修　　士　　論　　文

言語横断的なWebニュース記事の関連付け

Cross-Lingual Linking of Web News Articles

指導教員　**田浦 健次朗 准教授**

東京大学情報理工学系研究科
電子情報学専攻

氏 名　48-66446 **吉田 慎一郎**

提 出 日　**平成 20 年 2 月 4 日**

概 要

インターネットの普及と情報通信技術の発展に伴い、今では個人でも世界中の Web ニュースサイトにアクセスして世界中で報道されている Web ニュースを手に入れられることができるようになった。そのため、例えばある事件や事故についての報道がされているニュースを単純に集めてきたり、そこからニュースサイトごとに違った見方で報道されているのではないか、といったようにあるニュースを多角的に調べる、といったことが可能になってきている。それらの調査に、世界中のニュースを集めてきて、それらを体系的に事件や事故、トピックといったものでニュースを関連付けしておいて、必要なときに必要なニュースだけを取り出したいという需要が生じてきている。しかし、世界中のニュースを扱って処理するには、世界で使われている言語についての文法や単語の意味といった言語情報が必要となってくる。それらの情報を言語ごとに集めてきて使用するには大変手間がかかるものである。そこで、本論文では、Web ニュースを関連付ける手掛かりとして、まずは複数言語でニュースを提供しているサイトを対象とし、そのサイトにおいてニュースカテゴリとニュース記事を抽出するラッパーの作成とニュースカテゴリの対応からカテゴリごとにニュース記事を対応付けさせ、ニュース記事の対応付けには、言語情報をほとんど用いないことを前提として、単純なスペースか n-gram による単語区切りと、オンライン百科事典として有名であり誰でも使える Wikipedia を用いた単語判定と他言語への翻訳を用いた手法を提案する。本研究の手法では、使用した言語情報は Web ページの言語がどの言語であるかということと単語の区切りがスペースかそうでないかであることだけであり、ある程度のニュースの関連付けを確認できた。

目 次

図 目 次

第1章　序論

1.1　背景と目的

情報通信技術の発展に伴い、今日では日本にいながら世界中で報道されている様々なニュースについて簡単に知ることが出来るようになってきている。さらに、インターネット環境の普及により、様々な国の新聞やニュースサイトに簡単にアクセスし、ニュース記事を手に入れることができるようになった。

そのため、自国のニュースサイトで流されている内容が、海外のニュースサイトではどのように表現されているかといったことや、世界で起きたニュースについて、各国のニュースサイトではどのように扱っているかといったように、複数のニュースサイトを通じて、あるニュースを多角的に調べることが可能となってきている。

しかしながら、世界各国のニュースというのは、その国の言語で書かれていることが普通である。そのため自然言語処理の観点から言語横断的にそれぞれのニュース間の関係を調べるには十分に高性能な機械翻訳技術の実現が必要となってくる。しかし、世界中のニュースサイトで使われている言語においてそれぞれに十分な機械翻訳技術を用意するのは簡単なことではないので、十分な言語資源がない状況においても世界中のニュースから関連するニュースだけを効率よく取り出すという点は技術的な課題と言える。

また、一口に Web ニュースサイトからニュースを入手すると言っても、簡単にできることではない。ニュースの分析のためには当然ニュース本文やタイトルのみが欲しい情報であるのだが、現実のニュースサイトを見てみればわかるように、ニュースサイトというのはニュース本文のみを表示してくれているわけではない。ニュースページは必ずと言っていいほど HTML で装飾されており、そこにはほかのページへのリンクやバナー広告といったものが含まれている。ほかにも、ニュースサイトをトップページからクローリングすれば、そのクローリングされた Web ページのなかにはニュースページ以外のページが含まれている。例えば、トピックごとにまとめられたニュースへのリンクとそのページのタイトルやサマリーのみが表示されている Web ページである。そのため、ニュースページからニュース本文とタイトルのみを取り出すことも解決すべき課題と言える。

そこで、本研究では言語横断ニュースコーパスの構築を目標とし、次のような制約があるとする。

制約

1. ニュースは Web ニュースサイトから入手する

プレインなニュース本文のみを得ることは難しく、また日々発信されるニュースをリアルタイムに入手するには Web から収集するのが良いと考え、Web ページからニュース本文を入手することを前提とし、Web ページからニュース本文以外の情報は自前で切り取らなければならないとする。

2. 言語資源が限られている
対象とする言語すべてについて形態素解析器やほかの言語への辞書を用意するのは現実的ではないので、基本的に文法はわからず、辞書も持っていないこととする。

このような状況下において次の要求を満たすための言語横断ニュースコーパス構築手法を提案する。
要求

1. 幅広い言語をカバーする
日本語や英語、中国語と言った、この手の言語横断的な試みでよく対象とされる言語以外にもなるべく多くの言語を対象とする。

2. 関連するニュース同士が言語横断的にリンク付けられている
ある言語に閉じた中で関連するニュースをまとめることは Google News 等でやられているが、そのように言語内で閉じてしまわないで、ほかの言語で書かれていても関連するニュースであれば関連すると明示できるようにする。

以上のような要求を満たすために、本研究では教師なし学習による Web ページのテンプレート学習をする Webstemmer [10] の改良と Wikipedia を用いた対訳辞書の作成を行い、その辞書による単語一致でニュースを表す特徴を抽出し、関連付けを行うものとする。

本研究の貢献としては、様々な言語で書かれたニュースを言語横断的に関連付けしたニュースコーパスとすることで、複数のニュースサイトからある事象のみに注目した分析をするといったことに対して有用なリソースになる。例えば、複数のニュース源の差異を考慮したニュース分析の研究 [13] といったものにおいて、本研究で構築されたコーパスを提供できればさらに幅広い発見ができるであろうと期待される。また、ニュースということにこだわらずに多言語コーパスとして見れば、言語横断情報検索（Cross-Language Information Retrieval:CLIR）や言語横断質問応答（Cross-Language Question Answering:CLQA）といった研究がされている分野においての貴重な言語資源として使えることが期待される。

1.2 本研究の貢献

本論文では、様々な言語で書かれたニュースをある程度の関連性をもってまとめることによってニュースの分類がしやすいようにするための方法を述べる。それによって、ほかの言語横断情報検索

や言語横断質問応答といった研究分野やニュース自体を解析してなんらかの統計情報やニュース同士の比較による事件やトピックをどのようにとらえているかといった情報の検索に役に立つはずである。例えば、吉岡 [13] の研究では、ニュースの発信者は、主にその国の読者の興味（自国にどのような影響があるか）を反映したような記事を作成すると考えており、ニュース源のニュースの取り扱い方の違いを分析することにより、そのニュース源の想定読者の興味の違いを分析するという研究を行っている。このような場合に、本研究であらかじめニュースをある関連性をもってまとめてあれば、調べたいニュースだけを取り出してきて、それを調査対象とすることで計算時間の削減などが期待できる。

1.3 本論文の構成

第 2 章 関連研究 既存の Web ニュースサイトの収集方法や、解析方法について紹介し、それらに対する本研究の位置付けを述べる。

第 3 章 Web ニュース記事の言語横断関連付けのための基本アルゴリズム 多くの言語が使われる Web ニュースサイトにおいて、どのように言語横断的に関連するニュースを見つけてくるかの基本的なアルゴリズムを述べる。

第 4 章 多言語で展開しているニュースサイト解析のためのラッパー作成手法 本研究において使用する Web ニュースサイトの分析に使用するための情報を取り出すためのラッパー作成方法について述べる。

第 5 章 Web ニュースの言語横断関連付けのためのニュース比較 本研究で用いる言語横断的に Web ニュースを関連付けする手法について述べる。

第 6 章 実験と評価 本手法を実環境で実験した結果を述べる。

第 7 章 結論 本論文の寄与と今後の課題について述べる.

第2章　関連研究

関連研究については以下の２つにわけられる。

1. Web ページから必要な情報を取り出す手法

2. 言語横断的なニュース関連付け

まず、Web ニュースサイトからニュース本文のみを取り出すために必要になるのが Web ページから必要な情報のみを取り出す手法になる。これらの手法は、Web ページのどこが必要であるかを表すラッパーを作成し、それを用いて情報を取り出すもので、ラッパーの作成に教師あり学習を用いる手法と教師なし学習を用いる手法の２つがある。また、言語横断的にニュースサイトを関連付けするために必要な言語横断的なニュース関連付けについての手法がある。

2.1　Web ページから必要な情報を取り出す手法

2.1.1　教師あり学習を用いたもの

WIEN [1] , SoftMealy [2] , Stalker [3] は、Web ページをトークや文字列とみなして、学習によりデリミタベース抽出ルールを作成する。WIEN ではすべてのデータアイテムがどこにあるかの Header と Tail についてと、個別のアイテムをマークする Left と Right デリミタを学習する。WIEN ではタプル内のアイテムの喪失や変化といった事態に対応できない。SoftMealy では有限状態変換器 (finite-state tranducer) として抽出ルールを作成する。Staler でも境界条件による抽出ルールを作成する。

Web ページを DOM ツリーとみなしてインタラクティブにラッパーを作成するものとして W4F [4] 、XWrap [5] 、　Lixto [6] 、　Interactive Wrapper Generation with Minimal User Effort [9] があげられる。W4F は抽出ルールとして HTML Extraction Language(HEL) という言語を用いている。ユーザー補助のために、DOM ツリー情報を表示することができるが、ルールをユーザーが書かなくてはならないために、ユーザーにかかる負担が大きい。XWrap ではユーザーとシステムが GUI を通してインタラクティブに会話して、ルールの表現力を制限するあらかじめ決められたテンプレートに基づいたルールを作成することができる。しかし、ラッパー作成にはとても長い時間がかかる。Lixto では system-internal datalog-like rule-based language である Elog に基づいた抽出ルールを作

成する。Lixto ではいくつかの Web ページに対してテストしたり訓練したりする機能を提供しておらず、ひとつの訓練データのみを対象としている。Interactive Wrapper Generation with Minimal User Effort では DOM ツリー表現において、求めるデータの場所までの DOM パスとそこまでの各ノードに関する属性をもとにした抽出ルールを作成する。この手法ではさらに、いくつかの訓練データにおいて作成された複数のラッパーをほかのデータに適用してみた結果をユーザーに見せ、どのラッパーがユーザーの要望にあうのかどうかというインタラクティブな選択をすることで訓練データ数を少なくすることに成功している。

これらの手法においてはどの方法を用いても訓練データが必要である。一つの Web ニュースサイトを訓練データとして作成したラッパーがほかの Web ニュースサイトでも使える保障はなく、また、いくつかの Web ニュースサイトを訓練データとして有効なラッパーが作成できる保障もないため、Web ニュースサイトごとにラッパーが必要となると考えられる。Web ニュースサイトが数多くあることを考えると、その都度訓練データを作成することは実用的ではなく、さらに Web ニュースサイトがフォーマットを変えてきたときに再度訓練データを作成しなければならないことを考えると実用的ではないと考えられる。

2.1.2 教師なし学習を用いたもの

IEPAD [7] では規則的で連続するパターンを捕捉する PAT ツリーとよばれるデータ構造を用いてパターン発見をおこなっている。RoadRunner [8] では、サンプル HTML ページに対して、類似点と相違点に基づいたパターンを見つけ、関連する構造を特定するミスマッチを利用する。これらはどれも、単にパターンを発見し、それに基づいてデータを取り出してくるだけであるので、ニュースページに適用してもニュース以外の部分も抽出されてしまうことが予想される。

次に紹介する Webstemmer は Web ニュースサイトからニュース本文をとってくることに特化した手法で本研究でも使用するため、くわしく解説する。

Webstemmer は以下の仮定のもとにして Web ページの分析をしている。

1. すべてのニュースページは共通したいくつかのレイアウトが使用されている
2. 各ニュースページにはメインとなる本文が一つのみ含まれている
3. レイアウトは基本的に不変である

この仮定のもとに、あるニュースサイトの同一レイアウトをもつページをまとめ、その中で同じ部分を探し出す。バナー広告やほかのページへのリンクなどはレイアウトが同じなら大体同じ内容であるので、これらを取り除くことにより、メインの文章のみが抽出される。

Webstemmer の動作の概要を表すと次のようになる。

- あるサイトのページを異なるレイアウトごとに分類する（図 2.1 ）

- 同一レイアウトのページを並べる（図　2.2 ）
- 共通する部分を削除し、残ったものをニュース本文とタイトルとする（図 2.3 ）

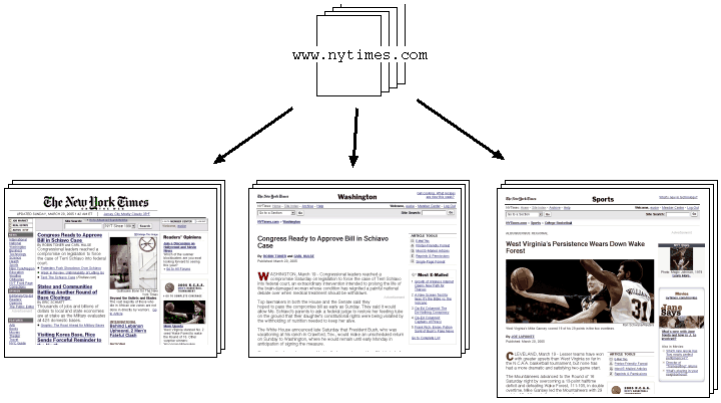


図 2.1: Cluster layouts

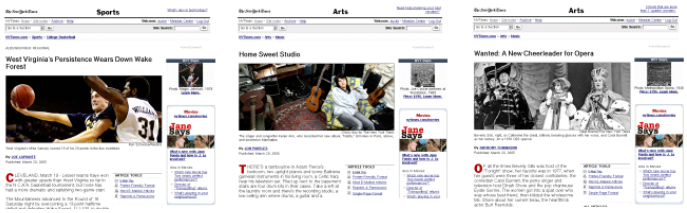


図 2.2: Align the pages whose layouts are similar

次に Webstemmer のレイアウト分析アルゴリズムについて説明する。Webstemmer のレイアウト分析は、似たようなページ構造（レイアウト）をもつページをまとめてクラスタリングすることによって、そのサイトが使っているフォーマットをパターンとして抽出している。クラスタリングを行うために、各 HTML ページのレイアウトを比較し、類似度を算出している。次に互いに類似度の高いページをクラスタとしてまとめ、クラスタ内のページに共通する要素をパターンとして抽出する。各段階における詳しい処理方法を順を追って説明する。

