．これらの単語は多くの記事で一般的に用いられる単語であると思われる．「赤外線天文学」での実験では「case」を含んだ文として，「As is the case for visible light telescopes, space is the ideal place for infrared telescopes.」という文が抽出された．しかしこの文の内容は日本語記事で既に書かれており，抽出されたのは誤りとなる．

一方，提案手法ではスコア計算で大きい重み A を付加するための共通重要語として「telescope」や「infrared」等が抽出された．これらの単語によって上記の文のスコアが高く設定され，閾値を超えたことで同内容の文として正しく判定されたと思われる．ベースラインではこのように一般的な単語を使用してしまうため，多くの文でスコアが似通った数値になり，差異情報を判定するのが難しくなったと考えられる．

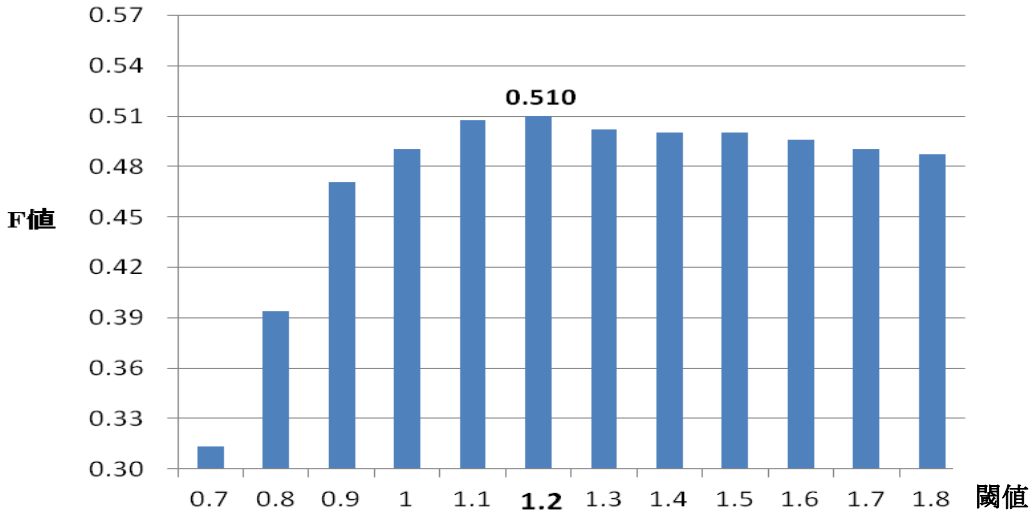

図 3：ベースライン手法を用いて閾値を変化させた F 値の結果

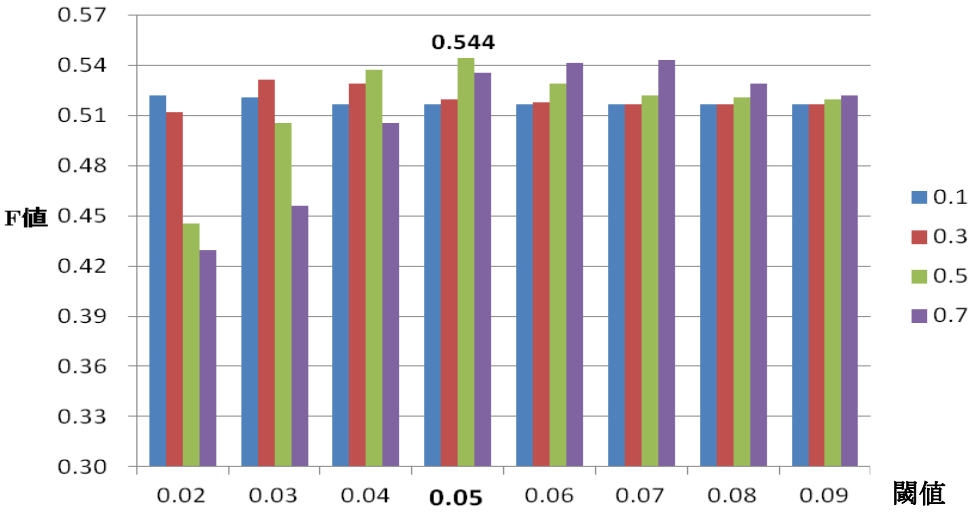

図 4：提案手法（上位 10%を重要語とした場合）で閾値を変化させた結果

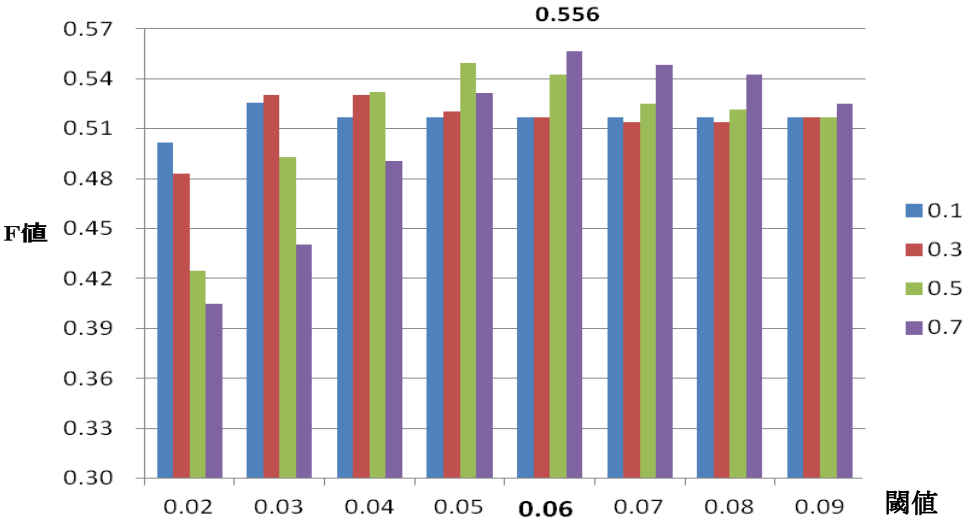

図 5：提案手法（上位 20%を重要語とした場合）で閾値を変化させた結果

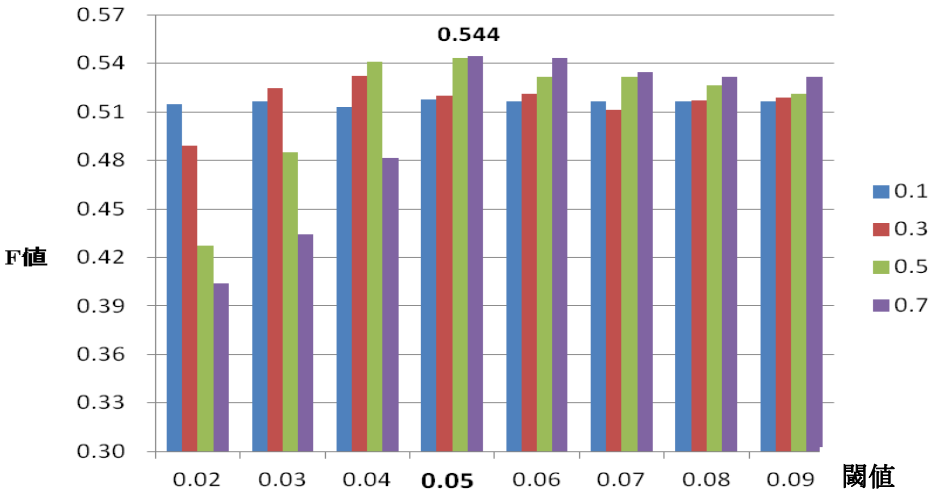

図 6：提案手法（上位 30%を重要語とした場合）で閾値を変化させた結果

提案手法については，上記のように「telescope」や「infrared」等，共通出現語の中でも特に重要な単語に絞って重みを大きく設定したことで，どちらの記事でも同じ内容が書かれている部分を多く取り出せたと考えられる．また，単語を限定したことで取り逃してしまう文についても，日本語記事で出現した重要語をスコア計算に利用することで，精度を向上させることができたと考えられる．また，Wikipedia は言語ごとに記述を行った著者が異なる場合があるため，日本語記事では 1 文で書かれている内容が英語記事では 2 つの文に分割されて記述されているという場合も考えられる．このような場合でも文ごとの単語数でスコアを調整したため，正しく抽出できたのではないかと考えられる．

7.まとめと今後の課題

本論文ではオンライン百科事典である Wikipedia において，閲覧している記事の内容を補完する差異情報を，同じ対象について述べている別の言語の記事から自動的に抽出する手法を提案し，実験によって，既存手法を用いた場合に比べ，その結果を向上させることができることを示した．

今後の課題として，今回の実験では，テスト記事での正解を「差異を含む文」と「同内容の文」といった区分で作成した．しかし，１つの文の中に閲覧記事と同じ内容と，差異が混在している場合がある．今後は，正解を「差異を含む文」と「部分的に差異を含む文」，「同内容の文」という区分で作成することで，より閲覧者に利用しやすい手法を目指すことなどが考えられる．

抽出された差異をその内容に応じて，閲覧記事の最も関連の高い部分に提示するということも改善案として考えられる．たとえば歴史に関する文であれば，閲覧記事の歴史部分へ，概要であれば概要部分へ提示することができれば，よりユーザの利用しやすいシステムになると考えられる．

