新聞の社説を教師信号とする文章の右翼度・左翼度判定

木村弦　筑波大学
金丸敏幸　独立行政法人 情報通信研究機構
村田真樹　独立行政法人 情報通信研究機構
掛谷英紀　筑波大学

概要　これまで、自然言語処理技術により、記事のジャンル別分類を実現する研究は多く行われている。しかし、同じ政治面の記事でも、新聞によって内容が大きく異なることも多い。そういった記事に含まれる政治的イデオロギーを自然言語処理によって分別する試みはほとんど行われていない。その理由として、政治的イデオロギーを測る客観的指標が得にくいことがある。本論文では、その指標として新聞の社説に着目する。具体的には、毎日新聞と読売新聞の過去の社説を教師信号とし、最大エントロピー法により文章を毎日新聞寄りか読売新聞寄りか判定するプログラムを作成した。学習で得られたプログラムに、テストデータとして毎日と読売の社説を入力したところ、高い正解率の判定結果が得られた。

キーワード：　メディア、右翼、左翼、イデオロギー、最大エントロピー法

1. はじめに

これまで、自然言語処理技術により、記事のジャンル別分類を実現する研究は多く行われている。しかし、同じ政治面の記事でも、新聞によって内容が大きく異なることも多い。そういった文中に含まれる政治的イデオロギーによって記事を分類したいという希望も、社会問題に関心の高い知識人層には根強く存在する。しかし、実際にはそのような分類を試みる研究はほとんど行われていない。その理由として、政治的イデオロギーを測る客観的指標が得にくいことがある。本論文では、その指標として新聞の社説に着目する。

日本の新聞は、左翼系のものとして朝日新聞、毎日新聞、東京新聞（中日新聞）、保守・右翼系のものとして読売新聞、産経新聞がある[1]。もちろん、ひとことに右翼・左翼といっても、その主張の中身は多岐にわたり、単純に一軸で分類することは難しい[2,3]。よって、安易に右翼・左翼のラベルを貼ることは、単なるレッテリングになるおそれもある。しかし、左翼系の新聞の社説および保守・右翼系の社説を教師信号にし、それらへの類似性をもとに主張の左右を分類することは、多くの人にとってそれなりにイメージしやすいものであろう。

そこで、本論文では、2 節で実験に用いたシステムについて、3 節で実験結果とその考察、4 節で最大エントロピー法における α について記す。

2. 最大エントロピー法を利用した社説の判定システム

2.1 システムの概要

本論文では、右翼系の教師信号として読売新聞の社説を、左翼系の教師信号として毎日新聞を用いたシステムを提案し評価する。

毎日と読売の社説データから、単語、熟語、末尾表現の 3 つの素性を抽出し、それらから学習データ及びテストデータを作成し、最大エントロピー法に基づいたプログラムで、学習データから社説の特徴を学習し、テストデータで毎日か読売かを判定させる[4]。最大エントロピー法のプログラムとしてはmaxentを利用した[5]。単語は名詞と動詞に限った。その他の単語は思想に関係ないものが多いと考えられたので、素性からは外した。ここで言う熟語とは、名詞が 2 つ以上連なったもの及び形容詞が係る名詞である。また、末尾表現とは、句点「。」から逆に数えて文字数 1～10 個までの部分である。例えば、「・・という意見が得られた。」という末尾だったら、「う意見が得られた。」、「意見が得られた。」、「見が得られた。」、・・・、「た。」、「。」といったものである。

2.2 クロスバリデーション

素性データを 10 個のセットに分割し、9 個のセットを用いて学習を行い、残った 1 個のセットを用いてテストを行う．この操作をすべてのセットが 1 個ずつテストデータとして用いられるようにくり返す方法をク

