E-046

ニュース記事の国別クラスタの作成と多国間対応

Clustering of News Articles in Each Country and Mapping Clusters between Different Countries

吉野 太郎† 斉藤 雄介† 山田 剛一† 絹川 博之†
Taro Yoshino Yusuke Saito Koichi Yamada Hiroshi Kinukawa

１．はじめに

現在，世界各国のニュースサイトから膨大な量のニュース記事が日々発信されている．それらの中には，複数の国が同一の国際的な話題について報じている記事があるが，各国の文化や思想などの違いによって，報道内容には違いが現れる．

そこで，多国間でニュース記事を比較することで，ユーザが多国間の感覚のズレや考え方の違いを比較できるシステムを開発する．

今回は，ある話題に対する各国の注目度の差を比較するために，国別の記事クラスタの作成と，それらの多国間の対応付けを行う．

２．クラスタの作成と多国間対応の目的

現在のシステムでは，記事の比較はフリーワード検索で行えるようになっている．この方法は，目的の話題が決まっていて，特に記事の内容を比較したい場合には有効である．しかし，各国が注目している話題は何か，またどの程度注目しているのかを知りたい場合には不向きである．

そこで，上記のような比較を可能にするために，国別の記事クラスタを作成し，それらの多国間での対応付けを行う．ある話題に対する注目度の高さは，その話題について報じている記事が属するクラスタの大きさで表されると考えられる．よって，この大きさを利用することで，ある話題に対する各国の注目度の差を比較することができる．

３．ニュース記事の収集・比較システム

3.1 ニュース記事の収集

ニュース記事の収集には Webstemmer を用いる．収集の対象とする国は日本・中国・台湾・イギリス・アメリカの5カ国で，ニュースサイトは記事内容の偏りを避けるために国内全域向けのものを選択する．

3.2 重要語の抽出

重要語の抽出には TermExtract[1]を用いる．TermExtract を用いることにより，複合語を重要語として利用することができる．また，重要語の重みには tf-idf値を用いることが一般的だが，本システムの検索用インデックスでは TermExtract が抽出した重要語を用いるため，tf-idf値の算出には tf値の代わりに TermExtract の算出したスコアを使用する．

†東京電機大学大学院　未来科学研究科
Graduate School of Science and Technology for Future Life, Tokyo Denki University

3.3 重要語の翻訳

言語横断的な検索を行うために日本語，中国語の重要語は英語に翻訳する．まず Wikipedia の言語間リンクを利用して作成した辞書で翻訳し，次に翻訳できなかった語を Google AJAX Language API で翻訳する．単一の辞書を用いるのではなくこのような手順で翻訳するのは，ニュース記事には新語や人名が多く含まれるためである．

3.4 索引化

索引化には Apache Lucene を用いる．索引化により，高速かつ効率的なフリーワード検索が可能になる．

3.5クラスタの作成と多国間対応

ニュース記事の国別クラスタを作成し，それらを多国間で対応付ける．クラスタの作成は，分割型階層的手法である Repeated Bisection法で行う．その際クラスタリングツールとして bayon[2]を用いる．bayon は分割ポイントの閾値を指定するだけで高速なクラスタ作成が可能なツールである．対応付けでは各国の記事を比較し，最も近い内容の記事を含むクラスタ同士を対応クラスタとする．

3.6 多国間のニュース記事の比較

ニュース記事の比較は以下の 2通りの方法で行う．

(1) フリーワード検索による比較表示

Apache Lucene による通常の検索を行い，国ごとに記事を表示して比較を可能にする．クエリには英語を使用する．

(2) クラスタを利用した比較表示

基準にする国を選択し，その国のクラスタを記事数順に並べ替え，上位のクラスタの記事数と代表記事を表示する，また，それらのクラスタに対応する他国のクラスタの記事数と代表記事も表示する．

４．実験と評価

4.1 実験対象

3.1 で述べた 5カ国から収集したニュース記事のうち，2010年4月10日から2010年4月11日の記事を用いて実験する．各国の記事数を次の表 1 に示す．

今回は英訳した記事タイトル中の語から，ストップワードを除外したものを用いて実験する．なお，ストップワードリストには，MySQL のデフォルトのリスト[2]を用いた．

表１． 実験に使用した各国の記事数

	日本	中国	台湾	アメリカ	イギリス
記事数	842	523	1375	613	362

4.2 国別クラスタの作成と評価

bayon を用いて各国の記事を国別にクラスタを作成し，F値により評価した. F値は情報検索の分野で伝統的に使用されている評価手法のひとつで，精度と再現率の調和平均から求めることができる．F値は以下の式で求められる．

$$P_{hk}=\frac{|A_h\cap C_k|}{C_k},\ R_{hk}=\frac{|A_h\cap C_k|}{A_h},\ F_{hk}=\frac{2P_{hk}R_{hk}}{P_{hk}+R_{hk}}$$

$$F\text{値}=\sum_{h=1}^{K}\frac{|A_h|}{N}\max_h F_{hk}$$

A_h : h番目の正解クラスタ

C_k : k番目のクラスタリング結果

N : 記事数，　P_{hk} : 精度，　R_{hk} : 再現率

なお，正解クラスタは 1人で作成したものを用いた．

F値により評価した結果を表２示す.