Step1 レイアウトの解析

各 Web ページを「レイアウトブロック（LayoutBlock）」と呼ばれる要素に分解する。木構造である HTML ページのタグを、HTML ブロック要素（table,div,tr など）のみに注目し、ページのレイアウトをこれらのブロックの一次元的な列としてみなす（図 2.4 ）。

Step2 類似度の計算

図 2.3: Remove same blocks and extract news body

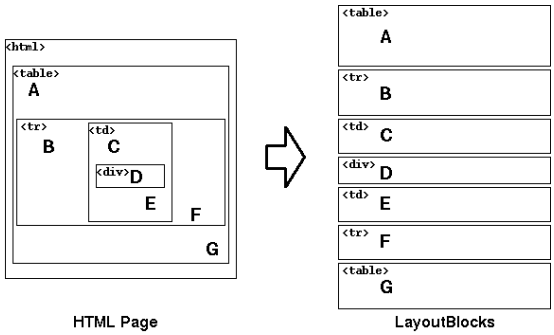


図 2.4: Parse HTML page to Layout Block

各ページのレイアウトの類似度を、それらのレイアウトブロックの列の編集距離を用いて計算する。二つのレイアウトブロック列に共通する最大の順列（図 2.5 ）と、元のブロック列との比率を類似度とする。

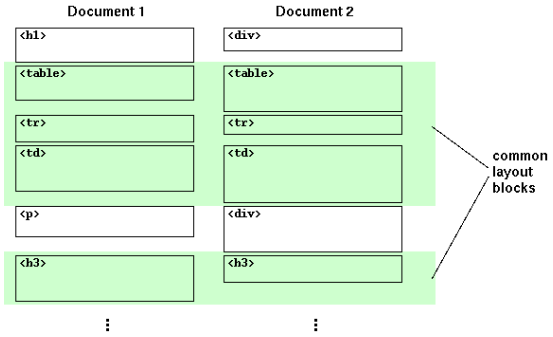


図 2.5: Comapre Layout Block and Calculate Similarity

Step3 クラスタリング

クラスタリングを行うために、N 個のページすべてのとりうる組み合わせに対して Step2 の類似度計算を行い、類似度があるスレッショルドより大きいブロック列ペアを一つのクラスタとしてまとめる操作を行う。クラスタリングが完了したら、各クラスタごとに共通するレイアウトブロックの HTML 要素を取り出し、それをレイアウトパターンとして使用する（図 2.6 ）。

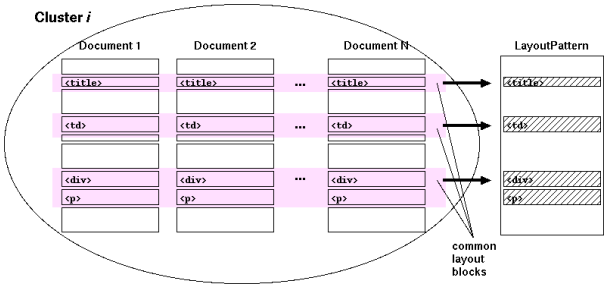


図 2.6: Cluster Layouts and extract pattern

Step4 バナーやリンク等の不要なものを取り除く

次に、抽出されたそれぞれのレイアウトパターンの中から、バナー広告やほかのページへのリンクと思われるブロック要素を取り除く。実際にはパターン中から完全に削除されるのではなく、あるレイアウトパターン中の各レイアウトブロックに DiffScore という値をつけ、この値が一定以下のブロックはニュース本文を抽出するときにスキップするようにしている（図　2.7 ）。DiffScore は Step3 で得られた各クラスタのそれぞれのブロックにある文字列を比較して、それが各ページごとに

どれくらい異なっているかを計算することにより定義される。

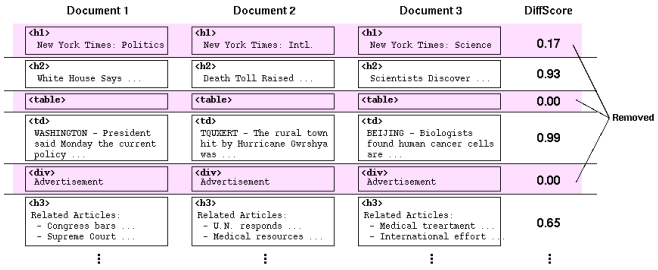


図 2.7: Remove ads and links

Step5 タイトルとニュース本文を発見する

Webstemmer ではページの本文をもっとも長いテキストが含まれるレイアウトブロックとして定義している。ニュースのタイトルについては、

1. 本文よりも前にある
2. 本文と類似している

という仮定のもとに抽出している（図 2.8 ）。

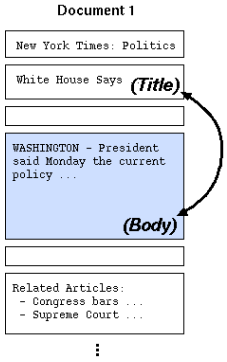


図 2.8: Extract news title

以上の処理によってレイアウトのパターンが生成される。生成されたパターンを使って与えれた Web ページからニュース本文とタイトルを抜き出すには、まず与えられた Web ページを分析と同様にレイアウトブロック列に変換し、それとそれぞれのレイアウトパターンのブロック列との類似度を計算する。その中で最も高い類似度をもつパターンをその Web ページのレイアウトとして判定し、パターン中で指定された本文ブロックとタイトルブロックからテキストを抽出してニュース本文とタイトルを得る（図 2.9 ）。

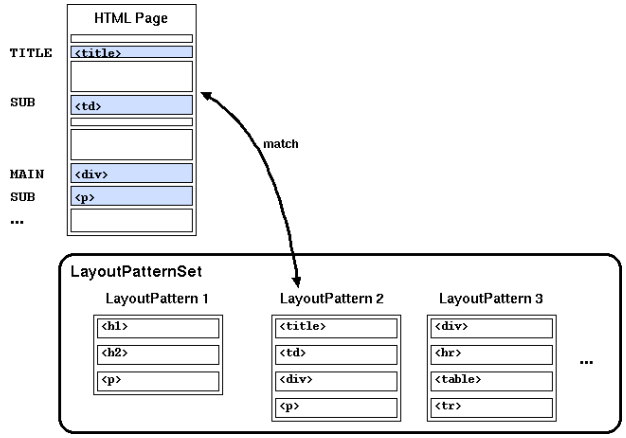


図 2.9: Adapt pattern and Extract news body and title

教師なし学習の手法では、データのラベル付けが必要ないため、多くの Web ニュースサイトにおいて適用するには教師あり学習の手法より良いと思われる。さらに、Web ニュースからニュース本文とタイトルを抜き出すことに特化した Webstemmer は最適であると思われる。

2.2 言語横断的なニュース関連付け

言語横断的なニュース関連付けの関連研究として、Integrating Cross-Lingually Relevant News Articles and Monolingual Web Documents in Bilingual Lexicon Acquisition [11] と Navigating Multilingual News Collections Using Automatically Extracted Information [12] などがある。Integrating Cross-Lingually Relevant News Articles and Monolingual Web Documents in Bilingual Lexicon Acquisition においては、図 2.10 のようなシステム構造をとっており、ニュース記事の対象とする言語として日本語と英語を用いており、英語の文章を MT の商用ソフトウェアを用いて日本語の文章へと機械翻訳している。日本語の文章と日本語化された英語の文章はそれぞれ単語列へと分解され、単語頻度ベクトルを作成する。そして、日本語の文章と英語の文章の類似度はその単語頻度ベクトルのコサインによって計算され、その値が一定以上ならそのニュースは関連していると判定される。

Navigating Multilingual News Collections Using Automatically Extracted Information においては、ニュース記事の対象とする言語は 20 個の EU 公式言語としている。まず、ニュース記事を凝縮型クラスタリング手法で階層的クラスタリングツリーを作成する。すべてのニュース間の類似度をドキュメントベクトルのコサインで計算する。その中で一番類似度の大きいもの同士をまとめクラスタとし、新しくできたクラスタは一つのドキュメントとみなす。この処理を繰り返し行い、すべて

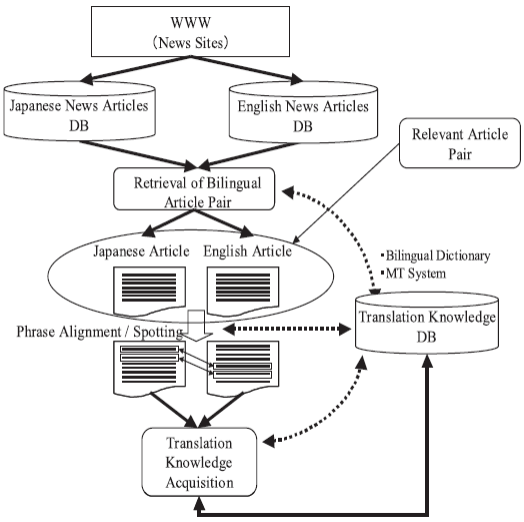


図 2.10: Integrating Cross-Lingually Relevant News Articles and Monolingual Web Documents in Bilingual Lexicon Acquisition

のニュースが一つのクラスタにまとまるまで続ける。こうすることでドキュメントの類似度に基づいた樹形図ができあがる。ドキュメントベクトルは名前リストと地名辞典にでてくる単語を軸とする。また、言語をまたいでニュースクラスター同士を関連付けするために、クラスタのキーワードを多言語シソーラスである Eurovoc と対応付けを行っている。この手法で使われる類似度の指標を図であらわしたものを図 2.11 に載せる。

両者の手法とも、英語を日本語へ翻訳する、多言語シソーラスである Eurovoc との対応付けを行う、といった十分に対象言語間における言語情報が手に入るということを前提としている。また Navigating Multilingual News Collections Using Automatically Extracted Information については、20 という多言語を対象としているものの、すべて欧州系の言語であるため、ほかの地域のニュースが得られにくいと考えられる。

欧州系に限らずにより多くの言語をカバーするためには、対訳辞書が入手しにくい、あるいは言語そのものの形態素解析器が入手できないような環境においても、関連するニュース間を言語をまたいでリンク付けできるような手法が必要である。

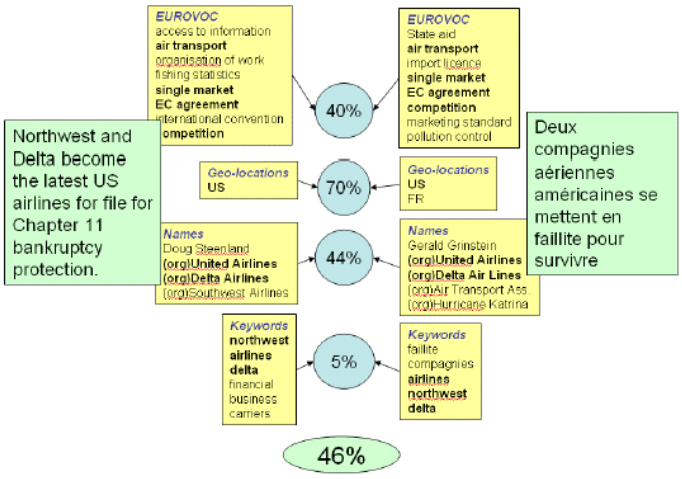


図 2.11: Navigating Multilingual News Collections Using Automatically Extracted Information

第3章　Webニュース記事の言語横断関連付けのための基本アルゴリズム

ここではWebニュースサイトから情報を取り出してきて、それぞれのニュースについてどのように言語横断的に関連しているかどうかを判定するアルゴリズムを述べる。

3.1　システム概要

本研究で実際に実装したシステムの概要図を図3.1に示す。システムのコンポーネントは大まかに以下の2つに分けられる。

- ラッパー作成&適用
  Webニュースサイトから Webニュース記事の関連付けのために必要な情報を抽出するために必要なコンポーネント。
- Webニュース記事の関連付けラッパーで抽出された情報をもとに、実際に Webニュース記事同士の関連付けを行う。ここで使う情報は、ラッパーから出力されるカテゴリとニュース記事へのリンクに含まれているアンカーテキストと URL である。

3.2　Webニュースサイトからの情報の取り出し

Webニュースサイトから URL を指定して特定のページを機械的にダウンロードするのはそんなに難しいことではないが、ダウンロードしたニュース記事ページから自動的にニュース情報を取り出すことは簡単ではない。なぜなら、Webページというのは基本的に HTML で記述されており、ブラウザ上で人間の目でみれば一目でどこがニュースタイトルでどこがニュース記事でどこがニュース記事へのリンクなのかといったわかることが、機械的に処理する機械にとってはただの文字列であり、そこに書かれている HTML を解釈して必要な情報を取得しなければならない。そこで、ダウンロードした Webニュース記事から Webニュース記事の関連付けに必要となる情報を取得するためのラッパーの作成と適用を行う。ラッパーで取り出す情報は、ニュース関連付けに必要となる、カテゴリへのリンクとニュース記事へのリンクで、そこからアンカーテキストと URL を出力する。

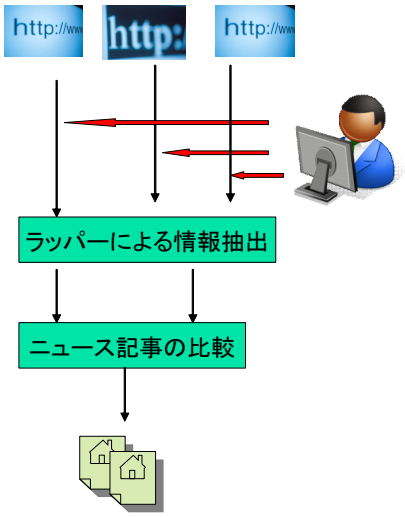


図 3.1: 本システムの概要図

3.3 Web ニュース記事の関連付け

3.3.1 複数言語でニュースを提供しているサイトの利用

単純にすべての Web ニュースサイトからすべてのニュースを取得してそれらを一つずつ比較していくという手法では、ニュース記事の多さや言語の多様性の問題で計算量が多く、無駄が多いものとなってしまう。そこで、それらを省くために本研究ではすでに複数言語でニュースを提供しているサイトを利用することを考える。これは同じ管轄のもとで管理されている Web ニュースサイトは同じニュースソースを使っていて、それらを元に各言語版のニュースを提供しているのではないかという期待に基づく。これを図で表したものを図 3.2 に載せる。もしそうであれば、そのニュースサイトにおいては各言語で提供されているニュースは言語が違えど同じ事件、事柄について書いてあることが多いと考えられる。そのため、無作為にニュースを取り出して比較するよりも関連しているニュースを見つけやすいというメリットがある。

