日中英ニュース記事比較のための収集と検索

斉藤 雄介† 山田 剛一† 絹川 博之† 中川 裕志‡
東京電機大学大学院 工学研究科† 東京大学 情報基盤センター‡

1. はじめに

現在、ウェブ上では、世界各国から膨大な量のニュースが発信され、世界中の人々に読まれている。そして、その中には言語は異なるが同様の内容を報じているニュースがある。さらに、同じ事件や事象を指す内容でも、筆者の国籍や立場によって視点や表現方法がさまざまで、自国にとって都合の悪い事柄については省かれたり、偏った報道がされたりする。

これらの差異を解析し比較を行うことで、国際的な感覚の違いを身近に感じられる環境をつくることが本研究の目的である。

今回は、日中英ニュース記事の比較のための収集方法と、多義語の曖昧性を考慮した検索方法について報告する。

2. 関連研究

多言語間のニュース検索に関する研究はいくつか行われている。日中間のニュース検索方法として、漢字を直接利用した方法が Fredric[1]によって提案されており、英語を経由した検索よりも良い結果が得られている。

また、単一言語の多国のニュース記事の違いを比較する研究として、吉岡[2]の研究がある。ここでは、入力された検索語との共起語の違いや時系列による分析が行われている。

3. ニュース記事の収集・検索

システムの概要を以下に示す（図1）。

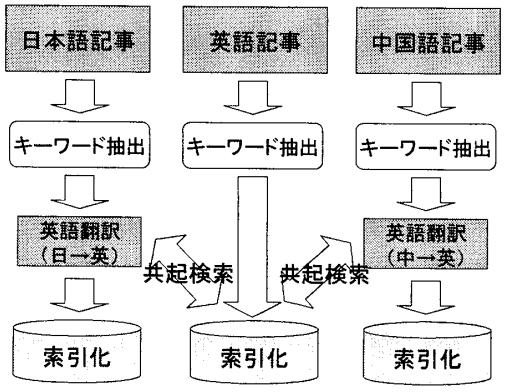

図1 システムの概要

3.1 ニュース収集

Webstemmer[3]を用いて、ニュース記事の収集を行った。Webstemmer とは、ニュースサイトから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。

また、収集先 URL は、世界中のニュースサイトを集めたリンク集である、NewspaperIndex.com[4]及び、Kidon Media-Link[5]を用いて取得している。

3.2 キーワード抽出

TermExtract[6]を用いて、本文からキーワードを抽出する。TermExtract とは、文章中から重要語を抽出するソフトウェアであり、重要語とともに算出したスコアを出力する。

本研究では、この重要語をキーワードとし、上位 100 語を使用した。

3.3 キーワード翻訳

収集したニュースは、全て世界標準語である英語に翻訳し、索引化を行う。 これにより、言語横断的に検索することができるようになる。

本研究では、本文から抽出したキーワード単位で翻訳を行う。辞書ベースには、EDICT[7]および CEDICT[8]を使用する。

新語・人名については、上記の辞書では翻訳をすることができない。そこで、Wikipedia[9]を用いて辞書を作成し、それを用いて翻訳を行った。Wikipedia には、他言語へのリンクがあり、それを利用ことで該当する言語に翻訳することができる（図2）。

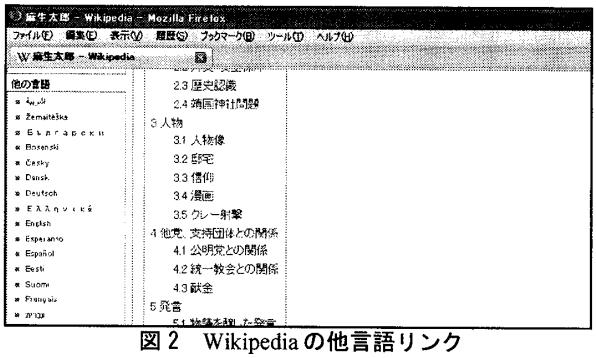

図2 Wikipedia の他言語リンク

また、日本語、及び中国語から英語への訳語は必ずしも 1 つではない。そこで英語記事を先に索引化し、これをコーパスとして利用することで、曖昧性を解消する。ここでは訳語候補のうち、共起する記事が一番多い語を訳語として採用した。

3.4 ニュース検索

検索・索引化には、全文検索エンジンである Apache Lucene[10] を用いて実装した。

英単語に翻訳したキーワードを Lucene によって索引化を行い、英単語をクエリとして用いることで、異なる言語間での類似記事の検索を試みる。Lucene では、TF-IDF 法、ベクトル空間モデルを用いた計算式を使っている。