すべての国で**閾値が** 0.3～0.4 **のとき** F値が最大となっている.

4.3 多国間対応と評価

2カ国の全組み合わせで，全記事タイトル同士について以下の式でスコアを計算し，スコアが最も高い組み合わせを対応記事とする.

$$\text{スコア}S=\frac{\text{共通語数}\times(\text{記事Aの語数}+\text{記事Bの語数})}{2\times\text{記事Aの語数}\times\text{記事Bの語数}}\quad\text{(式 1)}$$

また，対応記事を含むクラスタ同士を対応クラスタとし，多国間で同じ話題のクラスタ同士が対応付けられているかを人手で確認した．なお，確認は1人で行った.

正解数を評価数で割ったものを正解率とし，まとめたものを表３に示す.

スコアが 0.6 以上では 92% 以上の正解率となっているが，0.4 未満では 10% と低くなっている.

4.4 考察

国別クラスタの作成では，日本・中国・台湾の記事は翻訳していることがネックになるかと思われたが，もともと英語であるアメリカ・イギリスと比較しても特に F値は低いわけではない．これは，記事タイトルは簡潔で単純な文章であることが多いため，機械翻訳でも正しく翻訳しやすかったためと考えられる.

多国間対応でスコアの低い組み合わせが存在するのは，今回の実験では，本来は他国に対応記事が存在しない記事についても，必ずいずれかの記事と対応付けたためである．よって，スコアが 0.4 を下回っている組み合わせは，対応記事が存在しないために起こったものである.

同じ話題を報じた記事でも，報道内容が違えば当然ながら使用される語も変わる．しかし，今回は記事タイトル中の語，つまり話題をあらわす最低限の語のみを使用したため，翻訳の影響も小さく，国別クラスタの作成と多国間対応の両方で，人手で作成した正解に近づけることができたと考えられる.

表２． F尺度による国別クラスタの評価結果

	F値				
閾値	日本	中国	台湾	アメリカ	イギリス
0.2	0.879	0.872	0.861	0.926	0.801
0.3	0.885	0.885	0.870	0.930	0.795
0.4	0.901	0.886	0.864	0.933	0.836
0.5	0.807	0.836	0.738	0.737	0.690
0.6	0.534	0.500	0.453	0.459	0.539
0.7	0.384	0.335	0.323	0.311	0.369
0.8	0.324	0.269	0.256	0.225	0.291

表３． 多国間対応の正解率

スコア S	全体数	評価数	正解数	正解率
0 ≦ S ＜0.1	1523			
0.1 ≦ S ＜0.2	4206			
0.2 ≦ S ＜0.3	1623			
0.3 ≦ S ＜0.4	546	30	3	0.100
0.4 ≦ S ＜0.5	213	42	19	0.452
0.5 ≦ S ＜0.6	69	69	45	0.652
0.6 ≦ S ＜0.7	41	41	38	0.927
0.7 ≦ S ＜0.8	17	17	17	1.000
0.8 ≦ S	30	30	30	1.000

５．おわりに

本論文では，日本・中国・台湾・アメリカ・イギリスの5カ国のニュース記事を，記事タイトル中の語のみを用いて国別クラスタの作成と多国間の対応付けを行い，それぞれの結果を評価した. 国別クラスタの作成では F値が 0.83 以上, 多国間対応ではスコアが 0.6 以上のときに正解率が 92% 以上との結果を得られた.

今後はストップワードリストの充実や，本文中の語を用いた場合の評価や，国別クラスタの作成を式１で実験するなどして， F値や正解率を向上させ，ニュース記事の収集・比較システムを完成させたい.

謝辞

本システムの中で使用させていただいた Webstemmer, TermExtract，Apache Lucene，bayonの開発者の方々に深く感謝いたします.

参考文献

[1] 前田朗 : 専門用語（キーワード）自動抽出用Perlモジュール "TermExtract"の解説, http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html,

[2] fujisawa : mixi Engineers’ Blog » 軽量データクラスタリングツールbayon，mixi Engineers’ Blog , http://alpha.mixi.co.jp/blog/?p=1049

[3] MySQL : MySQL :: MySQL 5.1 リファレンスマニュアル :: 11.7.3 全文ストップワード, http://dev.mysql.com/doc/refman/5.1/ja/fulltext-stopwords.html