ロスバリデーションといい、本論文で一部用いた。

3. 社説の判定実験

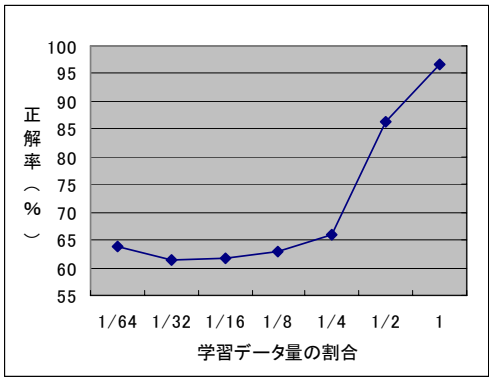
3.1 クロスバリデーションを用いた社説の判定実験

1991 年～2005 年までの社説データを使って、最大エントロピー法のクロスバリデーションで実験を行ったところ、正解率は 91.75%となった。

3.2 学習量を減らし社説の判定を行う実験

2004 年分の毎日と読売の社説データをテストデータ、2004 年分を除いた 1991～2005 年までの毎日と読売の社説データを学習データとし、学習データ量を減らしていく実験を行った。結果を表 1 に示す。学習データ量は 2004 年分を除いた 1991～2005 年の 15 年分を 1 とする。

表 1　学習量を減らす実験の結果



学習データ量の割合が 1/4 までは、正解率はほぼ横這いになっているが、学習データ量の割合が 1/2 から正解率が急に上がり、学習データ量が 1 になると正解率は 95%を超えている。判定を行う際に、学習データは 4、5 年分の社説では足りず、10 年分ほどは必要だとわかる。

3.3 被験者実験

1991 年の毎日と読売の記事から、各月の 5 日に載せられた社説を合計 42 件除き、読売か毎日かわからない状態で本文を読み、どちらか判定するというテストを、右翼、左翼に詳しい大学教員 1 名、大学院生の被験者 1 名に対し行った。

結果は大学教員については 42 問中 36 問正解で、正解率は 85.7%となり、大学院生については 42 問中 30 問正解で、71.5%の正解率となった。

次に、被験者実験に使った社説をテストデータとし、1991 年分を除いた 1992～2005 年分の毎日と読売の社説データを学習データとし、実験を行った。結果は正解率が 71.43%となった。正解率が前述のものと比べて低いが、これはおそらく、たまたま判定に困難なものが選ばれたためであろう。

4. 正規化したαのリスト

最大エントロピー法において、1991～2005 年までの毎日と読売の社説データを学習した時に、どの素性がテストデータを判定するのに重要になってくるかを示した変数αが算出される。αが高い素性を、αを正規化した上で、読売、毎日にわけて、それぞれ表 2、表 3 として一部示す。

表 2 の「か月」と表 3 の「カ月」や、表 2 の「反面」と表 3 の「半面」といった、新聞社の書き癖でαに差が出てしまうといった素性は、思想を反映しているとは言い難い。3.3 節の被験者となった大学教員に、どの素性が思想を反映しているか聞いたところ、表 2 の「わが国」、「国際社会」、「体制」、「平成」、「国益」、「昭和」、「阻止」、「市場経済化」、「着実」、「成長」は右翼の、表 3 の「市民」、「私たち」、「人々」、「キム」、「論理」、「庶民」、「金権」は左翼の思想を反映していると考えられるという意見が得られた。

5. おわりに

本論文では、新聞の社説を教師信号とする文章の判定システムを提案した。現段階では読売、毎日といった社説の判定では高い正解率が得られるが、右翼、左翼の判定をする段階には達していない。思想を反映しているとは言えない素性を排除し、また、新たな素性を加えるなどすれば、右翼、左翼の判定ができるようになると期待される。今後これらの点を改良したシステムを作成予定である。

参考文献

[1]Wikipedia 英語版
http://en.wikipedia.org/wiki/Main_Page
[2]掛谷英紀、日本の「リベラル」、新風舎、2002
[3]浅羽通明、左翼と右翼、幻冬舎、2006
[4]Eric Sven Ristad、Maximum Entropy Modeling for Natural Language 、 ACL/EACL Tutorial Program, Madrid、1997
[5]内山将夫氏、maxent
http://www2.nict.go.jp/x/x161/members/mutiyama/software.html

