E-045

同一事象に対する異新聞社記事間の類似点・相違点の検出

Detection of Similarities and Differences between Articles on the Same Matter by Different Newspaper Companies

三橋 靖大† 山田剛一† 絹川 博之†
Yasuhiro Mitsuhashi Koichi Yamada Hiroshi Kinukawa

１．はじめに

近年，コンピュータの普及により情報の受け渡しはアナログからデジタルに移行している．それに伴い新聞などの情報発信メディアも Web 上に進出してきている．これらの新聞では新聞社ごとの方針の違いなどによって，記事に取り上げられなかったり，記事に書かれている表現が異なったりする．

本研究では複数社の新聞記事間での比較を行い，新聞社ごとの違いを明らかにすることが目的である．本システムにより，単一メディアによる情報の偏りを防ぎ，柔軟な思考を助けることのできる環境を提供する．

今回は，同一事象に対する異新聞社記事間の類似点・相違点を検出するシステムの開発に取り組み，実験は朝日新聞社（以下朝日）と産経新聞社（以下産経）の２つの新聞社記事間を対象としている．

２．新聞社における記事の類似点・相違点

異新聞社記事間の類似点・相違点として以下の２つが挙げられる．

(a) ある事象に対する記事の有無
(b) 同一事象に対する記事の内容の違いや表現の違い

例えば，ある人が他の人に意見したとき，ある新聞社では「指摘した」と表現し，ある新聞社では「非難した」と表現するといった表現の違いや，記事の中である事象について触れているかいないかといった違いである．

３．異新聞社記事間類似点・相違点検出システム

類似点・相違点を検出するために，以下の (1)～(6) からなるシステムを提案する．（図１）

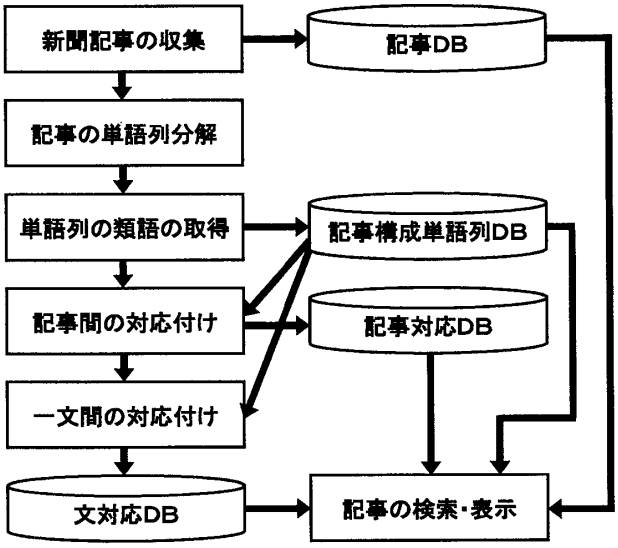

図１． 異新聞社記事間類似点・相違点検出システム

(1) 新聞記事の収集

比較対象となる新聞社のサイトから記事を収集し，タイトルと本文を抽出する．収集した記事は記事データベースに入れる．クローラには Webstemmer[1] を用いる．

(2) 記事の単語列分解

記事間の対応付けや文間の対応付けをする前段階として，記事を形態素解析によって短単位の単語列に分解する．形態素解析には MeCab[2] を用いる．

(3) 単語列の類語の取得

同一事象に対する記事中では，他社記事と同じ単語やその類語が使われることが多い．そのため，記事間の対応付けや文間の対応付けをするために (2) で生成した単語列の類語を取得する．取得した類語は記事構成単語列類語データベース（以下記事構成単語列データベース）に入れる．類語の取得には日本語 WordNet[3] を用いる．

類語を収集する形態素は「名詞」と「動詞」の２つに限定し，その種類も名詞は「一般」「固有名詞」「サ変接続」，動詞は「自立」に分類されたものに限定する．また，「運動する」のように「サ変接続の名詞」の直後に「サ変動詞」がある場合は名詞部分を動詞と見なし，関係のない記事が対応付けられることを防ぐために，日時などに使用される数字は除外する．

(4) 記事間の対応付け

新聞社間の違いを調べるために，異新聞社記事間で同一事象に対する記事を対応付ける．

(a) 同一事象に対する記事として対応付けられた記事群は，異新聞社記事間の類似点・相違点の検出に使用
(b) 対応のない記事は各社の取り扱う事象の違いの検出に使用

(3) で取得された類語群を使って２社の記事を１つずつ比較する．片方の記事の単語１つまたはその類語ともう片方の記事の同じ品詞の類語群とで一致しているものを検出する．この操作を「名詞」と「動詞」のすべてに対して行い，それぞれの一致度を得る．その後比較する側と比較される側を入れ替えて同じ操作を行う．この２つの結果を基に同一事象に対する記事であるかを判定し，その対応を記事対応データベースに入れる．

(5) 一文間の対応付け

同一事象に対する異新聞社記事間の類似点・相違点を探すために同一内容の文を対応付ける．

(a) 同一内容の文として対応付けられた文群は，類似点・相違点の検出に使用
(b) 対応のない文群は，各社の取り扱う事象の違いの検出に使用

記事対応データベースで同一事象に対する記事であると対応付けられている記事間で，一文ごとに対応付ける．比較単位を文単位にして (4) 記事間の対応付けと同様の方

†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

法で同一内容の文を対応付け，その対応を文対応データベースに入れる．

(6) 記事の検索と表示

検索方法として簡単なフリーワード検索を用意する．検索結果は記事タイトルの一覧で表示する．記事タイトルを選択することで同一事象の記事と共に表示する．片方の記事の文を選択すると，もう片方の記事の対応する文を強調表示する．これによって類似点・相違点を検出する．