3.3.2 ニュースカテゴリの利用

複数言語で提供しているニュースサイトにおいては、似たようなニュースカテゴリを持っていることが多い。例えば、経済、政治などである。言語が違っても同じカテゴリであるとみなされる場合には、そのカテゴリないでニュースの比較を行ったほうが関連しているニュースを見つけやすいと考えられる。そこで、ラッパーで出力されたニュースカテゴリの情報を用いて、ニュースカテゴリ

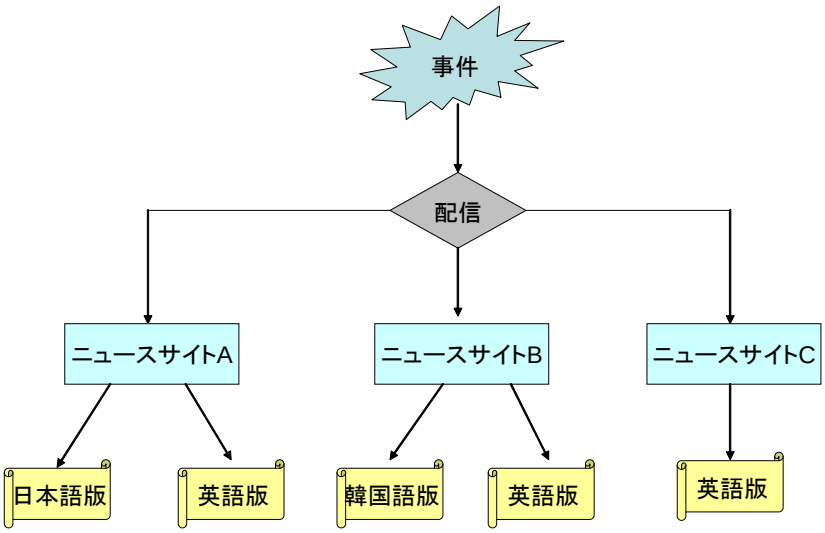


図 3.2: ある事件について取り扱っているニュースサイト

の対応付けを先に行い、対応付けされたニュースカテゴリないでニュース記事の比較を行うことで、比較コストの軽減と関連付け精度の向上が見込めると考えられる。

3.3.3 ニュースの比較

ここでは上記のように、複数言語でニュースを提供しているニュースサイトを対象とする。各言語のニュースを取り出してきてどれとどれが関連しているかを見つける手法として、ここではニュースのタイトルのみを取り出してそのタイトルを比較してある程度以上似ているならば関連しているとみなすことにする。

ここで、ニュースの本文全部を使わない理由として

- ニュースのタイトルは簡潔にニュース自体を表現している
  ニュースのタイトルは Web ニュースサイトを提供している側がそのニュースをどのように表現したら、簡潔にかつわかりやすくニュース本文を読まなくてもどのようなニュースなのかわかるようにつけているため、タイトルのみで十分にニュース記事の比較資料として使えると考えれる。

- ニュース本文全部を使うと長すぎる
  ニュースの本文はニュース記事によってかなりの長さになるため余計な単語やフレーズがでて

きてしまい、比較する際のノイズになってしまう可能性があることと、比較する計算コストが長さに比例して大きくなってしまうから。

が挙げられる。

3.3.4 他のニュースサイトへの拡大

上記の手法で、複数の言語で提供している Web ニュースサイトの中でのニュースの関連付けはできるのだが、すべての Web ニュースサイトが複数の言語で対応しているわけではないし、一つの Web ニュースサイトが多くの言語を提供しているわけでもない。そこで、複数の言語で提供している Web ニュースサイト同士や、単一の言語でしか提供していないニュースサイトのニュースについてもまとめてやる必要がある。ここでは、先ほど述べたような計算コストの増加や言語リソースの問題を回避するために、複数言語で提供している Web ニュースサイトにおいて関連しているニュースが見つかったものを核として、それらのニュースと関連しているかどうかを比較する手法をとることにする。これは、複数言語で提供されているニュースは需要が高いから複数言語で提供されているという考えのもとで、そのようなニュースはほかのニュースサイトでも配信されている可能性が高く、関連するニュースが見つけやすいと考えられるからである。

第4章　多言語で展開しているニュースサイト解析のためのラッパー作成手法

4.1　目的

ラッパーの作成の説明の前に、何のための、何を抽出するための、どのように、どのようなラッパーを作成するかという目的について整理する。本研究では、多言語でニュースを提供しているニュースサイトにおいて、各言語版のニュースサイトでそれぞれニュースを収集し、最終的に関係のあるニュースを関連つけるための第一歩として、各言語版のニュースサイト内でニュースカテゴリやニュース記事へのリンクを抽出する。例えば、http://www.intelasia.com/ というサイトでは英語版（図 4.1 ）とベトナム語版（図 4.2 ）がある。このようなサイトにおいて、左側にあるようなニュースカテゴリへのリンクや中央にあるようなニュースへのリンクをそれぞれ抽出するためのラッパー生成＆適用をゴールとする。そのさいに、多言語に対処できるようにするために、言語情報に頼らずに処理するため、用いる情報は HTML の DOM ツリーのみを扱うことにする。その抽出を手軽にできるようにするために、インタラクティブにユーザーがどのリンクを抽出すればよいかを指定でき、さらに一箇所リンクを指定すればそのリンク以外でもニュースカテゴリやニュースへのリンクを持っているであろうリンクを抽出できるようなルールを生成する。そのさいに、ニュースサイトの HTML 構造では、ニュースカテゴリやニュースへのリンクというのは DOM ツリー上で規則性をもって出現するという性質を用いて、リンクへのパス情報とどの DOM エレメントがサブツリーの親であるかという基本的な情報のみで構成されたラッパーを生成する。得られたラッパーをもとに、DOM ツリーを辿り、リンクの抽出を行なう。

4.2　インタラクティブな操作

ラッパーの生成を手軽に行なえるようにするために、インタラクティブな操作をできるようにする。そのために、本研究では javascript を用いて、ブックマークレットとして実装した。そうすることで、ユーザーは対象とする Web ニュースサイトをブラウザで訪れた後、すぐに操作を始めることができるようになる。ニュースサイトのどの部分を抽出したいかという指定は、ユーザーはマウスのクリックによって簡単に指定できる。ユーザーのマウスクリックを検知すると、javascript の機能によってその部分の DOM エレメントを取得でき、それを用いてラッパーの作成を始める。ラッパー

図 4.1: 英語版のニュースサイト

の作成が終わったら、ブラウザ上にラッパーを適用して抽出される部分を色のついた四角で囲んだ画面が表示される。ユーザーはこの結果を見て、作成されたラッパーが求めているものかどうかを判断することができる。先に挙げたニュースサイトで左側のニュースカテゴリの部分と中央にあるニュースへのリンク部分を選択した場合の結果を図 4.3 に載せる。もとの画面と比べて、抽出される部分が色付きの枠で囲まれているのがわかる。生成されたラッパーの情報は別ウィンドウが開かれてそちらに表示されている。

4.3 ラッパー作成

本節ではラッパーの生成方法について述べる。本手法では、Web ページを DOM ツリーとみなして処理する。ユーザーからのインタラクティブな操作によって得られた DOM エレメント（以下 E とする）を基に、以下の操作を行なう。

1. E について、ドキュメントのルートからのパス情報を取得する
2. サブツリーのルートとして E の親エレメントをセットする
3. サブツリーのルートから下にあるすべてのリーフノードを探索し、その中に E のパス情報と一致するものがあるかどうかを調べる
4. 一致した際に、E ではない DOM エレメントであることを確認する
5. ラッパーとして、E とサブツリーのルートのパス情報を出力する

図 4.2: ベトナム語版のニュースサイト

6. ラッパーが指定数見つかるまで、サブツリーのルートとして、いまのルートの親エレメントをセットして 3．にもどる

これらの操作によって、求めたい DOM エレメントのパス情報とどの DOM エレメントをルートとしたサブツリーを探索対象とするかというパス情報をもったラッパーが生成される。擬似ソースコードを図 4.4 に載せる。

例えば、図 4.3 の左側のニュースカテゴリの部分のラッパーは求めたい DOM エレメントのパス情報として *undefined / HTML / BODY / TABLE / TBODY / TR / TD / TABLE / TBODY / TR / TD / TABLE / TBODY / TR / TD / TABLE / TBODY / TR / TD / DIV / TABLE / TBODY / TR / TD / SPAN / A* 、サブツリーのルートのパス情報として *undefined / HTML / BODY / TABLE / TBODY / TR / TD / TABLE / TBODY / TR / TD / TABLE / TBODY / TR / TD / TABLE / TBODY / TR / TD / DIV / TABLE / TBODY* が出力される。ラッパーの適用は、二つのパス情報を用い、DOM ツリーを辿って条件に合った DOM エレメントを列挙するだけである。ここで、求めたい DOM エレメントのパス情報のみではなく、サブツリーのルートも情報として使用したのは、DOM エレメントのパス情報だけでは求める DOM エレメント以外にも合致する可能性があるため、ユーザーの指定した DOM エレメントとなるべく近い場所で共通の祖先を持つ DOM エレメントのみをなるべく取り出すためである。

4.4 ラッパーの適用

生成されたラッパー情報を使って求める DOM エレメントを取り出す手法は以下の手順である。

1. ラッパーを適用する対象となる HTML ページの DOM のルートエレメントを取り出す。

図 4.3: 英語版のニュースサイトでラッパーを生成&適用

2. ルートエレメントから順に下っていき、サブツリーのルートエレメントを見つける
   これは、上から下に幅優先探索でツリーを辿っていき、DOM エレメントごとにラッパーのサブツリーのルートパス情報とそのエレメントのルート情報を比べ、一致するものを見つけることでできる。

3. サブツリーのルートとなる DOM エレメントが見つかったら、そこをルートとして求める DOM エレメントを見つける
   サブツリーのルートエレメントから幅優先探索でツリーを辿っていき、DOM エレメントごとにラッパーの求める DOM エレメントのパス情報と比較していき、一致するものを見つける。

4. サブツリーの探索が終わったら再度手順 2 にもどり新しいサブツリー候補を見つける

この操作を繰り返すことで、ラッパーから求めたい DOM エレメントを発見することができる。ラッパー生成したときは、ユーザーの指定した DOM エレメントを基にしてそこから上に向かって

探索したが、ここではHTMLページのルートから下に向かって探索をかけている。そのため、ラッパーと同じサブツリーのパス情報とDOMエレメントのパス情報を持ったものはいくつもある可能性があるため、ラッパー生成時に求められたDOMエレメントより多くのものが見つかる可能性がある。

4.5 実装の詳細

ここでは実装したものについての詳細を述べる。

4.5.1 ラッパーの生成

説明の部分で述べたように、ラッパーの生成にはWebブラウザとインタラクティブに連携することができ、相性のよい言語としてJavascriptを用いた。JavascriptとWebブラウザではブックマークレット呼ばれる、現在見ているページに対して別のページのJavascriptを適用させるテクニックがあるので、それを用いてユーザーが見ているWebニュースサイトにそのまま適用させ、その画面上でクリック動作をすることでグラフィカルにラッパーの生成をできるようにしてある。

ラッパー生成Javascriptの適用には次のようなリンクを用いる。

```
javascript:(function(){var\%20s=document.createElement("script");
s.src="http://www.logos.ic.i.u-tokyo.ac.jp/~s-yoshida/js/wrapper_generation_from_mouse.js";
s.type="text/javascript";document.body.appendChild(s)})();
```

ここで、

```
http://www.logos.ic.i.u-tokyo.ac.jp/~s-yoshida/js/wrapper_generation_from_mouse.js
```

のリンクには実際にラッパー生成するJavascriptに繋がっている。

Javascriptを起動させると、ページの下部にstart interactionというボタンがでるので、そこをクリックしてからWebページ上のループしている求めたいニュースカテゴリやニュース記事へのリンクをクリックすることで、ラッパーが生成される。

計算して出てきた結果のDOMエレメントは表示しているWebページ上で色付きの四角で囲まれるので、ユーザーが求めたいものがでてきたかどうか一目でわかるようになっている。また、そのとき取得されるDOMエレメントのhref属性とText属性もラッパーの結果といっしょにWebページの違う部分に出力される。

4.5.2 ラッパーの適用

Javascript は Web ブラウザ等で使用される用で、ローカルで動く適当なものが見つからなかったので、ラッパーの適用には Java とそれで実装された HTML パーサの NekoHTML を用いている。

テスト段階では IE ブラウザを使用していたのだが、ブラウザでは HTML ページの記述において table 属性中に tbody 要素がなくても自動生成されるが、今回用いた NekoHTML パーサではそれがされないという実装上の違いがあったので、それを考慮にいれたラッパーの適用となっている。

```
/* root をサブツリーのルート、loop_tagをユーザーが指定したDOMエレメントのパス、
original_id をユーザーが指定したDOMエレメントのID、
root_found を見つかったサブツリーのルートを収める変数とする */
function create_wrapper( root , loop_tag , original_id , root_found , result )
{
   /* これ以上親がない場合帰る  */
   if( root == null )
        return false;

   /* 子ノードがいない場合、探索できないので親ノードへ移行  */
   if( root.childNodes == null )
      return find_loop( root.parentNode , loop_tag , original_id , root_found , result );
    /* rootをサブツリーのルートとして、探索する  */
    if( trace_subtree( root , loop_tag , original_id , root_found ) )
      {
         if( ! ( root in root_found ) )
            {
               root_found[ root_found.length ] = root;
            }
      }

    /* rootは探索が終わったので、rootの親ノードをサブツリーのルートとして探索させる  */
   if( create_wrapper( root.parentNode , loop_tag , original_id , root_found , result ) )
      return true;

}

/* rootをサブツリーのルートとし、loop_tagをもつリーフノードが存在するか調べる  */
 function trace_subtree( root , loop_tag , original_id , root_found )
{
    /* 子ノードがない場合探索終了  */
   if( root.childNodes == null )
      return false;
   var len = root.childNodes.length
   for( var i = 0 ; i < len ; i ++ )
      {
         var el = root.childNodes[i];
         var path = get_path(el);

            /*  子ノードのパスと求めるパスが一緒かどうか判定  */
         if( loop_tag == path )
            {
                  /* パスが一緒の場合、ユーザーから指定されたノードでないことを確かめる  */
               if( original_id != el.id )
                     return true;
            }
            /* 子ノードより下方向について探索させる  */
         if( trace_subtree( el , loop_tag , original_id , root_found ) )
            return true;
      }
}
```

