E-018

世界の多言語ニュースの収集と分類

Collecting and Indexing Global Multilingual News Articles

佐々木　慎† 山田　剛一† 絹川　博之† 中川　裕志‡
Makoto Sasaki Koichi Yamada Hiroshi Kinukawa Hiroshi Nakagawa

１．はじめに

現在、世界のさまざまな国からニュースが発信され、膨大な量のニュースが世界中の人々によって読まれている。そして、その中には言語は異なるが同様の事項を報じているものがあり、それらは同様の事項を報じているニュースであっても国によって報道のされ方に相違がある。報道をする人の国籍および文化によっても報道のされ方が異なる。たとえば、情報統制されている国では民主主義や人権保護といったニュースが公開されることはほとんどない。その他の国でも自国に都合の悪いニュースは報道しなかったり、自国の国益に関係の深いニュースは報道が激化したりする。これらの各国間の報道の相違を発見および抽出するのが本研究の目的である。

本研究では多くの国々から多言語記述のニュースを大規模に収集、分類、比較する。この多国籍かつ多言語のニュース発信源から収集したニュースを今後「多国多言語ニュース」と呼ぶ

ニュース収集では、過去の多言語ニュース検索[1][2][3]を上回る規模で、多国多言語ニュースの収集を行っている。分類は、ニュースのトピックごとに行い、それを時系列に並べて表示する。比較においては時系列にトピックごとに表示することによって行う。

このようにして、多言語ニュース検索のみならず、多言語間で、同様のトピックについて扱ったニュースを国籍・言語を問わず時系列に抽出・追跡・比較することが可能であることを示し、各国のニュースごとの報道のされ方の違いについて見ることが可能となる。本稿では収集と分類についての検討について報告する。

２．関連研究

既存の多国籍からの多言語記述のニュース記事検索・推薦・要約等に関する論文ではニュース収集が比較的小規模で実験・システム運用等が行われている。

NewsBlaster[1][2],NewsInEssence[3]においてニュース収集規模は NewsBlaster で 27 サイト、NewsInEssence においても 35 サイトと非常に小規模である。NewsBlaster は多国多言語のニュース記事を対象とした文書分類・要約・検索システムである。NewsInEssence は多国単言語ニュース記事・要約・検索システムである。多国単言語ニュースとは多国からのニュース源を利用しているが、言語は単一の言語であるものをいう。多国単言語ニュース間の記事の差異を見る研究としては吉岡[4]の研究がある。多国多言語間の Weblog を用いた研究として福原[5]らの研究がある。本研究ではこれらの関連研究における多国および多言語からのニュース収集、文書分類といった部分を同時に扱う。

既存の CLIR に関する研究では、対象の言語に直接辞書や機械翻訳で翻訳する方式と、直接対象の言語に翻訳せず、一度他の言語に翻訳した後に対象言語に翻訳するピボット方式が存在する。本研究では全ての語を英語に翻訳するため、ピボット方式が近いと言える。

3．収集・分類

ここでは、多国多言語ニュースの収集方法および、分類に適した形にコーパスを加工する方法について述べる。

3.1 多言語ニュース収集

Webstemmer[6]というニュース収集ソフトウェアを並列に動かし、多国多言語ニュースコーパスを作成している。収集する対象は取り扱っている言語数・国の数が多いのが理想的である。そのために世界中のニュースサイトのリンク集である、newspaperindex.com[7]および、Kidon Media-Link[8]を用い、世界 205 ヶ国から約 1500 サイト程度を収集対象としている。将来的には 15000 サイト程度に拡張予定である。

3.2 重要語抽出

収集した多国多言語ニュースから TermExtract[9]を用いて重要語を抽出する。これは、同一トピックを抽出するために必要なキーワードを抽出するために用いている。TermExtract の重要語抽出方式には「他と連接して複合語をなすような単語こそ、文中の核となる概念を表している」という考えに基づき、単語の連結回数を元に重要度を算出している。ためその言語によらず重要度を算出することができる。

3.3 翻訳

収集した多国多言語ニュースの重要語を和英・中英・欧英といった単語辞書を用いて語単位に英語に翻訳する。ニュース本文を翻訳せずに重要語のみを翻訳するのは言語によっては翻訳システムが存在しない、もしくは翻訳システムが成熟しきっておらず翻訳精度が悪いという問題によるものである。また、英語単語に翻訳する理由は英語が事実上の世界標準語であり、英語の辞書群が世界で最も充実しているという理由による。

新語・人名等については、単語辞書での対応が不可能であるので、Wikipedia[10]のデータベースを用いて翻訳する。Wikipedia は随時更新が行われており新語および人名といったものに随時対応が可能である。こうしてできた重要語を索引付けし検索可能とする。

