E-015

多国多言語ニュース記事の検索・比較システム

Retrieval System for Comparison of Multilateral Articles

斉藤　雄介†　山田　剛一†　絹川　博之†　中川　裕志‡
Yusuke Saito　Koichi Yamada　Hiroshi Kinukawa　Hiroshi Nakagawa

1. はじめに

現在、世界各国からインターネットを通じて膨大な量のニュースが毎日発信されている。各国間には国籍や文化の違いがあるため、同一の国際的なトピックスであっても、表現方法が異なったり、偏った報道がされたり、時には全く報じられないこともある。

本研究では、国・言語の異なるニュース記事に対して、言語横断的に検索をすることで、各国間との報道の違いを比較するためのシステムの開発を行う。本システムより、国内報道だけでは得られなかった情報を提示することで、各国間との感覚のズレや考え方の違いを身近に感じられる環境を提供する。

今回は、日本、中国、台湾（中華民国）、イギリス 、アメリカ の 5 つの国と地域のニュースサイトから得られる記事を対象とし、比較インタフェースの開発に取り組む。

2. ニュース記事の収集・検索

システムの流れを以下に示す（図 1）。

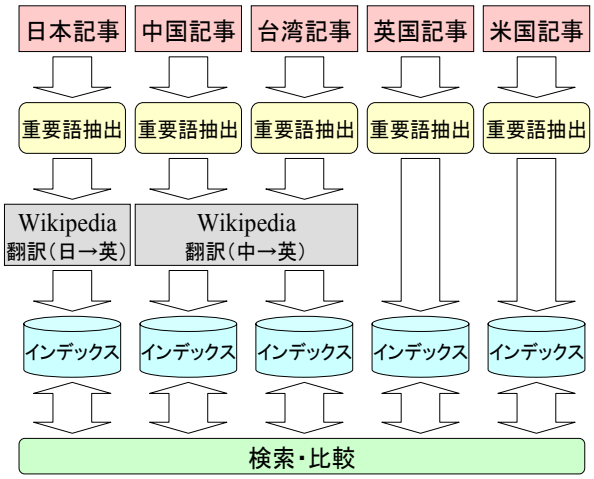

図 1 システムの流れ

2.1 ニュース記事の収集

今回対象としてる 5 つの国・地域の代表的なニュースサイトから記事を収集する。クローラとしては、Webstemmer[1]を用いる。Webstemmer は、ニュースサイトの URL を指定するだけで自動的に本文やタイトルを抽出するソフトウェアである。収集先のニュースサイトは、Wikipedia[2]などを参考に各国の代表的なニュースサイトを選択した。日本 5 サイト、中国 7 サイト、台湾 4 サイト、イギリス 4 サイト、アメリカ 6 サイトの 26 サイトから収集を行う（表 1）。それぞれの収集先は現地語で現地向け書かれたニュースサイトである。

† 東京電機大学大学院 工学研究科, Tokyo Denki University
‡ 東京大学 情報基盤センター, the University of Tokyo

表 1 収集先 URL（一部）

新聞社	URL	国
朝日新聞社	http://www.asahi.com	日
読売新聞社	http://www.yomiuri.co.jp	日
新華社	http://www.xinhuanet.com	中
中央社	http://www.cna.com.tw	台
タイムズ	http://www.timesonline.co.uk	英
ガーディアン	http://www.guardian.co.uk	英
ワシントン・ポスト	http://www.washingtonpost.com	米
ニューヨークタイムズ	http://www.nytimes.com	米

2.2 重要語抽出

重要語抽出ソフトウェアである TermExtract[3]を用いて、本文から重要語を抽出する。TermExtract は、形態素解析された日中英の各文章中から重要語を抽出するソフトウェアであり、重要語ともに算出したスコアを出力する。ここでのスコアは「単名詞（複合語を構成する最小単位の名詞）が他の単名詞と連結して複合語をなすことが多いほど、重要な概念を示す」という仮説に基づき計算される。本研究では、この重要語上位 100 語を使用する。

2.3 重要語翻訳

ニュース記事から抽出した重要語はすべて、世界で最も多くの国で使用されている言語である英語に翻訳し、インデクシングを行う。 これにより、言語横断的な検索・比較を可能にする。

本研究では、本文から抽出した重要語単位で翻訳を行う。翻訳は通常の辞書で行うと、新語・人名など、新たに増える単語に対しては翻訳をすることができないため、特にこれらの単語が多く出現するニュース記事には不向きである。そこで、Wikipedia を用いて辞書を自動作成し、それを用いた翻訳を行う。Wikipedia には言語間リンクとよばれる、他言語へのリンクがあり、それを利用することで目的の他言語に翻訳する（図 2）。

