E-058

ニュース記事の国別クラスタの作成と多国間対応の実験評価
Clustering of News Articles in Each Country, Mapping Clusters between Different Countries and Their Evaluation

吉野 太郎† 山田 剛一† 絹川 博之†
Taro Yoshino 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配信されている．それらの中には，複数の国が同じ話題について報じた記事もあるが，その報道姿勢や内容には，各国の思想や文化等による違いが表れる．

そこで，各国のニュース記事を比較することで，ユーザが各国の価値観の違いを比較できるシステムを開発する．

ある話題に対する各国の注目度の差は，各国の同じ話題を取り上げた記事数に表れると考えられる．本稿ではこの仮定に基づき，話題に対する注目度の比較のために，国別の記事クラスタの作成および記事クラスタの対応付けを行い，実験評価する．

2．ニュース記事の収集・比較システムの概要

2.1 ニュース記事の収集

ニュース記事の収集には Webstemmer[1] を用いる．収集対象の国は日本・中国・台湾・イギリス・アメリカの5カ国で，各国の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した．また，各ニュースサイトのカテゴリから，「健康」「住まい」といった比較に適さないカテゴリは除外した．

2.2 重要語の抽出

重要語の抽出には ChaSen と TermExtract[2]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．重要語の重みには tf-idf値を用いることが一般的であるが，本システムでは TermExtract が抽出した重要語を用いるため，重みの算出には tf値の代わりに TermExtract の算出したスコアを用いる．

2.3 重要語の翻訳

言語横断的な検索を可能にするために日本語，中国語の重要語を英語に翻訳する．ただし，ニュース記事は人名や新語を多く含むため，単一の辞書を用いて翻訳することは困難である．そこで，まずは Wikipedia の言語間リンクを利用して作成した辞書で翻訳を試み，翻訳できなかった語は Google AJAX Language API で翻訳する．それぞれ人名が充実している，新語への対応が早いという強みがある．

2.4 索引化

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

†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

2.5 クラスタの作成と多国間対応付け

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．クラスタリングは Repeated Bisection法で行う．その際クラスタリングツールとして bayon[3]を用いる．多国間の対応付けでは，まず各国の各クラスタの中心ベクトルを利用し，2国間で同じ話題のクラスタ同士を対応付ける．次に，2国の全組み合わせの対応付け結果から，同じクラスタを含む対応付け結果同士を結合し，多国間の対応とする．

2.6 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

英語を検索クエリとしてニュース記事検索を行い，目的とする話題について書かれた各国の記事内容を表示することで，ユーザ自身による比較を可能にする．

(2) クラスタを利用した比較

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，上位のクラスタに含まれる記事数の多少を図で表示し，併せて代表記事も表示する，また，それらのクラスタに対応する他国のクラスタについても記事数を図で出すことで，各国の話題に対する注目度を，直感的に理解しやすい絵として比較することを可能にする．表示イメージを図 1 に示す．

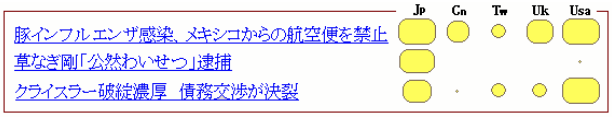

図 1． 話題注目度比較のイメージ

3．実験と評価

3.1 実験対象

2.1 で述べた 5カ国のニュースサイトから収集したニュース記事のうち，2011年6月8日から2011年6月11日の4日間の記事を用いて実験する．各国の記事数を表 1 に示す．

また，今回の実験に使用する重要語を抽出する範囲は，記事タイトルのみ・タイトルおよび本文の第 1段落まで・タイトルおよび本文の第 2段落までの 3通りとした．

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	1655	1403	1738	908	978

3.2 実験と評価

2.5 で述べた通り，まず各国のニュース記事を bayon を用いてクラスタリングする．次に，得られた各国の各クラスタを中心ベクトルを利用して 2国の全組み合わせについて対応付ける．クラスタの類似度 S は以下の式で計算する．