† 東京電機大学大学院
‡ 東京大学

3.4 索引付け

英語単語に翻訳した重要語を用いて索引付けをすることにより、類似した多国多言語ニュースを、多言語で検索することができる。その結果を、ニュース発信時刻を用いて時系列順に配列表示すると、過去および未来への多国多言語ニュース追跡が可能となり、時系列的な流れを把握することができる。ここでの索引付けおよび検索には全文検索エンジン Apache Lucene[11]を用いている。

3.5 スコアリング法

本研究の収集および分類方式では、TermExtract によって重要語を抽出する際に同時に重要度が得られるため、検索の際のスコアリングに従来の tf-idf 法の tf の代わりに重要度を利用することができる。よって、以下のような方式でスコアリングを行っている。

$$score(q,d) = Norm(q) \cdot \sum_{t\ in\ q} weight(t,d) \cdot idf(t)$$

$weight(t,d)$: 記事d中の単語tの重要語重み

$$idf(t) = 1 + \log \frac{\text{記事総数}}{\text{重要語が出現した記事}+1}$$

$weight(t,d)$: 記事d中の単語tの重要語重み

$Norm(q)$: $score(q,d)$が最大1.0になるよう正規化するための値

q : 検索語

ここまでに述べた工程により、多国多言語ニュースを検索および分類に適した形として持つことが可能となる。その工程を図１に示す。

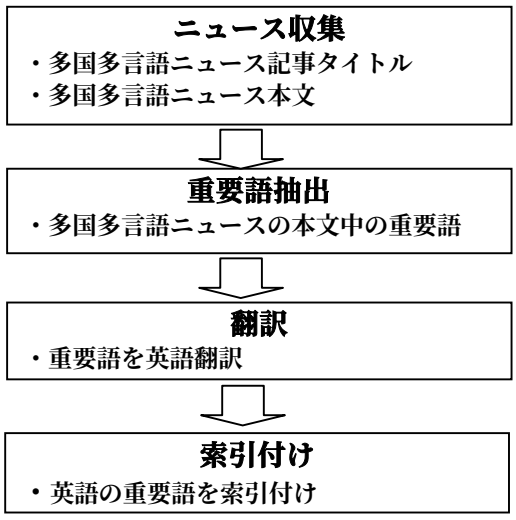

図１　多国多言語ニュース収集と加工の流れ

4．考察

本方式によって多国多言語ニュースの収集と分類を行った際の問題点について考察する

4.1　ニュース収集に関して

多国多言語ニュースを収集するにあたり、収集速度が問題となる。現在、1500 サイトの巡回に 8 時間程度の時間を要している。サイト数が増加した場合に実用に耐えなくなる恐れがある。また、Webstemmer により、ニュースサイトのリニューアル等でサイト構造が変化した場合収集が不可能になる恐れがある。これらの問題に関しては、マシンの並列化や収集ソフトウェアの改善で対処する予定である。

4.2　重要語抽出に関して

重要語抽出のプロセスにおいて、日本語や中国語のような膠着語の場合には形態素解析を行い、形態素ごとに分割した後に重要語抽出をする必要がある。英語のようにあらかじめ分割されている言語においては連接を見る方式は有利であるが、膠着語の場合には形態素解析システムが入手可能である保証がなく不利である。

4.3　翻訳に関して

語単位で翻訳することにより、現在の翻訳システムの持つ精度の悪さは回避されている。しかし、本システムでは同義語の存在や、辞書にも Wikipedia にも存在しない単語が考慮されていない。

5．おわりに

本研究では、世界の多国多言語ニュースの収集と分類手法について提案した。本手法によって、多国籍のニュースサイトから多言語ニュースを収集し、言語に関係なく検索・分類することが可能となる。

参考文献

[1] Kathleen McKeown, Regina Barzilay, John Chen, David Elson, David Evans, Judith Klavans, Ani Nenkova, Barry Schiffman, and Sergey Sigelman.Columbia's newsblaster: New features and future directions. In Proceedings of NAACL-HLT'03, 2003.

[2] David Kirk Evans, Judith L. Klavans,Kathleen R. McKeown.Columbia Newsblaster: Multilingual News Summarization on the Web.HLT-NAACL 2004.

[3] Dragomir Radev, Jahna Otterbacher,Adam Winkel, and Sasha Blair-Goldensohn.NewsInEssence: Summarizing ONLINE NEWSTOPICS COMMUNICATIONS OF THE ACM October 2005/Vol. 48, No. 10.

[4] 吉岡真治. 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. 第13回年次大会発表論文集 pp.32-35. 言語処理学会,2007.

[5] 福原知宏，宇津呂武仁，中川裕志. 複数言語間の語彙出現傾向比較による言語横断型ウェブログ関心解析システムの開発. 第13回年次大会発表論文集 pp.40-43. 言語処理学会,2007.

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

[7] newspaperindex.com http://www.newspaperindex.com/

[8] Kidon Media-Link http://www.kidon.com/media-link/index.php

[9] TermExtract http://gensen.dl.itc.u-tokyo.ac.jp/

[10] Wikipedia http://ja.wikipedia.org/

[11] Apache Lucene http://lucene.apache.org/java/docs/