参考文献

[1] 藤原裕也，鈴木優，小西幸男，灘本明代．異言語 Wikipedia を用いた補完情報の提示手法の提案．DEIM 2013 第 5 回データ工学と情報マネジメントに関するフォーラム

[2] 難波英嗣，奥村学．ここまで来たテキスト自動要約．情報処理 Vol.43 No.12, 2002.

[3] Barzilay. R, McKeown, K. R. Elhadad, N: Information Fusionin the Context of Multi-Document Summarization, Proceedings of the37th Annual Meeting of the Association for Computational Linguistics, pp.550-557(1999).

[4] 上田良寛，小山剛弘:共通意味断片の抽出による複数文書要約:言語処理学会第 6 回年次大会，pp.360-363(2000)

[5] Angela Fahrni, Vivi Nastase, Michael Strube:HITS' Graph-based System at theNTCIR-9 Cross-lingual Link Discovery Task: Proceedings of NTCIR-9 Workshop Meeting, December 6-9, 2011,

[6] Maofu Liu, Le Kang, Shuang Yang, Hong Zhang: WUST EN-CS Crosslink System at NTCIR-9 CLLD Task: Proceedings of NTCIR-9 Workshop Meeting, December 6-9, 2011,

[7] Jungi Kim and, Iryna Gurevych: UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery: Proceedings of NTCIR-9 Workshop Meeting, December 6-9, 2011,

[8] Petr Knoth, Lukas Zilka, Zdenek Zdrahal: KMI, The Open University at NTCIR-9 CrossLink:

[9] Cross-Lingual Link Discovery in Wikipedia Using Explicit

[10] Semantic Analysis:Proceedings of NTCIR-9 Workshop Meeting, December 6-9, 2011,

[11] mecab - Japanese morphological analyzer. (オンライン) https://code.google.com/p/mecab/.

[12] TreeTagger-a language independent part-of-speech tagger. (オンライン) http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.