図 4.4: ラッパー作成擬似ソースコード

第5章　Webニュースの言語横断関連付けのためのニュース比較

この章では、どのように言語横断的に Web ニュースを関連付けするか説明する。まず、本研究ではどのようなことが重要かについて述べる。

5.1　言語横断的なニュースコーパスを構築するための要求

最終的な目標は言語横断的なニュースコーパスを構築し、そのニュースコーパスにおいてニュースがある関連性をもってまとまって欲しいというものである。そのようなニュースコーパスを構築するための要求として以下が挙げられる。

- 対応する言語が多い
  限られた言語しか対応していないのでは、リソースとしては乏しいものになってしまう。また、単純に対応する言語が多いほうがニュースコーパスを利用する側にとって調査できる余地が大きく、より広い範囲でニュース記事の比較が可能というのもある。
- 言語に関する知識を事前にあまり知らなくてよい
  対応する言語が多いにこしたことはないのだが、それに伴ってそれぞれの言語について膨大な知識を事前に取得しなければならないというのでは話にならない。そこで、対応する言語についての事前知識をあまり必要としないことが望ましい。
- 対応するニュースサイトが多い
  これも上記と同様の理由で、多くのニュースサイトを利用できたほうがそれだけ多くのニュース記事を集められており、ニュースコーパスの情報リソースとして役に立つはずである。

これらの要求を満たすために、本研究では Wikipedia を用いた単語翻訳手法を用いることにする。

5.2 ニュースタイトルの比較

5.2.1 ニュースタイトルの比較時に用いる Wikipedia 単語翻訳

本研究でもちいる Wikipedia による単語翻訳について述べる。Wikipedia はウィキメディア財団が運営しており、コピーレフトなライセンスの下、インターネット上の誰もが無料で自由に編集に参加することができ、世界各国の言語で展開されているオンライン百科事典である。Wikipedia には様々なジャンルにおいて多くの記事が載っており、それらの記事にはかなり詳細なことが書かれていることも少なくない。ただし、誰でも自由に書き換えできるため、記事自体の情報の信頼性は必ずしも保障されていない。そのため、記事に載っていることすべてを信じてしまうことは危険でもある。そこで、ここでは記事のタイトルと翻訳リンクのみを使用することにする。記事のタイトルはそれ自体が何の記事であるかということを表す標識であり、この部分に嘘の情報が入り込む余地はかなり少ないといってよい。仮に嘘のタイトルがついていたとしても、その記事には何の情報も載っていないか、あるいは運営側からの削除があると期待できる。次に Wikipedia 記事の翻訳リンクについて述べる。wikipedia の記事には、ページの左下のほうにその記事の他の言語へ翻訳された単語について書かれている記事へのリンクがある。そのリンクを辿ることで元の記事のほかの言語の表現を知ることができる。例として、日本語で首相という単語について Wikipedia で調べた場合の英語版へのリンクを辿ったときの結果を図 5.1 に載せる。

5.2.2 言語知識を用いない単語の切り出し

上記の手法でニュースタイトル内の文字列を単語として、それを目的の言語への翻訳ができる。しかし、ニュースタイトル内の文字列の中でどの範囲の文字列を単語として見るかという問題が残っている。ここでは、言語知識をあまり用いないということを目標にしているため、事前にどこが単語であるかということを知ることは難しい。そこで、我々の手法では以下の二つの手法を用いる。

- n-gram を用いた単語区切り
  ある一定の数 (n) を決め、1 から n まで順に数を上げていき、その数だけ文字を区切っていき、それを単語としてみる。

- スペースを用いた単語区切り
  単語がスペースで区切られていることを前提にして、スペースがあればそれらに挟まれている文字列は一つの単語であるとする。

n-gram を用いた単語区切りとスペース区切りによる単語区切りを用いることにする。二つの手法を使う理由は、言語の中には日本語のように単語の区切りが見ただけでは分からず、文字列がずらずらと並べられているものと、英語のように基本的に単語はスペースで区切られていて、単語の羅列

図 5.1: 日本語の「首相」を Wikipedia を用いて英語へ翻訳した例

で文章が構成されているものの二つがあるからである。どの言語であっても、スペース区切りの言語であるか区切りのない言語であるはパッと見でわかりやすく、文法や単語辞書のような高度な言語知識も必要なく適応できるため、対応する言語を広く取りたいという目的に合致する。

5.2.3 違う言語間でのニュースタイトル比較

言語が違うニュースタイトル文において、それを比較して関係があるかどうかを調べる方法には次の 2 種類が考えられる。

1. 片方の言語の文をもう片方の言語へ翻訳する
2. 両方の言語の文を別の一つの言語へ翻訳する

たいていの場合は 1 番の手法が用いられるが、それは両方の言語についての知識が豊富であり、形態素解析や対訳辞書が得られる場合に可能であって、そのような知識を用いないとなるとそう簡単ではない。今回用いるのは単純なルールに則った単語区切りと区切られた単語が Wikipedia 上でほかの言語への単語翻訳があるかどうかというだけであるので、ある言語からある言語へ一発で翻訳単語が見つかる可能性はあまり高くないと考えられる。Wikipedia 上では英語の記事が圧倒的に多く、ほかの言語はそれ以下の量であるので、ある言語の単語からほかの言語への翻訳のなかで一番存在する可能性が高い言語は英語であると考えられる。そこで、我々の手法では 2 番の手法を用い、両方の言語の文を英単語へと翻訳、変換することにする。また、対応する言語を広げていったときに、1 番の手法ではそれぞれの言語同士で翻訳しなおさなければならないというデメリットがあるが、2 番の手法では同一の英単語ベクトルになっているので翻訳の手間が省け、検索しやすいというメリットが生まれる。

5.2.4 同一言語間でのニュースタイトル比較

言語が同じニュースタイトル文では、英語に翻訳せずにそのままの文で比較することにする。比較するには、文中に同じ単語がでてくるかどうかで類似度の判定を行う。その際に使う単語は、Wikipedia に出てくるものを使う。単語翻訳の際にも Wikipedia を使ったが、このときには翻訳リンクがなくても Wikipedia に記事があるかどうかのみを単語の判定基準とする。これは、Wikipedia にはある言語である単語の記事があっても必ずしも翻訳リンクがあるわけではなく、記事のみしかない場合もあるので、単語判定のみを使うことでより多くの単語を使用することができるからである。

5.2.5 単語の思みつけ

共通してでてきた単語の数だけで類似度の指標としてしまうと、ニュースによくでてくるような単語ばっかりがあつまって意味のない結果がでてきてしまう可能性がある。特定のニュースにしかで

てこない単語にはたとえ一単語しか出てこなくてもその重みを増して計算してあげたほうがより精度が高くなると考えられる。そこで、本研究では、一単語の重み付けとして、idf(Inverse Document Frequency) を用いた。

$$\text{単語の重み} = log(\frac{\text{ドキュメント数}}{\text{その単語の出てくるドキュメント数}}) \tag{5.1}$$

この値を用い、共通の単語がでてくるごとに類似度として加算していき、最終的な値をその文書ペアの類似度とする。

5.2.6 ニュースカテゴリの利用

以上の手法で、あとは全体全でニュース記事の比較をすることもできるのだが、ここではさらに効率よくマッチングさせるためにニュースカテゴリを用いる。ニュースカテゴリは、ニュースサイトの提供者がニュースをある程度分野分けしてくれておいてくれるものなので、このニュースカテゴリ内でニュース記事をマッチングさせれば計算コストの軽減とマッチング精度の向上が見込めると考えられる。

ニュースカテゴリの抽出自体はラッパーの項で説明してあるので、ここではラッパーを適用した結果出てくるアンカーテキストと URL の使用方法について述べる。アンカーテキストは、上記の Wikipedia の翻訳手法を用いて英単語列に変換する。そして、URL のほうは/で区切って単語にして英単語列に加える。そして、出てきた英単語列をサイトのカテゴリ同士で比較し、一致する英単語が多い順にカテゴリの対応とみなす。実際は、カテゴリの対応は一単語でなされる場合が多いので、ほかの英単語列とは違う単語で一致したものを優先するようにしてある。これは、URL で共通してでてきやすい単語（ex.list,index）をわかりやすいく排除するためである。

5.3 実装の詳細

5.3.1 Wikipedia の利用

説明では、単語に区切ったあとは Wikipedia を引いて、その結果をもって単語であるかどうか、英語へのリンクがあるかどうかというのを調べてチェックすると述べたが、すべてのニュースタイトルにおいて、n-gram やスペース区切りで文を区切ると、単語以外の意味のない文字列が取り出されることのほうが多い。そのため、すべての単語候補文字列で Wikipedia を引くということをすると Wikipedia に非常に迷惑がかかってしまう。そこで、Wikipedia ではオフラインでも利用できるように Wikipedia 上のデータをダウンロードできるようになっているのでこれを用いることにする。

我々の手法で用いる Wikipedia の情報は、単語であるかどうか、英語へのリンクがあるかどうかであるので提供されているデータの中で、全項目のページ名一覧である all-titles-in-ns0.gz と記事ペー

ジについての情報である page.sql.gz と各記事ページについてのほかの言語へのリンク情報を含んでいる langlinks.sql.gz を用いる。

ページ名一覧についてはそのまま用い、データベースにすべて入れてそこから検索できるようにして、単語かどうか判定できるようにする。ほかの言語へのリンク情報についてはそのままでは使いやすい形になっていないので次のような処理を施し、使いやすい形にする。

1. page.sql.gz と langlinks.sql.gz を MySQL のデータベース上に入力する
2. ある単語について、もとの言語での記事タイトルとその英語版でのタイトルについての情報を取得する
   これは 1 でデータを入力したデータベース上で、次のような SQL 文を実行すればよい。

```
SELECT page.page_title , langlinks.ll_title FROM page,langlinks
WHERE langlinks.ll_from = page.page_id AND langlinks.ll_lang = "en";
```

これで、ある言語について、Wikipedia 上に存在する単語情報と単語から英単語への変換についてのデータがでてくるのでこれらを用いて実際の Wikipedia へのアクセスは最小限にとどめる。

5.3.2 単語チェック

単語のチェックには基本的に得られた文字列をそのまま使い、スペース区切りか n-gram で区切って単語としてそのまま Wikipedia のデータベース上にあるかどうか調べるのだが、テストしていた段階で全角英数字が使われているとその単語が半角では存在していても見つからないという事態が起きた。これは日本語を処理する場合特有な例なのかもしれないが、このような場合に全角英数字の単語が得られないというのは比較する際に大きな情報欠落と考えられるので、すべてのチェックする文章に対して全角英数字は事前に半角英数字に変換しておくという処置を施した。

また、数字については、Wikipedia 上にもいくつかの数字データについてはデータベース上にあるが、当然すべての数字を網羅しているわけではないので、数値列については無条件で単語であるという判定をしている。

第6章　実験と評価

ここでは実際に実験してみた結果と評価について述べる。

6.1　実験対象

今回実験の対象としたサイトは

http://www.yonhapnews.co.kr/ （韓国語版）http://japanese.yonhapnews.co.kr/ （日本語版）のペアと

http://www.joins.com/ (韓国語) http://japanese.joins.com/ (日本語) のペア

である。この二つのサイトペアは、ドメインを見てのとおり同じところが運営していると考えられるので、同様のニュースを扱っている可能性が高いと考えられる。以下、上のペアのサイトを yonhapnews、下のペアのサイトを joins と称す。

6.2　Wikipedia の検証

本手法は、単語のチェックと英単語の導出という点で大きく Wikipedia に依存している。そこで、まず Wikipedia についてのデータをまとめる。（2008/2/2 日現在）Wikipedia で扱われている言語でもっとも記事数の多い言語はやはり英語であり、その記事数は約 220 万件である。そして、今回扱う日本語の記事数は 46 万件で、韓国語の記事数は 5 万件である。英語に比べると少ないが、記事数自体はそこまで少ないわけではないといえる。そして、日本語の単語から英語の単語へ飛べる数は 203,144 件であった。この中には数字やひらがな、カタカナといった大したことのない語の記事も含まれているが 20 万の単語が翻訳できると考えれば、翻訳辞書としては十分であると考えられる。得られた日本語の単語翻訳についての例を図 6.1 に示す。次に、韓国語の単語から英語の単語へ飛べる数は 58332 件であった。総記事数より多いのは、Wikipedia には曖昧性回避という機能があり、一つの単語で複数の意味が取れるものは複数の検索結果が出現するため、一つの韓国語単語から複数の英単語へリンクされていることがあるためである。とくに数字にこの傾向が多いため、もともとの記事数があまり多くなかった韓国語にはその影響がわかりやすくでてきたものと考えられる。日本語にくらべると少ないが、約 5 万件の翻訳リンクが得られると考えれば十分に使えるものだと思われる。得られた韓国語の単語翻訳についての例を図 6.2 に示す。