図 2 Wikipedia の言語間リンク

また、原文はユーザに表示するために Google AJAX

Language API[4]を用いて翻訳する。Google AJAX Language API は google のサービスの一つで Javascript で提供されている API である。

2.4 ニュース記事の検索

検索及びインデクシングには、全文検索エンジンである Apache Lucene[5] によって実装する。検索には英単語に翻訳した重要語を用いる。一日あたりの平均収集記事数を以下に示す（表 2）。

表 2　一日あたりの平均収集記事数

	記事数		サイト数
	国・地域	言語	
日本	2029	2029	5
中国	3171	5206	7
台湾	2035		4
英国	758	1352	4
米国	594		6

記事の検索方法には、クエリによる「フリーワード検索」とあらかじめ比較対象の記事を推薦する「多国ニュース記事クラスタ比較表示」の 2 通りのいずれかの方法により比較する記事を選ぶ。

2.4.1 フリーワード検索

「フリーワード検索」では、英語をクエリとして通常の Lucene による検索を行う。Lucene では TF-IDF 法を基としたベクトル空間モデルを用いて、文書の類似度を求めることができる。

本研究では、TermExtract で得られた重要語を利用するため、tf の代わりに TermExtract で得られるスコアを用いて、以下の式でスコアリングを行う。

$$score(q,d) = queryNorm(q) * \sum_{t \ in \ d} \{ weight(t,d) * idf(t)^2 * lengthNorm(t \ in \ d) \}$$

q：検索語　***t***：単語　***d***：文書

idf(t)
log { 全文書数／(t を含む文書数＋1) } ＋1
queryNorm(q)
score(q,d)の値が 0.0 から 1.0 までに収まるように正規化するための関数。
weigth(t, d)
d 中の t の重み。
lengthNorm(t in d)
t を含む d のトークン数の平方根の逆数が返される。

2.4.2 多国ニュース記事クラスタ比較表示

多国間で比較したい記事のトピックが決まっているときにはフリーワード検索が有効であるが、そうではなく今話題となっているトピックの中から比較すべきトピックを知りたい場合には、それを推薦することが有効である。

本システムでは国・地域ごとに記事集合をクラスタリングし、各クラスタに含まれる記事数やそれらの記事の発信元のサイト数などからクラスタをスコアリングする。これにより、国・地域ごとの主要なトピックを俯瞰することができる。

また、ある国で主要なトピックとなっているものが他国でどのような扱いをされているのかを可視化する手法として、次の方法を採っている。まず基準となる国(これを軸と呼ぶ)を選択し、その国におけるスコアの高いクラスタから順に、他国の記事のクラスタで類似度の高いものを探す。軸となる国でスコアの高いトピックが他国でどのようなスコアとなっているのかを図 3 のように可視化する。これにより、多国間における注目度の違いを明らかにすることができる。

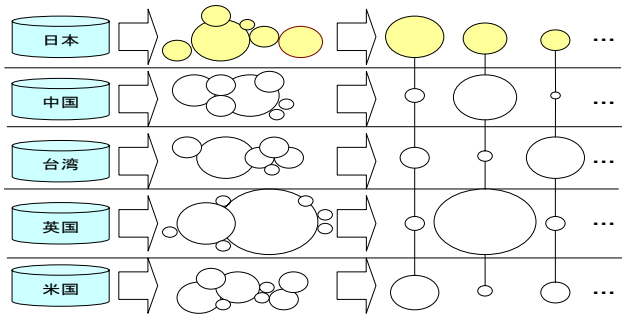

図 3　上位 3 クラスタの場合（軸＝日本）

3. ユーザインタフェース

画面イメージを図 4 に示す。多国ニュース記事クラスタ比較表示では、軸とした国・地域内でクラスタスコアが上位 3 位のクラスタとその代表記事を表示し、さらにそのクラスタと最も関連のある各国の記事クラスタのスコアをクラスタアイコンの大きさで図示する。

図 4　画面イメージ図

4 おわりに

本研究では、多国間のニュース記事の検索および比較方法について提案した。今後、ここで述べた方法を実装し、実験・評価する。また、研究対象とする国・地域の拡大、翻訳性能の改善等についても検討していく。

謝辞

本研究で使用した、Webstemmer、Apache Lucene、TermExtract、Wikipedia、Google AJAX Language API を開発された方々に感謝いたします。

参考文献

[1] Webstemmer, http://www.unixuser.org/~euske/python/webstemmer/index-j.html
[2] Wikipedia, http://ja.wikipedia.org/wiki/
[3] TermExtract, http://gensen.dl.itc.u-tokyo.ac.jp/
[4] Google AJAX Language API, http://code.google.com/intl/ja/apis/ajaxlanguage/
[5] Apache Lucene, http://lucene.apache.org/