Collection and Retrieval for Comparison of Japanese, Chinese and English News Articles
† Yusuke Saito, Koichi Yamada, Hiroshi Kinukawa, Tokyo Denki University
‡ Hiroshi Nakagawa, the University of Tokyo

本研究では、TermExtract で得られたキーワードを利用するため、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った。

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

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。

4. 本システムを用いた記事の分析

本システムを用いて、2008 年 5 月から 6 月の 2 ヶ月間に収集したニュース記事の分析を行った。

収集した各言語の記事数、サイト数は以下のとおりである。

- 日本語記事　16,230 件　5 サイト
- 中国語記事　6,460 件　3 サイト
- 英語記事　16,840 件　14 サイト

収集した各言語の記事の中からランダムに記事を抽出し、3 言語全てのサイトで話題となりそうなキーワードを抽出した。そのキーワードをクエリとして本システムを用いて検索を行い、3 言語による結果を比較した。

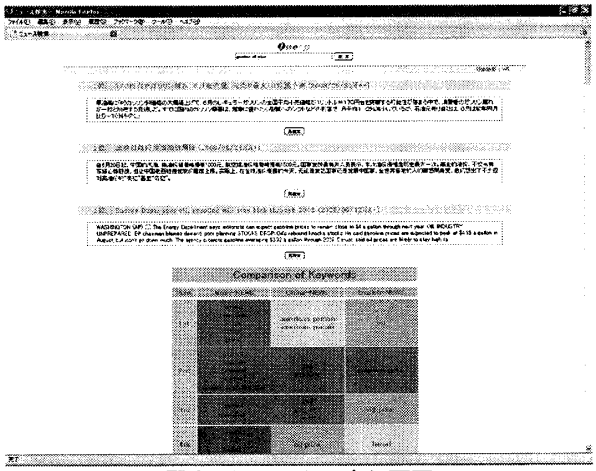

図 3　システムの実行画面

クエリと一番類似度が高い記事を並べ、各クエリと同じ単語を含む項目は同じ色で強調して表示した。各言語、上位に来るクエリや共起する語が違うことが一目わかるようになっている。

例として、「gasoline」「price」「oil」の 3 つをクエリとして検索した（図 3）。中国語の記事では、「ameri can people」や「vietnam」や「motorcycle」などが共起したことから、海外での原油の価格高騰の問題についても言及されていることや、英語記事では「future oil」や「next year」などが共起したことから、今後の原油価格について述べられていることがわかる。

5. 考察

本システムによって多言語ニュース記事の検索を行って生じた問題点について考察する。

5.1 翻訳について

「サイクロン」は、英語では「cyclone」、中国語では「气旋」と表現される。しかし、中国語では、ミャンマーで起きたサイクロンは「强热带风暴」、和訳すると強い熱帯暴風雨（中国での台風の表現）と表記される。つまり、サイクロンや台風に相当する訳語は存在するが、中国では表現方法が異なるため、同じ英単語ではうまく検索にヒットしないという問題が生じた。

5.2 検索結果について

今回の実験では、3 言語のサイト間で共通の話題を少数しか見つけることができなかった。世界共通の問題と思われる記事であっても、視点や切り口が異なるため、単純にクエリを与えて検索するだけでは、抽出することは難しい。

また、検索性能においても、辞書に存在しない固有名詞がある場合やクロール先のサイト数が少ないことによって、性能が低下したことが考えられる。

6. おわりに

本研究では、多言語間のニュース記事の収集・検索方法について提案した。今回の実験で、翻訳性能が予想以上に検索結果に影響することがわかり、この問題について再検討していきたい。

謝辞

本研究において、ニュース記事収集として Webstemmer、検索エンジンとして Apache Lucene、キーワード抽出として TermExtract、辞書として EDICT、CEDICT、Wikipedia を利用させていただきました。これらのソフトウェアを開発された方々には深く感謝いたします。

参考文献

[1] Fredric C. Gey : How Similar are Chinese and Japanese for Cross-Language InformationRetrieval?, NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan
[2] 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
[3] Webstemmer, http://www.unixuser.org/~euske/python/webstemmer/index-j.html
[4] NewspaperIndex.com, http://www.newspaperindex.com/
[5] KidonMedia-Link, http://www.kidon.com/media-link/index.php
[6] TermExtract, http://gensen.dl.itc.u-tokyo.ac.jp/
[7] EDICT, http://www.csse.monash.edu.au/~jwb/j_edict.html
[8] CEDICT, http://www.mandarintools.com/cedict.html
[9] Wikipedia, http://ja.wikipedia.org/wiki/
[10] Apache Lucene, http://lucene.apache.org/