鳥類 Bird
バード Bird (disambiguation)
トンチャイ・メーキンタイ Bird McIntyre
バード・オブ・プレイ Bird Of Prey (Star Trek)
渡り鳥 Bird migration
猛禽類 Bird of prey
バード・オン・ワイヤー Bird on a Wire (film)
バードストライク Bird strike
ひよこ Bird#Breeding
オオタニワタリ Bird's Nest Fern
燕の巣 Bird's nest soup
鳥瞰図 Bird's-eye view
バードキャッチャー Birdcatcher
バーディー_(ストリートファイター) Birdie (Street Fighter)
バーディ・キム Birdie Kim
バードランド Birdland (jazz club)
鳥黐 Birdlime
鳥人大系 Birdman Anthology
鳥人間コンテスト選手権大会 Birdman Rally
フウチョウ族_(Sibley) Birds of Paradise
ゴッサム・シティ・エンジェル Birds of Prey (TV series)
野鳥観察 Birdwatching
トリバネチョウ Birdwing
バーディー Birdy
鉄腕バーディー Birdy the Mighty
複屈折 Birefringence
ビレンドラ・ビール・ビクラム・シャー・デーヴ Birendra of Nepal
ビルギット・フィッシャー Birgit Fischer
ビルギット・ニルソン Birgit Nilsson
ビルギット・プリンツ Birgit Prinz
ビルカ Birka
ビルケナウ Birkenau
ビルケニア Birkenia
バーキン Birkin (surname)
ビルキルカラ Birkirkara
ビルラ Birla family
バーマン Birman
バーミンガム Birmingham
バーミンガム_(クレーター) Birmingham (crater)
バーミンガム_(曖昧さ回避) Birmingham (disambiguation)
バーミンガム・シティFC Birmingham City F.C.
バーミングハム国際空港 Birmingham International Airport (U.S.)
バーミンガム国際空港 Birmingham International Airport (United Kingdom)
バーミンガム美術館 Birmingham Museum & Art Gallery
バーミンガム・ニューストリート駅 Birmingham New Street station

図 6.1: Wikipedia を用いて日本語単語から英語単語への変換結果の例

래리_플린트_(영화) The People vs. Larry Flynt
래브라도 Labrador
래브라도_반도 Labrador Peninsula
래브라도_해 Labrador Sea
래빗_펀치 Rabbit punch
래즈베리 Raspberry
래티스_세미컨덕터 Lattice Semiconductor
래퍼해녹 Rappahannock
래퍼해녹_강_(미국) Rappahannock River
랜달_개릿 Randall Garrett
랜덤하우스 Random House
랜드_그리드_배열 Land grid array
랜드_앤_프리덤 Land and Freedom
랜드마크 Landmark
랜드마크_(동음이의) Landmark (disambiguation)
랜드마크_(홍콩) The Landmark (Hong Kong)
랜드샛_계획 Landsat program
랜드윈드 Jiangling Motors Landwind
랜디_커투어 Randy Couture
랜디와_샤론_마쉬_부부 Randy and Sharon Marsh
랜바이어푸흘귄기흘 Llanfairpwllgwyngyll
랜스_미사일 MGM-52 Lance
랜시드 Rancid (band)
랜싱 Lansing, Michigan
랜턴_(포켓몬) List of Pokémon (161-180)#Lanturn
랠리 Rallying
랠프_윌도_에머슨 Ralph Waldo Emerson
램 Random access memory
램_디스크 RAM disk
램_상주_프로그램 Terminate and Stay Resident
램버스_회의 Lambeth Conferences
램제트 Ramjet
램지_캠벨 Ramsey Campbell
램지의_정리 Ramsey's theorem
램페이지 Rampage (arcade game)
랩_(음악) Rapping

図 6.2: Wikipedia を用いて韓国語単語から英語単語への変換結果の例

次に、日本語から飛んでいける英単語と韓国語から飛んでいける英単語がどのくらい同じものであるのかを調べた。もし、これでほとんど共通する部分がなかったら使い物にならないからである。検索した結果、34157 件であった。対訳辞書として用いるとして 3 万件であれば、そこそこ使えるのではないかという期待ができる。この結果で得られた Wikipedia 上で同じ英単語へリンクされる日本語の単語と韓国語の単語を、翻訳された英単語といっしょにしたものの例を図 6.3 に示す。

また、日本語から英語を経由せずに直接韓国語に飛んだ場合はどの程度翻訳できるかを、Wikipedia のデータを入れた SQL に次の SQL コマンドを投入して調査した。

```
SELECT page.page_title , langlinks.ll_title FROM page,langlinks
WHERE langlinks.ll_from = page.page_id AND langlinks.ll_lang = "ko";
```

その結果、41456 件得られた。日本語から英語を経由せずに直接韓国語へのリンクを探ったほうが多くの翻訳が得られる結果となった。Wikipedia の翻訳リンクは手動で張られるため、途中でほかの言語を挟むとリンクが張られていない可能性が高くなるので少なくなるのは当然であるのだが、82 %しか違わないため、圧倒的な記事数を持っている英語を基準言語として使用するというのは、ほかの色々な言語と混ぜ合わせて使うときにはそこそこ妥当であると考えられる。

6.3 ラッパーの生成と適用

6.3.1 ニュースカテゴリへのリンク

yonhapnews の韓国語のサイトでニュースカテゴリの抽出で得られた結果を図 6.4 に示す。ニュースカテゴリの部分が緑色の四角で囲まれ、それからラッパーの生成とカテゴリのアンカーテキストとリンク URL 情報がうまく抽出されているのがわかる。

次に yonhapnews の日本語のサイトでニュースカテゴリの抽出で得られた結果を図 6.5 に示す。ニュースカテゴリの部分が緑色の四角で囲まれ、そこからラッパーの生成とリンク URL の情報が抽出されているのがわかる。しかし、韓国語版と違い、アンカーテキストの情報が抽出されていない。これは、このサイトではリンクに使っている DOM エレメントがテキスト情報ではなく IMG タグによるイメージによって構成されているためであり、ここからテキスト情報が抽出できなかったためである。しかし、このサイトの場合はリンク URL からそれぞれのリンクがどのようなカテゴリに属しているかがわかるため、それを用いて判断することができる。

同様にして、joins の韓国語のサイトと日本語のサイトでニュースカテゴリの抽出で得られた結果を図 6.6 と図 6.7 に示す。

Marcha Real スペインの国歌:ja,스페인의_국가:ko,
Jeolla 全羅道:ja,전라도:ko,
Natsumi Hinata 日向夏美:ja,히나타_나츠미:ko,
Eiður Guðjohnsen エイドゥル・グジョンセン:ja,아이두르_구드욘센:ko,
1638 1638년:ko,1638年:ja,
Category:33 deaths 33年没:ja,33년_죽음:ko,
Anders Hejlsberg 아네르스_하일스베르:ko,アンダース・ヘルスバーグ:ja,
Aubrey Beardsley オーブリー・ビアズリー:ja,오브리_비어즐리:ko,
8 8年:ja,8년:ko,
I (kana) い:ko,い:ja,
Category:1836 deaths 1836年没:ja,1836년_죽음:ko,
Category:Egyptian culture エジプトの文化:ja,이집트의_문화:ko,
Category:Compositions by Max Bruch ブルッフの楽曲:ja,막스_브루흐의_작품:ko,
1st millennium 제일천년기:ko,1千年紀:ja,
Gian Lorenzo Bernini ジャン・ロレンツォ・ベルニーニ:ja,잔_로렌초_베르니니:ko,
Blood plasma 血漿:ja,혈장:ko,
Category:731 deaths 731년_죽음:ko,731年没:ja,
Analogue electronics アナログ回路:ja,아날로그_회로:ko,
Category:Venezuelan culture 베네수엘라의_문화:ko,ベネズエラの文化:ja,
Cartridge (firearms) 탄약:ko,実包:ja,
Category:682 deaths 682年没:ja,682년_죽음:ko,
Main sequence 主系列星:ja,주계열성:ko,
Automaton オートマタ:ja,자동기계:ko,
Category:Computing output devices 컴퓨터_출력_장치:ko,出力機器:ja,
Category:Cities in Lithuania 리투아니아의_도시:ko,リトアニアの都市:ja,
Oz (programming language) Oz_(プログラミング言語):ja,오즈_(프로그래밍_언어):ko,
Category:Cities in Switzerland スイスの都市:ja,스위스의_도시:ko,
Category:1554 births 1554년_태어남:ko,1554年生:ja,
Reformed churches 개혁교회:ko,改革派:ja,
1973 1973년:ko,1973年:ja,
Military of the Republic of China 중화민국의_군사:ko,中華民国国軍:ja,
Archimedes アルキメデス:ja,아르키메데스:ko,
Cantabria 칸타브리아_지방:ko,カンタブリア州:ja,
Category:Transportation in Canada カナダの交通:ja,캐나다의_교통:ko,
Taitō, Tokyo 다이토_구:ko,台東区:ja,
294 294년:ko,294年:ja,
Category:817 deaths 817年没:ja,817년_죽음:ko,
Paul Kagame ポール・カガメ:ja,폴_카가메:ko,
SQ SQ:ja,SQ:ko,
Negaraku マレーシアの国歌:ja,말레이시아의_국가:ko,
Kaká 카카:ko,リカルド・イゼクソン・ドス・サントス・レイチ:ja,
Halmahera ハルマヘラ島:ja,할마헤라_섬:ko,
Template:Doi Doi:ja,Doi:ko,
Yū Kobayashi 小林ゆう:ja,고바야시_유:ko,

図 6.3: Wikipedia を用いて同じ英単語へリンクされている日本語の単語と韓国語の単語の例

得られたラッパー
HTML/BODY/DIV/TABLE/TBODY/TR/TD/TABLE/TBODY/TR/TD/A
HTML/BODY/DIV/TABLE/TBODY/TR/TD/TABLE/TBODY/TR/TD
抽出されたアンカーテキストとリンクURL
연합속보:http://www.yonhapnews.co.kr/bulletin/0200000001.html
경제:http://www.yonhapnews.co.kr/economy/index.html
증권:http://www.yonhapnews.co.kr/stock/index.html
정치/민족뉴스:http://www.yonhapnews.co.kr/politics/index.html
국제:http://www.yonhapnews.co.kr/international/index.html
사회:http://www.yonhapnews.co.kr/society/index.html
전국:http://www.yonhapnews.co.kr/local/index.html
문화:http://www.yonhapnews.co.kr/culture/index.html
스포츠:http://www.yonhapnews.co.kr/sports/index.html
연예:http://www.yonhapnews.co.kr/entertainment/index.html

図 6.4: yonhapnews 韓国語のサイトのニュースカテゴリへのラッパー生成

得られたラッパー
HTML/BODY/TABLE/TBODY/TR/TD/TABLE/TBODY/TR/TD/A/IMG
HTML/BODY/TABLE/TBODY/TR/TD/TABLE/TBODY/TR
抽出されたアンカーテキストとリンクURL
:http://japanese.yonhapnews.co.kr/headline/0200000001.html
:http://japanese.yonhapnews.co.kr/Politics2/0900000001.html
:http://japanese.yonhapnews.co.kr/northkorea/0300000001.html
:http://japanese.yonhapnews.co.kr/relation/0400000001.html
:http://japanese.yonhapnews.co.kr/economy/0500000001.html
:http://japanese.yonhapnews.co.kr/society/0800000001.html
:http://japanese.yonhapnews.co.kr/Locality/3000000001.html
:http://japanese.yonhapnews.co.kr/itscience/0600000001.html
:http://japanese.yonhapnews.co.kr/sports/0700000001.html

図 6.5: yonhapnews 日本語のサイトのニュースカテゴリへのラッパー生成

抽出されたWrapper
undefined/HTML/BODY/DIV/DIV/UL/LI/A
undefined/HTML/BODY/DIV/DIV/UL
抽出されたアンカーテキストとURL
최신기사:http://news.joins.com/list/total_list01.html?cloc=home|top|list
경제:http://news.joins.com/money/?cloc=home|top|list
사회:http://news.joins.com/life/?cloc=home|top|life
정치:http://news.joins.com/politics/?cloc=home|top|politics
지구촌:http://news.joins.com/world/?cloc=home|top|world
문화:http://news.joins.com/culture/?cloc=home|top|culture
IT과학:http://news.joins.com/infotech/?cloc=home|top|infotech
사설 칼럼:http://news.joins.com/opinion/?cloc=home|top|opinion
핫이슈:http://news.joins.com/issue/?cloc=home|top|issue
화제:http://news.joins.com/list/topic_list01.html?cloc=home|top|topic
J-only:http://service.joins.com/asp/ijonly_list.asp
보도자료:http://press.joins.com/?cloc=home|top|press

図 6.6: joins 韓国語のサイトのニュースカテゴリへのラッパー生成

抽出されたWrapper
HTML/BODY/DIV/DIV/DIV/UL/LI/A
HTML/BODY/DIV/DIV/DIV/UL
抽出されたアンカーテキストとURL
記事一覧:http://japanese.joins.com/list/list.php
経済 :http://japanese.joins.com/biz/
日本・国際:http://japanese.joins.com/nihon/
北朝鮮・政治:http://japanese.joins.com/nk/
社会・文化:http://japanese.joins.com/cul/
スポーツ:http://japanese.joins.com/spo/
社説・コラム:http://japanese.joins.com/list/list.php?servcode=100
フォトニュース:http://japanese.joins.com/etc/photonews/
動画:http://japanese.joins.com/etc/medianews/
ニュース特集:http://japanese.joins.com/series/

図 6.7: joins 日本語のサイトのニュースカテゴリへのラッパー生成

6.3.2 ニュースカテゴリの抽出

yonhapnews と joins において、ラッパーによるニュースカテゴリの抽出の結果、カテゴリの対応付けができたものを次に示す。yonhapnews については、次の 4 つのカテゴリが対応付けされた。

- Economy
  http://japanese.yonhapnews.co.kr/economy/0500000001.html
  http://www.yonhapnews.co.kr/economy/index.html
- Politics
  http://japanese.yonhapnews.co.kr/Politics2/0900000001.html
  http://www.yonhapnews.co.kr/politics/index.html
- Society
  http://japanese.yonhapnews.co.kr/society/0800000001.html
  http://www.yonhapnews.co.kr/society/index.html
- Sports
  http://www.yonhapnews.co.kr/sports/index.html
  http://japanese.yonhapnews.co.kr/sports/0700000001.html

日本語のサイトには 8 つのカテゴリ、韓国語のサイトには 9 つのカテゴリがあるので、約半分が抽出されているのがわかる。

joins については、次の 3 つのカテゴリが対応付けされた。

- Economy
  http://news.joins.com/money/
  http://japanese.joins.com/biz/
- Politics
  http://news.joins.com/politics/
  http://japanese.joins.com/nk/
- Sports
  http://news.joins.com/list/sports_list01.html
  http://japanese.joins.com/spo/

日本語のサイトには 6 つのカテゴリ、韓国語のサイトには 8 つのカテゴリがあるので、約半分が抽出されているのがわかる。

カテゴリの対応付けには、Economy や Politics、Sports などの良くある単語での一致によって決められるというのがわかった。国際関係のカテゴリは international や world などで、科学関係は science や technology、tech などで表記ゆれがあり、このままの手法ではうまくいかないことがわかった。手動で対応つけるか、これらの表記ゆれを吸収するなんらかの手法が必要となる。

