E-060

同一事象に対する異新聞社記事間の相違点検出のための文間対応とその評価

Matching Sentences in each Article about the Same Matter between 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，コンピュータとネットワーク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．

本研究の最終目標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相違点検出のための文間対応システムの開発とその評価に取り組んだ．実験は朝日新聞社（以下朝日）と産経新聞社（以下産経）の２つの新聞社記事を対象としている．

２．新聞社における記事の相違点

異新聞社記事間の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．異新聞社記事相違点検出システム

相違点を検出するために，以下の (1)～(5) からなるシステムを提案する．（図１）

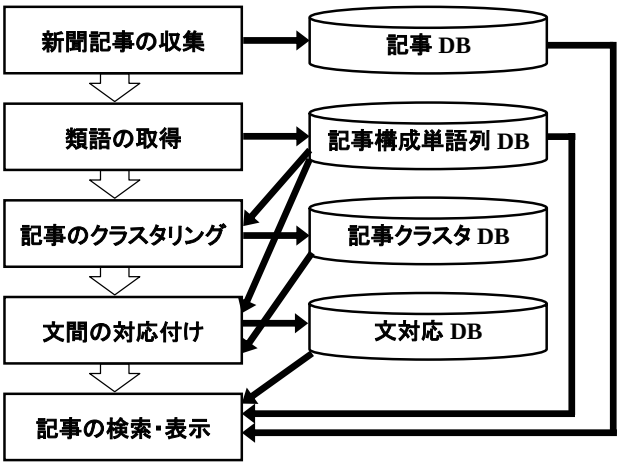

図１．異新聞社記事間相違点検出システム

(1) 新聞記事の収集

Webstemmer[1] を用いて比較対象となる新聞社のサイトから記事を収集し，記事データベースに入れる．

(2) 類語の取得

同一事象に対する記事中の同一内容の文の多くでは，他社記事と同じ単語やその類語が使われる．類語を用いて文間の対応をとるために記事中の単語の類語を取得する．

まず，記事を MeCab[2] を用いて記事を形態素解析する．類語を取得する形態素は「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．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．

(3) 記事のクラスタリング

異事象に対する記事中に存在する類似文間で対応が取られることを防ぐために，文間の対応付けの前に記事をクラスタリングする．(2)で取得した形態素のうち類語の取得に使った形態素と同じ「名詞」「動詞」とその TF*IDF の値を使って記事をクラスタリングする．クラスタリングツールには bayon[4] を用いる．

(4) 文間の対応付け

同一事象に対する異新聞社記事間の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，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

記事クラスタデータベースで同一クラスタに所属する記事間で，１～３文ごとに対応付ける．同一クラスタに所属する２社の記事間で(3)で取得された類語群を使い，まず１文ずつ比較する．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つの結果を元に同一内容文であるかを判定し，その対応を文対応データベースに入れる．同一内容文でないと判定された場合は片方を１文増やして同様の操作を行う．これを双方１～３文の間で繰り返し同一内容文を対応付ける．ただし，形態素数が一定数以下の文は対応を取らない．

(5) 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．これによって相違点を提示する．

†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

４．評価実験

4.1 記事のクラスタリング実験

(1) 実験対象

「政治」カテゴリに属する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記事についてクラスタリングする．IDF は2009年11月30日から2011年2月1日までの朝日82852記事と2009年12月3日から2011年1月16日までの産経76156記事から作成した．

(2) 評価方法

bayon を用いて記事をクラスタリングし，F値により評価した．F値は以下の式で求められる．閾値とは bayon でクラスタリングする際に使用する値である．

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

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

A_h : h番目の正解クラスタ

C_k : k番目の実験結果クラスタ

P : 適合率, R : 再現率, N : 記事数

(3) 実験結果

表１．記事クラスタリングの性能

閾値	F 値	適合率	再現率
0.80	0.527311	0.690345	0.598901
1.20	0.500512	0.580442	0.637363
1.60	0.469976	0.436822	0.71978
2.00	0.470203	0.42328	0.774725

4.2 文間の対応付け実験

(1) 実験対象

記事のクラスタリングと同一のものを使用する．

(2) 評価方法

対象となる朝日の記事文と産経の記事文との一致度を求め，その値 C が閾値 Cth 以上であるとき同一内容文であると判断する．以下に値 C を求める式を示す．

$$C = C_{hk} \times C_{kh}$$

$$C_{hk} = \frac{N_{hk}+V_{hk}\times w}{N_h+V_h\times w}, C_{kh} = \frac{N_{kh}+V_{kh}\times w}{N_k+V_k\times w}$$

C ： ２つの記事文の一致率

C_{hk}: 記事文hから記事文kを見たときの一致率

C_{kh}: 記事文kから記事文hを見たときの一致率

N_{hk}: 記事文hの名詞とその類語で記 事文kの名詞の類語と一致する単語数

V_{hk}: 記事文hの動詞とその類語で記 事文kの動詞の類語と一致する単語数

N_{kh}: 記事文kの名詞とその類語で記 事文hの名詞の類語と一致する単語数

V_{kh}: 記事文kの動詞とその類語で記 事文hの動詞の類語と一致する単語数

N_h: 記事文hの名詞の単語数, N_k: 記事文kの名詞の単語数

V_h: 記事文hの動詞の単語数, V_k: 記事文kの動詞の単語数

w ：動詞の重み

また，事前に人手で文間対応の正解対応を作成する．この正解対応を使い実験結果の適合率と再現率を求める．以下に適合率と再現率を求める式を示す．

$$\text{適合率} = \frac{\text{実験結果の文間対応の正解数}}{\text{実験結果の文間対応の数}}$$

$$\text{再現率} = \frac{\text{実験結果の文間対応の正解数}}{\text{正解対応の文間対応の数}}$$

(3) 実験結果

表２．異新聞社記事の文間対応付けの性能

Cth	対応数	正解数	適合率	再現率	F 値
0.30	22	16	0.727	0.340	0.464
0.35	18	15	0.833	0.319	0.462
0.40	13	11	0.846	0.234	0.367

５．考察

5.1 記事のクラスタリングについて

文間の正解対応はすべて同一クラスタ内に対応する文が含まれる記事が所属しているが，適合率・再現率はあまり良い結果は得られていない．実験対象記事数が少ないことが原因の可能性もある．また，インタビュー記事など複数の話題が書かれている記事，【橋本日記】【鳩山日誌】など個人の予定が書かれているだけの記事，【主張】などの筆者の意見が書かれている記事などクラスタに分けるのが難しい記事などがあるためだと考えられる．これらの記事をクラスタリングの前に除いておくことでクラスタリングの性能は上がると考えられる．

5.2 文間対応について

今回使用した閾値 Cth は以前同様の方式で実験した記事間対応の結果に基づいて決めたものであり，文間対応においてはこの値が適切ではない可能性がある．今後実験をしていくなかで適切な閾値を求める必要がある．

６．おわりに

本稿では，同一事象に対する異新聞社記事間の相違点検出のための文間対応の方式について，異事象に対する記事中の文間対応を防いだ方式を提案し，それを評価した．

今後は記事のクラスタリングと文間対応双方の適合率・再現率の向上を目指す．また，実験対象記事を増やした実験をしていく．

謝辞

本研究で使用した Webstemmer, MeCab, 日本語WodNet, bayon を開発された方々，ご協力いただいた計算言語学研究室の皆様に感謝いたします．

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/

[4] bayon
http://code.google.com/p/bayon/