4．評価実験

4.1 記事間の対応付け

(1) 実験対象

WebStemmer によって収集された 2009 年 10 月 1 日から 2009 年 10 月 5 日までの朝日 68 記事と産経 172 記事について記事同士を対応付ける．

(2) 評価方法

対象となる朝日の記事と産経の記事との一致度を求め，その値 P が 閾値 P_{th} 以上であるか，朝日から産経を見たときの一致度 M と産経から朝日を見たときの一致度 N が共に閾値 MN_{th} 以上であるとき同じ記事であると判断する．以下に値 P, M, N を求める式を示す．

$$M = \frac{AN + AV \times w}{NA + VA \times w}, N = \frac{SN + SV \times w}{NS + VS \times w}$$

$$P = M \times N$$

M：朝日の記事から産経の記事を見たときの一致率
N：産経の記事から朝日の記事を見たときの一致率
P：２つの記事の一致率，w:動詞の重み
AN：朝日の記事から産経の記事を見たときの名詞の一致数
AV：朝日の記事から産経の記事を見たときの動詞の一致数
NA：朝日の記事の名詞数，VA:朝日の記事の動詞数
SN：産経の記事から朝日の記事を見たときの名詞の一致数
SV：産経の記事から朝日の記事を見たときの動詞の一致数
NS：産経の記事の名詞数，VS:産経の記事の動詞数

また，事前に人手で朝日の記事から見て同一内容の産経の記事を調べ，その対応で正解集合を作成する．この正解集合を使い実験結果の評価と再現率を求める．以下に精度と再現率を求める式を示す．正解集合は 47 個である．

$$\text{精　度} = \frac{\text{実験結果の記事間の対応の正解数}}{\text{実験結果の記事間の対応の数}}$$

$$\text{再現率} = \frac{\text{実験結果の記事間の対応の正解数}}{\text{正解集合の記事間の対応の数}}$$

表１．　異新聞社記事間の対応付けの性能

P_{th}	MN_{th}	一致数	正解数	適合率	再現率	F 値
0.38	未使用	56	29	0.518	0.617	0.563
0.39	未使用	41	27	0.659	0.574	0.614
0.40	未使用	38	25	0.658	0.532	0.588
0.41	未使用	34	24	0.706	0.511	0.593
0.42	未使用	31	21	0.677	0.447	0.538
0.43	未使用	27	19	0.704	0.404	0.514
0.38	0.576	64	32	0.500	0.681	0.577
0.39	0.615	42	28	0.667	0.596	0.629
0.40	0.615	40	27	0.675	0.574	0.621
0.41	0.583	50	30	0.600	0.638	0.619
0.42	0.615	36	25	0.694	0.532	0.602
0.43	0.615	34	25	0.735	0.532	0.617

(3) 実験結果

動詞の重みを w＝5 としたときの精度・再現率およびその F 値を表１に示し，動詞の重み w について 1,3,5 の各々 F 値が最も良い結果を表２に示す．

表２．　動詞の重み w と対応付けの性能

w	P_{th}	MN_{th}	一致数	正解数	精度	再現率	F 値
1	0.43	0.600	37	24	0.649	0.511	0.571
3	0.43	0.600	39	27	0.692	0.575	0.628
5	0.39	0.615	42	28	0.667	0.596	0.629

5．評価・考察

5.1 対応付けに使用した類語の種類について

動詞の中には新聞記事に使われやすい単語がある．例えば「述べる」という動詞は，誰かの発言を載せた記事であれば，そのほとんどでその類語が使われている．そのため発言内容が違う場合でも P の値が上がってしまう．また，このような単語が多くある記事は，全く関係ない記事と対応付けられてしまうときがある．これを防ぐにはこのような現象を起こす動詞を調べて使用しないようにするか，重みを下げる必要がある．

5.2 対応付けに使用した閾値について

(1) 閾値 MN_{th} について

M,N の閾値 MN_{th} を用いることで，精度・再現率を向上することができた．極端に低い値を指定した場合精度が下がることがあるが，適切な値を使用すればこの方式は有効であるといえる．

(2) 最適な閾値について

動詞の重み w = 5, P_{th} = 0.39, MN_{th} = 0.615 のとき最も良い F 値を得ることができた．しかし，今回実験で使用した新聞記事数がまだ少ないため，これが適切な値であるとは断言できない．今後使用する記事を増やして実験的に最適値を求める必要がある．

6．おわりに

本研究では，同一事象に対する異新聞社記事間の類似点・相違点を検出する方法について提案した．記事間の対応付けについては，評価実験を実施して結果を得ることができた．

今後は，記事間の対応付けの精度・再現率の向上と，一文間の対応付けの評価実験の実施およびその精度・再現率の向上を目指す．また，記事の検索および表示するシステムを開発していく．

謝辞

本研究で使用した Webstemmer，MeCab，日本語 WodNet を開発された方々に深く感謝いたします．

参考文献

[1] Webstemmer
http://www.unixuser.org/~euske/python/webstemmer/index-j.html
[2] MeCab
http://mecab.sourceforge.net/
[3] 日本語WordNet
http://nlpwww.nict.go.jp/wn-ja/