6.3.3 ニュース記事へのリンク

同様にして、次にニュース記事へのリンク部分を抽出した結果を示す。まず、yonhapnews 韓国版の経済欄と思われる

http://www.yonhapnews.co.kr/economy/index.html

においてニュース記事へのリンクを抽出した結果を図 6.8 に示す。ニュース記事へのリンクの部分が緑色か黄色の四角で囲まれ、それからラッパーの生成とニュース記事へのアンカーテキストとリンク URL 情報がうまく抽出されているのがわかる。

次に、yonhapnews 日本語版の経済欄と思われる

http://japanese.yonhapnews.co.kr/economy/0500000001.html

においててニュース記事へのリンクを抽出した結果を図 6.9 に示す。

ニュース記事へのリンクの部分が緑色か黄色の四角で囲まれ、それからラッパーの生成とニュース記事へのアンカーテキストとリンク URL 情報がうまく抽出されているのがわかる。

同様にして、joins 韓国語版と joins 日本語版の経済欄と思われる

http://news.joins.com/money/

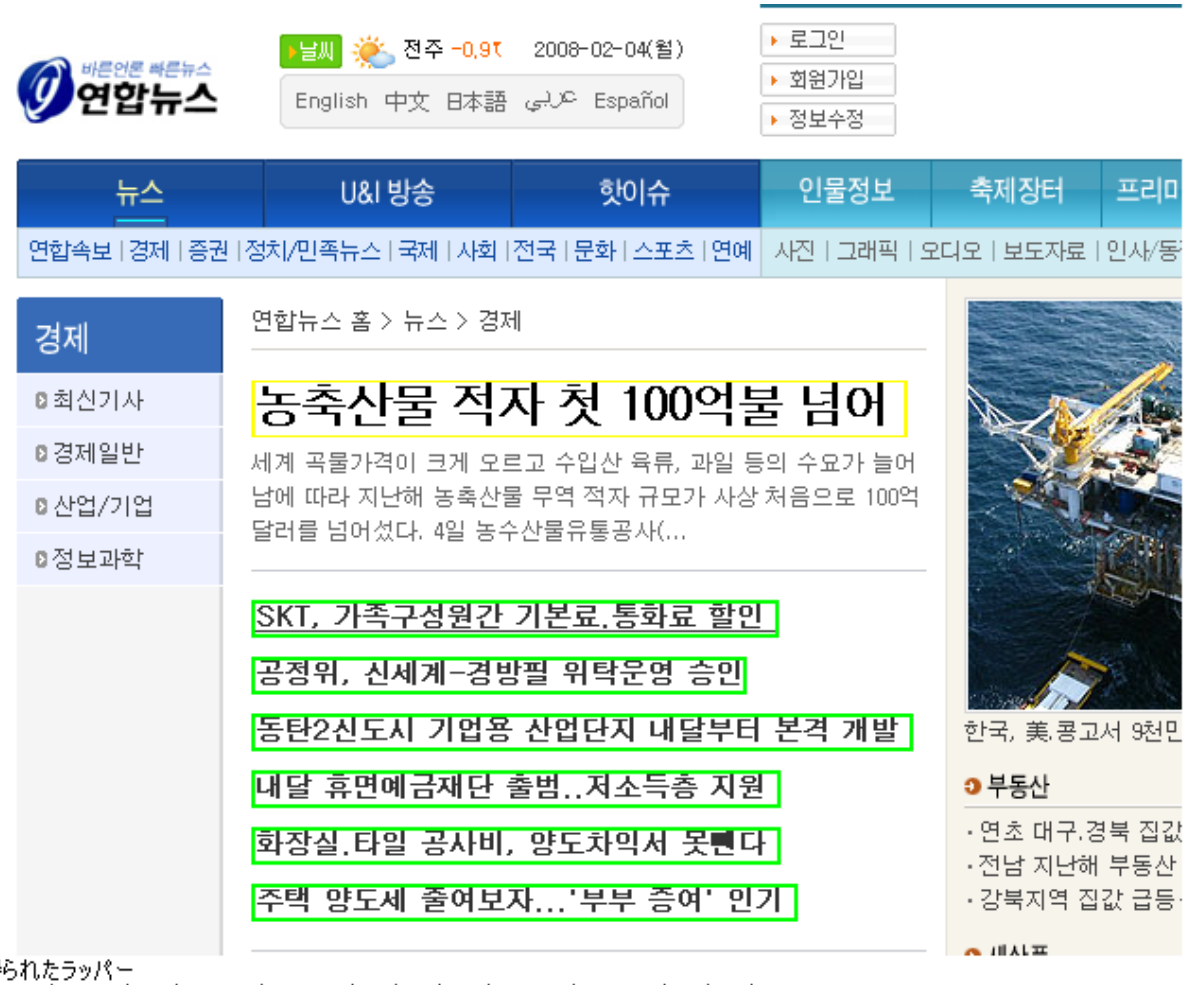
http://japanese.joins.com/biz/

についてニュース記事へのリンクを抽出した結果を図 6.10 と図 6.11 に示す。

どちらもニュース記事へのリンクの部分が緑色か黄色の四角で囲まれ、それからラッパーの生成とニュース記事へのアンカーテキストとリンク URL 情報がうまく抽出されているのがわかる。

6.3.4 結果の考察

今回行った実験では、リンクの DOM 要素として、テキスト情報を扱っていればうまくアンカーテキストを取り出せることがわかった。また、この日韓ニュースサイトでは存在しなかったが、Javascript を用いて動的に HTML ページを構築しているサイトではうまく動かないケースがあった。



得られたラッパー
HTML/BODY/DIV/TABLE/TBODY/TR/TD/DIV/TABLE/TBODY/TR/TD/A
HTML/BODY/DIV/TABLE/TBODY/TR/TD
抽出されたアンカーテキストとリンクURL
농축산물 적자 첫 100억불 넘어
:http://www.yonhapnews.co.kr/economy/2008/02/04/0309000000AKR20080202025800002.HTML
SKT, 가족구성원간 기본료.통화료 할인
:http://www.yonhapnews.co.kr/economy/2008/02/04/0303000000AKR20080204041800006.HTML
공정위, 신세계-경방필 위탁운영
승인:http://www.yonhapnews.co.kr/economy/2008/02/04/0301000000AKR20080204043800002.HTML
동탄2신도시 기업용 산업단지 내달부터 본격 개발
:http://www.yonhapnews.co.kr/economy/2008/02/04/0301000000AKR20080202000900003.HTML
내달 휴면예금재단 출범..저소득층 지원
:http://www.yonhapnews.co.kr/economy/2008/02/04/0301000000AKR20080130213200048.HTML
화장실.타일 공사비, 양도차익서 못뺀다
:http://www.yonhapnews.co.kr/economy/2008/02/04/0301000000AKR20080201206700002.HTML
주택 양도세 줄여보자...'부부 증여' 인기
:http://www.yonhapnews.co.kr/economy/2008/02/04/0301000000AKR20080202043600003.HTML
韓美 금리차 확대로 채권시장에 '핫머니' 홍수
:http://www.yonhapnews.co.kr/economy/2008/02/04/0310000000AKR20080202045400008.HTML
"국가별.지역별 대외통상정책 수립해야"< KIEP >
:http://www.yonhapnews.co.kr/economy/2008/02/04/0301000000AKR20080202038800002.HTML
1월 외환보유액 3억5천만달러 감소
:http://www.yonhapnews.co.kr/economy/2008/02/04/0301000000AKR20080201199200048.HTML
식약청, 포도주 발암성 물질 기준 마련
:http://www.yonhapnews.co.kr/economy/2008/02/04/0303000000AKR20080203035500006.HTML

図 6.8: yonhapnews 韓国語のサイトのニュース記事へのラッパー生成

得られたラッパー
HTML/BODY/TABLE/TBODY/TR/TD/TABLE/TBODY/TR/TD/TABLE/TBODY/TR/TD/A
HTML/BODY/TABLE/TBODY/TR/TD/TABLE/TBODY/TR/TD/TABLE/TBODY
抽出されたアンカーテキストとリンクURL
生活経済苦痛指数が昨年10月から急騰、LG経済研:http://japanese.yonhapnews.co.kr/economy/2008/02/03/0500000000AJP20080203000700882.HTML
韓国人の日本訪問、韓国訪問日本人数を初めて超える:http://japanese.yonhapnews.co.kr/economy/2008/02/03/0500000000AJP20080203000400882.HTML
現代自インド第2工場完工、年産規模60万台に :http://japanese.yonhapnews.co.kr/economy/2008/02/03/0500000000AJP20080203000500882.HTML
1月の海外建設受注53億ドル、昨年の2倍に:http://japanese.yonhapnews.co.kr/economy/2008/02/03/0500000000AJP20080203000100882.HTML
外換銀の売却承認問題は引き続き留保、金融監督委:http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201003700882.HTML
1月の完成車販売台数は17%増、現代自がけん引:http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201003300882.HTML
1月の消費者物価3．9%上昇、生活物価は5．1%:http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201003200882.HTML
総合株価指数(1日)1634．53ポイント:http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201002900882.HTML
ウォン・ドル相場(1日)944．20ウォン:http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201002800882.HTML
EUとのFTA交渉、中核争点を除く7割が妥結 :http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201001700882.HTML
石油公社とサムスン物産、メキシコ湾の油田を確保:http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201002100882.HTML
1月輸入額が過去最大、貿易赤字34億ドルに拡大:http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201001600882.HTML
ハイニックス、Q4営業損失3180億ウォン:http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201001000882.HTML
現代自、米NFLスーパーボウル生中継でCM放映:http://japanese.yonhapnews.co.kr/economy/2008/02/01/0500000000AJP20080201000500882.HTML
航空2社の昨年業績、増収増益も純利益は減少:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131004000882.HTML
EUとのFTA交渉、知的財産権で歩み寄り気配:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131003900882.HTML
開城工業団地の通行時間拡大、施行後数時間で中止:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131003800882.HTML
現代重工業が過去最高の業績、売上高15兆ウォン:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131003500882.HTML
12月の産業生産、前年同月比12．4%増:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131003100882.HTML
ネットバンキングの利用件数、昨年は40%急増:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131002800882.HTML
総合株価指数(31日)1624．68ポイント:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131003000882.HTML
ウォン・ドル相場(31日)943．70ウォン:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131002900882.HTML
銀行界がカザフに注目、新韓銀・国民銀が進出:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131001700882.HTML
SKT昨年の営業利益16%減、売上・純利益は増加:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131001400882.HTML
急激な経済萎縮の可能性低い、財政経済部次官:http://japanese.yonhapnews.co.kr/economy/2008/01/31/0500000000AJP20080131000700882.HTML

図 6.9: yonhapnews 日本語のサイトのニュース記事へのラッパー生成

抽出されたWrapper
HTML/BODY/DIV/DIV/DL/DT/A
HTML/BODY/DIV/DIV
抽出されたアンカーテキストとURL
[머니브리핑] 2월 13일:http://news.joins.com/article/3038862.html?ctg=11
[분양Memo] 인천 서구 검단 힐스테이트 2차 外:http://news.joins.com/article/3038852.html?ctg=11
[REALESTATE] 택지지구 땅 투자 문턱 높아진다:http://news.joins.com/article/3038851.html?ctg=11
[REALESTATE] 택지지구 상가 살 사람 볼 수 없…:http://news.joins.com/article/3038850.html?ctg=11
[REALESTATE] 'MB 시대' 투자전략 궁금하다면…:http://news.joins.com/article/3038849.html?ctg=11
[분양 하이라이트] 대전 서남부신도시 '엘드 수목토'&#…:http://news.joins.com/article/3038848.html?ctg=11
[REALESTATE] 다세대·연립 살 사람 셀 수 없네:http://news.joins.com/article/3038847.html?ctg=11
[동정] 윤동한 한국콜마 회장:http://news.joins.com/article/3038846.html?ctg=11
[인사] 하나대투증권 外:http://news.joins.com/article/3038845.html?ctg=11

図 6.10: joins 韓国語のサイトのニュース記事へのラッパー生成

抽出されたWrapper
HTML/BODY/DIV/DIV/DIV/UL/LI/A
HTML/BODY/DIV/DIV/DIV/UL
抽出されたアンカーテキストとURL
原油流出で成果給消える… 三星重工業:http://japanese.joins.com/article/article.php?aid=95815&servcode=300&sectcode=300
空港貴賓室が‘企業の戦場’に？:http://japanese.joins.com/article/article.php?aid=95725&servcode=300&sectcode=300
1月の物価上昇率は3.9%…40カ月ぶり最高水準:http://japanese.joins.com/article/article.php?aid=95678&servcode=300&sectcode=300
高齢化社会に向けて‘快適’再建築を準備中:http://japanese.joins.com/article/article.php?aid=95668&servcode=300&sectcode=300
GM大宇、「マティス」最大53万ウォン値下げへ:http://japanese.joins.com/article/article.php?aid=95666&servcode=300&sectcode=300
世界は‘ヒューマン新都市’建設中…東京・多摩市:http://japanese.joins.com/article/article.php?aid=95665&servcode=300&sectcode=300
携帯大手3社「傷だけの競争」:http://japanese.joins.com/article/article.php?aid=95653&servcode=300&sectcode=300

図 6.11: joins 韓国語のサイトのニュース記事へのラッパー生成

6.4 ニュースの関連付け

6.4.1 言語横断的な関連付け

上記のサイトにおいて、2007/12/1 から 2007/12/31 までの 31 日間分のデータを実際に処理した結果を述べる。まず、yonhapnews において、カテゴリーへのリンク部分を抽出してそこから同じカテゴリーであると簡単にわかるものとして economy,society,politics の 3 つが存在した。そこで、この 3 つのカテゴリーを対象にしてニュース記事の関連付けを行うことにした。

まず、economy について処理した結果を述べる。このカテゴリでは、日本語記事数 232 個、韓国語記事数 368 個であった。そのうち、リストにでてきた TOP20 のうち、8 個が実際に人手で確認してかなり近いトピックについて書いてあると思われるものが見つかった。それを

韓国語タイトル、
（スペース）Excite で韓国語を日本語に翻訳した結果、
韓国語タイトルの英単語ベクトル、
日本語タイトル、
日本語タイトルの英単語ベクトル、
共通してでてきた英単語