表2　正規化したαの読売の上位の素性69件

素性	読売のα	毎日のα
様々	0.846543	0.153457
か月	0.808666	0.191334
か国	0.803032	0.196968
だ。	0.791651	0.208349
十	0.787233	0.212767
し	0.730453	0.269547
読売新聞	0.729940	0.270060
巡る	0.723369	0.276631
こたえる	0.701828	0.298172
二	0.695382	0.304618
い。	0.689249	0.310751
わが国	0.675948	0.324052
反面	0.667481	0.332519
あり方	0.667372	0.332628
ものだ。	0.666433	0.333567
小泉首相	0.665713	0.334287
いる	0.659888	0.340112
国際社会	0.656330	0.343670
アメリカ	0.655642	0.344358
とされる。	0.649588	0.350412
言える	0.648066	0.351934
はたん	0.646242	0.353758
二十一世紀	0.643718	0.356282
こたえ	0.641847	0.358153
以前	0.641793	0.358207
する	0.640600	0.359400
橋本首相	0.635825	0.364175
昭和	0.634243	0.365757
めざす	0.632864	0.367136
狙い	0.632844	0.367156
破綻	0.631008	0.368992
体制	0.630516	0.369484
ロ	0.629400	0.370600
て欲しい。	0.629141	0.370859
図	0.628904	0.371096
欲しい。	0.628297	0.371703
のだろう。	0.624584	0.375416
阻止	0.624483	0.375517
市場経済化	0.621054	0.378946
中学	0.620011	0.379989
るものだ。	0.619727	0.380273
言え	0.618542	0.381458
かも知れない。	0.617844	0.382156
も知れない。	0.617785	0.382215
成長	0.616439	0.383561
明した。	0.616232	0.383768
か所	0.616230	0.383770
読売新聞社	0.613839	0.386161
た。	0.612553	0.387447
一因	0.612228	0.387772
巡っ	0.611777	0.388223
デフレ	0.610417	0.389583
国益	0.609268	0.390732
三	0.609141	0.390859
えるだろう。	0.604049	0.395951
言う	0.602797	0.397203
迅速	0.602756	0.397244
促し	0.601786	0.398214
」だ。	0.601551	0.398449
姿勢	0.601290	0.398710
平成	0.601023	0.398977
来月	0.599360	0.400640
自己	0.599092	0.400908
一定	0.598937	0.401063
着実	0.597578	0.402422
侵略	0.597434	0.402566
問だ。	0.596885	0.403115
え	0.596422	0.403578
見れ	0.595991	0.404009

表 3　正規化したαの毎日の上位の素性 69 件

素性	読売のα	毎日のα
純一郎	0.144214	0.855786
カ月	0.183680	0.816320
カ国	0.188303	0.811697
であろう。	0.195336	0.804664
であった。	0.240191	0.759809
半面	0.240938	0.759062
のである。	0.262559	0.737441
応え	0.268184	0.731816
富市	0.270238	0.729762
日	0.276289	0.723711
９	0.281804	0.718196
０	0.286728	0.713272
わけ	0.289693	0.710307
いわ	0.293004	0.706996
市民	0.293820	0.706180
私たち	0.295355	0.704645
１	0.295796	0.704204
とき	0.296828	0.703172
さまざま	0.297635	0.702365
小泉純一郎首相	0.302512	0.697488
たち	0.307753	0.692247
いえ	0.310420	0.689580
４月	0.316128	0.683872
ひとつ	0.319574	0.680426
破たん	0.323194	0.676806
３	0.325440	0.674560
村山富市首相	0.325599	0.674401
人々	0.326792	0.673208
応える	0.328715	0.671285
)。	0.328963	0.671037
あろう。	0.333379	0.666621
年度	0.334148	0.665852
２１世紀	0.338001	0.661999
なのである。	0.343768	0.656232
２	0.349052	0.650948
カ所	0.349354	0.650646
いま	0.350275	0.649725
のだ。	0.353197	0.646803
キム	0.353370	0.646630
分かる	0.353404	0.646596
ここ	0.356162	0.643838
３月	0.356627	0.643373
名	0.356946	0.643054
位置付け	0.358242	0.641758
いっ	0.363166	0.636834
毎日新聞	0.364094	0.635906
分から	0.364404	0.635596
護熙	0.364502	0.635498
るのだ。	0.365800	0.634200
１月	0.365935	0.634065
米国	0.366157	0.633843
６	0.366958	0.633042
９月	0.369869	0.630131
金権	0.370148	0.629852
やる	0.373586	0.626414
行っ	0.373812	0.626188
私	0.374099	0.625901
庶民	0.374432	0.625568
論理	0.374476	0.625524
アップ	0.374544	0.625456
と。	0.374633	0.625367
気	0.375584	0.624416
このこと	0.377657	0.622343
４	0.377796	0.622204
在り方	0.377971	0.622029
なのだ。	0.378353	0.621647
大蔵省	0.379715	0.620285
恵三	0.381849	0.618151
基づく	0.382624	0.617376