$$S = \frac{n(N_A + N_B)}{2N_A N_B}$$

N_A: クラスタAの語数
N_B: クラスタBの語数
n：A,Bに共通する語数

また，2国間での対応付けの結果を，類似度ごとに無作為に 50件ずつ（50件に満たない場合は全件）抽出し，精度を評価した．精度は以下の式で求めた．

$$\text{精度} = \frac{\text{同じ話題を正しく対応 付けた件数}}{\text{評価した対応付け件数}}$$

類似度の範囲ごとの対応付け件数を表２に，評価結果を表３に示す．

重要語の抽出範囲を記事タイトルのみに設定した場合は，どの 2国を組み合わせた場合でも，話題が同じであれば類似度が高く算出された．しかし，抽出範囲を本文の第1段落まであるいは第 2段落までに設定した場合は，アメリカとイギリス，中国と台湾の 2通りの組み合わせ，つまり元の言語が同じである組み合わせ以外は，同じ話題であっても類似度が低く算出された．

表２． 類似度の範囲ごとの対応付け件数

類似度S	タイトルのみ	第1段落まで	第2段落まで
0.9 ≦ S	17	1	7
0.8 ≦ S ＜0.9	2	8	3
0.7 ≦ S ＜0.8	8	3	1
0.6 ≦ S ＜0.7	31	2	3
0.5 ≦ S ＜0.6	115	7	5
0.4 ≦ S ＜0.5	510	7	8
0.3 ≦ S ＜0.4	2859	80	35
0.2 ≦ S ＜0.3	6850	583	216
0.1 ≦ S ＜0.2	2705	7654	4206
0 ≦ S ＜0.1	1116	4878	8776

表３． 対応付け結果の精度による評価

類似度S	タイトルのみ	第1段落まで	第2段落まで
0.9 ≦ S	1.00	1.00	1.00
0.8 ≦ S ＜0.9	1.00	1.00	1.00
0.7 ≦ S ＜0.8	1.00	1.00	1.00
0.6 ≦ S ＜0.7	1.00	1.00	1.00
0.5 ≦ S ＜0.6	0.92	0.86	0.80
0.4 ≦ S ＜0.5	0.66	0.43	1.00
0.3 ≦ S ＜0.4	0.12	0.54	0.63
0.2 ≦ S ＜0.3	0.04	0.38	0.64
0.1 ≦ S ＜0.2		0.08	0.24
0 ≦ S ＜0.1			

４．考察

重要語の抽出範囲を広げたときに，言語を超えて対応付けた結果の類似度が低く算出される理由は，ニュース記事の本文は複文を含む文章が多いために翻訳精度が落ち，重要語が一致しない場合があるためと考えられる．類似度の平均値が下がる理由は，翻訳精度の問題に加えて，1文が長くなることで迂遠な表現が増えたり，国によって記事の論旨展開の順序が異なったりするためである．また，語数が増えることで，同じ話題のクラスタ同士の重要語が一致した場合と，無関係なクラスタ同士のノイズとなる語が一致した場合で，類似度に明確な差が表れにくくなっている．

タイトルのみから重要語を抽出した場合は，少ない語数で簡潔にその記事の話題を表していることが多いため，上記のような問題は起きにくい．しかし，日本のニュース記事のタイトルでは，例えば「世界銀行」を「世銀」と表記するように，しばしば重要語が省略される場合がある．上記の例を Google 翻訳で翻訳すると，本来は「world bank」と翻訳されるべきところが「segin」と翻訳されていた．このような場合，正しく翻訳されていれば一致していたはずの語が一致しないことで，類似度が大きく下がってしまうことがある．

また，今回は定量的な評価を行えていないが，一定の類似度を閾値としたときの再現率については，記事タイトルのみから重要語を抽出した場合の方が優れていることを確認した．

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クラスタリングし，得られたクラスタを対応付け，精度で評価した．その結果，重要語をタイトルのみから抽出した場合に，類似度 0.4 以上で精度66%という結果を得られた．

今後の課題としては，各ニュースサイトのカテゴリ情報の利用や，略語のような表記ゆれに対応するために Wikipedia の転送設定を利用するなどして精度を向上させることと，再現率についての定量的な評価を行うことが挙げられる．また，重要語を本文からも抽出した場合に，国の組み合わせによって類似度の閾値を変えたときに，精度がどう変わるかを調査する必要があると考えられる．

謝辞

本システムの中で使用させていただいた Webstemmer, ChaSen, 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] 新山祐介：Webstemmer, http://www.unixuser.org/~euske/python/webstemmer/
[2] 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
[3] 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