という順番で上から順に挙げていったものを図 6.12 に示す。

この結果から、ほとんどのタイトルの関連性の判定は数文字の英語単語か数字に大きく影響されることがわかる。

また、あまり関係しないと思われるものもかなりの数見つかった。それらはほとんどニュース記事を表しているとはいえない単語で関連していると見なされたもので、それらの単語は主に、Gold,Year,Dollar,Moon、それと一桁の数字であった。これらは、日本語でいうと金、年、月、ドルと言ったものであるので、ほかの語といっしょに判断して初めて関連しているかどうかわかるという類のものであった。とくに経済欄では株価や株価指数、為替の値で多くの数字が出てくるのでそれらで誤認定されることが多くあることがわかった。

次に、politics について処理した結果について述べる。politics カテゴリには日本語記事数 269 個、韓国語記事数 60 個であった。economy と同様に、TOP20 のなかにかなり近いと思われるものが 7 個見つかったので図 6.13 に示す。

これらの結果では、第 17 代目ということや、大統領、個人名、BBK という問題といったニュースを見るうえで重要なキーワードが含まれているかどうかというものが指標になった。

また、得られた結果を眺めることで、この時期の政治欄では BBK 問題と第 17 代大韓民国大統領に就任した李明博が世間で話題になっていることがよくわかった。

次に、society について処理した結果について述べる。このカテゴリには日本語記事数 265 個、韓

韓-印 CEPA 협상, 상품양허 등 '목표수준' 근접
韓-印 CEPA 交渉, 商品譲歩など '目標数与えた' 近接
[-, C, E, P, A, Domestic sheep, ' (disambiguation), Neck, Zune, ' (disambiguation), C,CE, EP, PA, A, Product (business), ' (disambigu
(disambiguation), CE, CEP, EPA, Pa, Product (business), Cep, CEPA]
インドとのCEPA締結、18日から9回目の交渉
[To (kana), No (kana), C, E, P, A, 1, 8, Day, Ka (kana), Ra (kana), 9, Eye, No (kana),
CE, EP, PA, 18, India, CEP, EPA, CEPA]
[Pa, Cep, CEPA]

현대차, 러시아에 연산 10만대 규모 공장 건설
現代車, ロシアに演算 10万台規模工場建設
[CHA, Kite, Mountain, 1, 0, Headlands and bays, Hyundai (disambiguation), Bogie, Kite, Mountain, 1 (number), 10, Factory, Constru
10 (number), Factory, Factory, Construction, Construction, Russia, Factory, Construction]
現代自がロシアに完成車工場新設、年産10万台規模
[Ni (kana), Field (physics), Xin Dynasty, Year, 1, 0, Hyundai (disambiguation), Factory, 10, Russia]
[Hyundai (disambiguation), 10, Factory, Russia, Factory, Factory, Russia, Factory]

내년 수출 4천130억달러, 11.4% 증가 전망
来年輸出 4千130億ドル, 11.4% 増加見込み
[Year, 4, 1, 3, 0, Moon, 1, 1, 4, Year, 4 (number), 13, 30, Dollar, 1 (number), 11, 1., .4, 130, 11 (number), 11., 1.4, 11.4]
来年の輸出見通し、4130億ドルで11．4%増
[Year, No (kana), Shi (kana), 4, 1, 3, 0, 100000000 (number), 1, 1, 4, Export, 41, 13, 30, Dollar, 11, 1., .4, 413, 130, 11., 1.4, 4130,
[Year, Year, Dollar, 130, 11.4]

美 11월 소비자지출 3년래 최대 증가
美 11 月消費者支出 3年来最大増加
[1, 1, Cattle, 3, Year, 1 (number), 11, January, Cattle, Consumption (economics), Visa, 3 (number), 3, 11 (number), November, Ja
Consumption (economics), Consumer, 3, November, November, Consumer, November]
11月の消費者物価、3年来最高の3．5%上昇
[1, 1, Moon, No (kana), 3, Year, No (kana), 3, 5, 11, Category:January, Consumption (economics), Price index, 3, 3., .5, Category:
Consumer, 3.5]
[3, 3, 3, Year, 11, Consumption (economics), 3, 3, 3, Consumption (economics), Consumer, 3, 3, 3, Consumer]

개인 부채 700조 돌파..1인당 1천477만원
個人負債 700兆突破..1人当り 1千477万ウォン
[Dog, Phosphorus, 7, 0, 0, 1, Phosphorus, 1, 4, 7, 7, Headlands and bays, Phosphorus, 7 (number), 70, 00, ..., .1, 1 (number), 47, 77
700, ..1, 477, 700 (number)]
個人負債700兆ウォン突破、1人1477万ウォン
[Person, 7, 0, 0, trillion, U (disambiguation), 1, Person, 1, 4, 7, 7, U (disambiguation), Individual, Liability, 70, 00, 14, 47, 77, 700, K
147, 477, Korean won, 1477]
[1, 1, 1, 1, 700, 477]

나라 재산 6천조원..10년새 2배로
国財産 6千兆ウォン..10年の間 2倍で
[Ash (analytical chemistry), Mountain, 6, 1, 0, Year, Bird, 2, Country, Ash (analytical chemistry), Estate (law), Mountain, 6 (numbe
Year zero, Bird, 2 (number), Country, Estate (law), Estate (law), ..1, .10, 10, Estate (law), ..10]
国の財産6000兆ウォン、10年間で2倍に
[State (disambiguation), No (kana), Good (economics and accounting), 6, 0, 0, 0, trillion, U (disambiguation), 1, 0, Year, 2, Ni (kana
60, 00, 00, 10, Year zero, 600, 000, Korean won, 10, 6000]
[6, 2, 10, 10, Year zero, 10, 10]

10만원권 보조도안에 무궁화.대동여지도
10万ウォン券補助図案にムクゲ.テドンヨ地図
[1, 0, Headlands and bays, 10, Illuminance, Map, Hibiscus syriacus, Map, Hibiscus syriacus]
高額紙幣の補助図案に、ムクゲや大東輿地図など
[Forehead, Paper, No (kana), Ni (kana), Ya (kana), East, Land, Na (kana), Banknote, Map, Hibiscus syriacus]
[Map, Hibiscus syriacus, Map, Hibiscus syriacus]

거침없는 해외건설 수주..올해 380억달러 전망
障りない海外建設受注..今年 380億ドル見込み
[3, 8, 0, Moon, Construction, ..., 3 (number), 38, 80, Dollar, Construction, 38 (number), 380, Dollar, 380 (number)]
海外建設受注先月までで過去最高の355億ドル
[Sea, Moon, Ma (kana), No (kana), 3, 5, 5, 100000000 (number), Overseas, Construction, Past, 35, 55, Dollar, 355]
[3, Moon, Construction, Dollar, Construction, Dollar]

図 6.12: yonhapnews 経済欄における日本語と韓国語のタイトルでほとんど同じトピックを持つもの

鄭 "'반부패연대'로 선거혁명"
鄭 "'反腐敗連帯'で選挙革命"
[' (disambiguation), Kite, ' (disambiguation), Election, Revolution, Election]
反腐敗連帯で選挙革命を、鄭東泳候補が呼びかけ
[Muraji, Obi (sash), Wo (kana), Zheng (state), East, Ka (kana), Ke (kana), Decomposition, Election, Revolution, Ke (kana), Chung D
(kana)]
[Election, Revolution, Election]

李당선자 " 선진화 원년 다 함께 열어가자"
李当選者 " 先進化元年もろともに開いて行こう"
[Year, Heat, Evolution, Year, Heat, Evolution]
「来年は先進化元年」李明博当選者が新年の辞
[Year, Ha (kana), Yuan, Year, Ming Dynasty, Xin Dynasty, Year, No (kana), Evolution, Lee Myung-bak]
[Year, Year, Year, Evolution, Year, Year, Year, Evolution]

'BBK 검사 탄핵소추안' 본회의 보고
'BBK 検事弾劾訴追案' 本会議報告
[' (disambiguation), B, B, K, Nucleus, Cattle, ' (disambiguation), Bonn, Hoe (dish), BB, BK, K, Impeachment, ' (disambiguation), Bon
Impeachment]
BBK事件担当検事に対する弾劾訴追案、新党が提出
[B, B, K, Ni (kana), Pair, Su (kana), Ru_(kana), Xin Dynasty, BB, BK, Impeachment, New Party, BBK]
[Impeachment, BBK, Impeachment]

이명박 제17대 대통령 당선
李明博第17代大統領当選
[1, 7, 17, Lee Myung-bak, President, Lee Myung-bak, President, President, President]
第17代大統領に李明博氏、10年ぶりの政権交代
[1, 7, Ni (kana), Ming Dynasty, 1, 0, Year, Ri (kana), No (kana), 17, 10, Year zero, Regime, President, Lee Myung-bak, 10]
[17, Lee Myung-bak, President, Lee Myung-bak, President, President, President]

부시 대통령,김정일에 친서 보내
ブッシュ大統領,金正日に親書送って
[Gim (Korean food), Bush, Bush, President, Kim Jong-il, President]
「義務果たす」ブッシュ大統領の親書に北朝鮮が返答
[Ta (kana), Su (kana), No (kana), Ni (kana), North, Morning, Duty, Korea, President, Bush]
[Bush, Bush, President, President]

사면.복권 김우중.박지원.한화갑 포함될 듯
赦免.復権キム・ウジュン.パク・チウォン.ハン・ファガップ含まれるよう
[Gim (Korean food), Han, Lottery, Gim (Korean food), Lottery, Kim Woo-jung, Kim Woo-jung]
年末特別赦免、金宇中・朴智元氏ら含まれる見通し
[Year, Gold, Naka, Yuan, Ra (kana), Ma (kana), Re_(kana), Ru_(kana), Shi (kana), Kim Woo-jung]
[Kim Woo-jung, Kim Woo-jung]

鄭-文 단일화 사실상 '무산'
鄭-文 一本化事実上 '無産'
[-, Yarn, ' (disambiguation), Mountain, ' (disambiguation), Fact, ' (disambiguation), ' (disambiguation), Fact]
鄭・文候補一本化事実上白紙、劇的反転の可能性残し
[Zheng (state), Sentence (linguistics), Book, White, Paper, No (kana), Noh, Shi (kana), Ippon, Fact]
[Fact, Fact]

図 6.13: yonhapnews 政治欄における日本語と韓国語のタイトルでかなり近いトピックを持つもの

国語記事数 317 個あった。eeconomy と同様に、TOP20 のなかにかなり近いと思われるものが 7 個見つかったので図 6.14 に示す。

この結果からも、同じトピックであろうとわかるものは数字に依存する部分が多いというのがわかった。また、個人名や英単語があると比較的合致させやすいというのもわかった。

つぎに、joins において、カテゴリの分類が分かるものとして、Economy と Politics があるので、これらに適用した結果について述べる。まず、economy について処理した結果について述べる。このカテゴリには日本語記事数 96 個、韓国語記事数 670 個あった。このうち、TOP20 のなかにかなり近いと思われるものは 5 個見つかったので図 6.15 に示す 。

politics カテゴリについては、日本語記事数が 167 個、韓国語記事数が 886 個であった。このうち、TOP20 のなかにかなり近いと思われるものは 4 個見つかったので、図 6.16 に示す。

6.4.2 考察

日本語から得られた英単語列と韓国語から得られた英単語列を比べて、共通する英単語が多いものから順に取り出すという手法である程度、関連しているニュースを取り出す手掛かりにすることはできた。しかし、関連しているニュースであっても、共通する単語が訳すまでもなく両方の文中に出現する短い英単語や数字列であったり、特定の個人名や少数の一致単語のみでしか判定できないため、Year、Month、Dollar といったニュース中に多く出現するが関連付けにはあまり寄与しないものや、似たような数字列による誤認定の数が多く、かなりのゴミが混ざってしまっている。すべて混ぜこぜになっている中から関連しているニュースを見つけるよりは手間が減ってはいるが、以前この結果の中から確実に関連しているニュースを取り出すのは人の手に頼る部分が多く、手間のかかるものになっているのが現状である。

2020년 농가인구 10명중 6명 60대 이상
2020年農家人口 10人の中 6人 60代以上
[2, 0, 2, 0, Year, Phosphorus, 1, 0, 6, 6, 0, 20, 02, 20, Year zero, Year, Population, 1 (number), 10, 6 (number), 6 (number), 60, 20
Year zero, Population, 10 (number),
60 (number), 2020, 20, 2020]
2020年には農家人口の6割が高齢者に、統計協会
[2, 0, 2, 0, Year, Ni (kana), Ha (kana), Person, Mouth, No (kana), 6, Ni (kana), 20, 02, 20, Year zero, Farmer, Kenin (Japanese feuda
Population, statistic, 202, 020, 20,
Old_age, 2020, 2020]
[6, 6, Year zero, Population, Year zero, Population, 2020, 2020, 2020, 2020]

한국 15세학생 학업성취도, OECD 5-9위
韓国 15セハックセング学業成就も, OECD 5-9位
[Han, 1, 5, O, E, C, D, 5, -, 9, Wei, Korea, 1 (number), 15, Student, O, OE, EC, CD, D, 5 (number), Wei, Korea, 15 (number), Studer
ECD, CD (disambiguation), OECD]
OECD学習到達度調査、韓国は加盟国中5～9位
[O, E, C, D, Degree, Key (music), Han (state), State (disambiguation), Ha (kana), State (disambiguation), Naka, 5, 9, OE, EC, CD, Le
ECD, OECD]
[5, 5, 9, Oe, OECD]

"500대 기업 CEO 출신, 서울대가 32위"
"500大企業 CEO 出身, ソウル大が 32位"
[5, 0, 0, C, E, O, Deity, 3, 2, Wei, 50, 00, C, CE, EO, O, 3 (number), 32, 500, CE, CEO, EO, 32 (number), CEO, CEO, CEO]
500大企業CEOの出身大学、ソウル大は32位
[5, 0, 0, Karma, C, E, O, No (kana), U (disambiguation), Ha (kana), 3, 2, 50, 00, Company (law), CE, EO, University, Saw (film), Ur, 32
[32, 500, CEO, CEO, CEO, CEO]

"아르바이트 청소년, 체불 · 욕설 · 폭력 시달려"
"バイト青少年, 滞払・悪口・暴力苦しんで"
[Bar, Cattle, Year, Fire, Moon, Are, Boy, Byte, Byte, Part time]
問題多い青少年アルバイト環境、給与未払いや悪口
[I (kana), Blue, Year, Sheep (zodiac), I (kana), Ya (kana), Evil, Mouth, Boy, al, Environment, Salary, Aruba, Part time]
[Year, Boy, Part time]

김우중 · 박지원 · 한화갑 등 75명 특별사면
キム・ウジュン・パク・チウォン・ハン・ファガップなど 75人特別社なら
[Gim (Korean food), Han, 7, 5, 7 (number), 75, Kim Woo-jung, 75 (number)]
年末の特別赦免、対象者に金宇中氏ら75人
[Year, No (kana), Pair, Ni (kana), Gold, Naka, Ra (kana), 7, 5, Person, 75, Kim Woo-jung]
[75, Kim Woo-jung]

인기가수 싸이 현역으로 입대할 듯
人気歌手サイ現役に入隊するよう
[Phosphorus, Mouth, Giga-, Singer, PSY, Mouth, Singer, PSY, PSY, PSY]
歌手のPSYさん、きょう陸軍訓練所に入所
[Song, Hand, No (kana), P, S, Y, Sa (kana), N (kana), Ki (kana), U (kana), Ni (kana), Singer, PS, SY, Army, Training, PSY]
[Singer, PSY, Singer, PSY, PSY, PSY]
4.32273600446995

정부-공노총, 공무원 노동절 휴무 검토
政府-ゴングノーチョング, 公務員メーデー休務検討
[-, Government, May Day, May Day, May Day, May Day]
公務員のメーデー休日化、政府と労組が暫定合意
[No (kana), Day, To (kana), Shibaraku, Gō (volume), Day (disambiguation), Vacation, Government, Consensus, Civil service, Mede
Day]
[Government, May Day, May Day, May Day, May Day]

図 6.14: yonhapnews 社会欄における日本語と韓国語のタイトルでかなり近いトピックを持つもの

“32인치 PDP 잘 팔리네” LG 소형 인기없다 통념 깨
“32インチ PDP よく売れるのね” LG 小型人気ない通念ごま
[3, 2, Phosphorus, P, D, P, Arm, Li (unit), L, G, Cattle, Phosphorus, 32, Inch, P, PD, DP, L, LG, Cattle, Phosphorus, PDP, LG (disambiguation)]
LG「32インチPDPが販売好調」… 月20万台
[L, G, 3, 2, P, D, P, Key (music), Moon, 2, 0, LG, 32, PD, DP, Sales, 20, Inch, PDP]
[32, Inch, LG, PDP]

한국 여성 근로시간 OECD 중 최장
韓国女性勤労時間 OECD の中で最長
[Han, Liver, O, E, C, D, Korea, Woman, Time, Liver, O, OE, EC, CD, D, Korea, Woman, Woman, Time, Oe, OEC, ECD, CD (disambigu
Woman, OECD]
韓国女性の勤務時間 OECDで最長
[Han (state), State (disambiguation), No (kana), O, E, C, D, Woman, Category:Time, OE, EC, CD, OEC, ECD, OECD]
[Woman, Woman, Woman, Oe, Woman, OECD]

[글로벌IT] 일본, 아이팟 못 만든 이유는 …
[グローバルIT] 日本, アイパッ作る事ができなかった理由は …
[I, T, Bonn, Nail (fastener), Headlands and bays, IT, Japan, Nail (fastener), Headlands and bays, Japan, IPod]
＜グローバルIT＞日本、iPodを作れない理由は…
[I, T, Day, Book, i, P, o, d, Wo (kana), Re_(kana), Na (kana), I (kana), Ha (kana), Rho (Italy), IT, Japan, iP, Po, od, iPo, Pod, iPod]
[IT, Japan, Japan, IPod]

[글로벌IT] 일본, 아이팟 못 만든 이유는 …
[グローバルIT] 日本, アイパッ作る事ができなかった理由は …
[I, T, Bonn, Nail (fastener), Headlands and bays, IT, Japan, Nail (fastener), Headlands and bays, Japan, IPod]
＜グローバルIT＞日本、iPodを作れない理由は…
[I, T, Day, Book, i, P, o, d, Wo (kana), Re_(kana), Na (kana), I (kana), Ha (kana), Rho (Italy), IT, Japan, iP, Po, od, iPo, Pod, iPod]
[IT, Japan, Japan, IPod]

2012년 국민소득 3만7000달러 될 것
2012年国民所得 3万7000ドルになること
[2, 0, 1, 2, Year, Cattle, 3, Headlands and bays, 7, 0, 0, 0, Moon, 20, 01, 12, 2, Year, Nation, 3 (number), 70, 00, 00, Dollar, 201, 01
Nation, 700, 000, Dollar, 2012, 7000,
2012]
「韓国の国民所得、2012年には日本水準に」
[Han (state), State (disambiguation), No (kana), State (disambiguation), 2, 0, 1, 2, Year, Ni (kana), Ha (kana), Day, Book, Water, Ni (
Nation, 20, 01, 12, 2, Japan, 201, 012
, 12, Measures of national income and output, 2012, 2012]
[Year, Year, Nation, Nation, 2012, 2012, 2012, 2012]

図 6.15: joins 経済欄における日本語と韓国語のタイトルでかなり近いトピックを持つもの

[Joins풍향계] "차기 정부의 '한반도 대운하' 건설…
[Joins風向計] "次期政府の '韓半島大運河' 建設…
[J, o, i, n, s, CHA, ' (disambiguation), Han, ' (disambiguation), Jo, oi, in, ns, Government, ' (disambiguation), Peninsula, Canal, '
(disambiguation), Construction, Joi, oin, ins,
Government, Korean Peninsula, Grand Canal, Construction, Join, oins, Joins]
＜Joins風向計＞「次期政府の韓半島大運河建設に賛成」42．3%
[J, o, i, n, s, Category:Winds, No (kana), Han (state), Island, Luck, Ni (kana), 4, 2, 3, Jo, oi, in, ns, Wind direction, Government, Pe
Canal, Construction, Template:Suppo
rt, 42, 2., .3, Joi, oin, ins, Grand Canal of China, 42., 2.3, Join, oins, 42.3, Joins]
[Government, Peninsula, Canal, Construction, Government, Construction, Joins]

"반드시 투표" 67%…5년전 대비 13.5%p↓
"必ず投票" 67%…5年前備え 13.5%p
[6, 7, 5, Year, 1, 3, 5, p, Voting, 6 (number), 67, 5, 1 (number), 13, 3., .5, Voting, 13., 3.5, 13.5]
「必ず投票」67%…5年前より13．5%↓
[6, 7, 5, Year, Yo (kana), Ri (kana), 1, 3, 5, Voting, 67, 5, 13, 3., .5, 13., 3.5, 13.5]
[Year, Voting, 67, Voting, 13.5]

[Joins풍향계] 국민 75.7% "이명박 당선자 집…
[Joins風向計] 国民 75.7% "李明博当選者家…
[J, o, i, n, s, 7, 5, 7, House, Jo, oi, in, ns, Nation, 7 (number), 75, 5., .7, House, Joi, oin, ins, Nation, Nation, 75 (number), 75., 5.7,
-bak, Join, oins, 75.7, Lee My
ung-bak, Joins]
＜Joins風向計＞国民75．7%「李氏が政権取れば国家的利得」
[J, o, i, n, s, Category:Winds, State (disambiguation), 7, 5, 7, Re_(kana), State (disambiguation), Jo, oi, in, ns, Wind direction, Natio
Li (李), Regime, Country, Ga
in, Joi, oin, ins, 75., 5.7, Join, oins, 75.7, Joins]
[Nation, Nation, Nation, 75.7, Joins]

김흥국 윤형주 등 가수 50여명 李 지지
キム・フングク輪形株など歌手 50人余り 李 支持
[Gim (Korean food), 5, 0, Singer, 5 (number), 50, Singer, 50 (number)]
歌手50人余「李明博氏支持」宣言
[Song, Hand, 5, 0, Person, Ming Dynasty, Singer, 50, Lee Myung-bak]
[Singer, 50, Singer]

図 6.16: joins 政治欄における日本語と韓国語のタイトルでかなり近いトピックを持つもの

第7章　結論

7.1　本論文のまとめ

本論文では、Web ニュースサイトのニュース記事を言語横断的に関連付ける手法について述べた。

7.1.1　ラッパーの作成

我々の手法では、複数の言語でニュースを提供しているサイトを対象として、そのサイトにおいて同じニュースソースからニュース記事を書いていることを期待してそれぞれの言語においてラッパーを作成し、ニュースカテゴリとニュース記事へのリンクをそれぞれ抽出してきて、それらをニュース記事を対応付けるために用いる。そのため、すべてのニュースにおいて全対全で比較するよりも少ない範囲での比較をすることができ、計算量の削減になる。また、ラッパーの生成は Web ブラウザ上でインタラクティブに行うことができ、ユーザーフレンドリーになっている。

7.1.2　ニュース関連付け

我々の手法は、言語情報をほとんど用いずにニュースの関連付けをすることを目指した手法である。そのため、文章を単語で区切るときには単純な n-gram かスペースで区切ることとし、単語かどうかの判定は Web 上で今現在、そしてこれからも世界中の人が更新し続けるであろう Wikipedia を用いて行っている。また、同様に Wikipedia のある記事のページからその記事の他言語版へと飛ぶ機能を利用し、単語の翻訳を行っている。これらを用いることで、対象とする言語についての情報はほとんどもちいることなく文章の処理に成功している。言語横断的な関連付けに関しては、対象とする言語から最も記事数の多い英語への翻訳を探すことで、文章を英単語列に変換し、英単語列同士で比較することで実現している。これは、対象とする言語を増やしていっても翻訳にかかる時間が増えないというメリットがある。また、英語を途中経由してもさほど単語数は減らないという点で計算時間を考慮すると言語間で一対一対応で翻訳を探すよりは有利であると考えられる。

7.2　今後の課題

これから解決しなければならない課題として以下が挙げられる。

- 対象とする言語の拡大
  今回は日本語と韓国語しかできなかったが、実用的な言語横断コーパスの構築のためにはもっと幅広い言語について実験、検証してみる必要がある。

- 精度の向上
  今回の実験では、関連するかどうかわかるためにはある程度の長さの数値か英単語、個人名といった単語が入っていなければならなかった。また、それらが含まれていても必ずしも関連するとはいえないゴミも多く、この中から実際に関連するものを見つけるのは楽とはいえない。あまり言語情報を用いない、という本研究の目標の中で、精度を向上させるにはニュースのタイトルのみではなく、ニュース本文も用いたニュースの比較やほかの指標が必要となってくると考えられる。

参考文献

[1] N.Kushmerick, D.S. Weld, and R¿ B. Doorenbos. Wrapper induction for information extraction. In Proc. of the Int. Joint Conf. on Artificial Intelligence, 1997.

[2] C.Hsu and M.Dung. Generating finite-state transducers for semi-structured data extraction from the web. Informatin Systems, 23(8):521-538,1998.

[3] I. Muslea, S.minton, and C.Knoblock. A hierarchical approach to wrapper induction. In Proc. of the Third Int. Conf. on Autonomous Agents(Agents'99) , 1999.

[4] A. Sahuguet and F. Azavant. Building light-weight wrappers for legacy web data-sources using W4F. In The VLDB Journal, 1998.

[5] L.Liu, C.Pu , and W.Han. XWLAP: An XML-enabled wrapper construction system for web information sources. In IEEE Int. Conf. on Data Engineering,2000.

[6] R.Baumgartner, S. Flesca, and G. Gottlob. Declarative information extraction, Web crawling, and recursive wrapping with Lixto. In Proc, of Int. Conf. on Logic Programming and Nonmonotonic Reasoning,Vienna, Austria, 2001.

[7] C.Chang and S. Lui. Iepad: Information extraction based on pattern discovery. In Proc. of the Int. World Wide Web Conf, 2001.

[8] V.Crescenzi, G.Mecca, and P.Merialdo. Roadrunner: Towards automatic data extraction from large web sites. In Proc. of 27th Int. Conf. on Very Large Data Bases, 2001.

[9] Utku Irmak, Torsten Suel. Interactive Wrapper Generation with Minimal User Effort International World Wide Web Conference archive Proceedings of the 15th international conference on World Wide Web, 2006

[10] Webstemmer. http://www.unixuser.org/ euske/python/ webstemmer/index-j.html

[11] Integrating Cross-Lingually Relevant News Articles and Monolingual Web Documents in Bilingual Lexicon Acquisition, Takehito Utsuro, Kohei Hino, Mitsuhiro Kida, Seiichi Nakagawa, and

Satoshi Sato, Proceedings of the 20th International Conference on Computational Linguistics, pp. 1036-1042, August 2004.

[12] Navigating Multilingual News Collections Using Automatically Extracted Information R Steinberger, B Pouliquen, C Ignat. Information Technology Interfaces, 2005. 27th International Conference on Publication Date: June 20-23, 2005

[13] 複数のニュース源の差異を考慮したニュース分析の研究. 吉岡　真治. 第 13 回言語処理学会年次大会併設ワークショップ大規模 Web 研究基盤上での自然言語処理情報検索研究. 2007.

発表文献

1. 吉田 慎一郎，田浦 健次朗，近山 隆．言語横断的なニュース関連付けのためのラッパー作成．ソフトウェア科学会第 24 回大会特別セッション：情報爆発,，奈良，2007 年 9 月．

謝辞

本研究を進めるにあたり、近山隆教授、ならびに田浦健次朗准教授には大変お世話になりました。近山隆教授には、適切な助言、貴重なご指摘を何度もいただきました。また、田浦健次朗准教授には、普段から研究の方向性について貴重なご意見をいただき、ミーティングでは大切なポイントを教えていただき、困ったときには相談にもらい、研究の進め方について丁寧な指導をしていただきました。心より感謝申し上げます。

横山大作さんや、鴨志田良和さん、斉藤秀雄さんには普段からお世話になり、ミーティング等の議論の場において様々なアドバイスを頂き、多くの知識を享受して頂きました。

同期の斉藤大君、関谷岳史君は研究に行き詰ったとき、よく話相手になってくれました。

その他、研究室の皆様には、研究を進めるにあたり日頃から助けていただきました。

本当にお世話になりました。ありがとうございました。

平成 20 年　2 月　4 日