新聞の社説を教師信号とする文章の右翼度・左翼度判定 第２報

畑中允宏　筑波大学
金丸敏幸　情報通信研究機構
村田真樹　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、文章をジャンル別に分類する研究は多く行われている。しかし、同じジャンルの文章でも、その政治的・思想的スタンスは大きく異なることも多い。そのような文章中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。毎日新聞・読売新聞・日経新聞の社説を教師信号とし、最大エントロピー法により入力した文章がどの新聞の論調に最も近い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朝日新聞の社説を入力したところ、高い確率で毎日新聞寄りと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、文章をジャンル別に分類する研究は多く行われている。しかし、同じ政治というジャンルに属する文章でも、書き手の主義主張によって内容は大きく異なる場合がある。その主張に現れているイデオロギーに基づき文章を分類する試みはほとんど行われていない。その理由として、イデオロギーを測る客観的指標が得にくいことがある。

本論文ではこのイデオロギーの左右を峻別する指標として新聞の社説に着目する。一般的に日本の大手新聞社は、朝日新聞・毎日新聞がリベラル・左派、読売新聞・産経新聞が保守・右派と位置付けられている[1]。そこで、これらの新聞の社説を教師信号とし、それらへの類似性をもとに文章の右翼度・左翼度を判定することを試みる。2 章で実験に用いたシステムについて説明し、3 章で実験結果、4 章で考察を述べ、５章でまとめを行う。

2.　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の社説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[2]。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[3]。

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。熟語は名詞が 2 つ以上連なったもの及び形容詞が係る名詞のみを用いた。また、末尾表現は、句点「。」から逆に数えて文字数 3~7 個までの部分を採用した。例えば、「・・という意見もある。」という末尾ならば、「いう意見もある。」、「う意見もある。」、「意見もある。」、「見もある。」、「もある。」がそれに該当する。

3.　実験

3.1 判定結果の確信度分布

本論文の実験では、複数の社説からなるテストデータの社説の１つ１つについて、最大エントロピー法による機械学習で得られたプログラムにもとづき、各社の社説である確率を算出し、その確率(確信度)がより高い新聞社であると判定する。

1991~2005 年の 15 年分の毎日新聞と読売新聞の社説データを用いて、10 分割のクロスバリデーションにて実験を行った結果、正解率は 91.7%となった。このとき、それぞれ判定において確信度がどのような数値になっていたかを、頻度分布として表したのが図１である。確信度 50%以上が正解、50%未満が不正解を意味する。この図から、高い確信度で正解しているケー

スが多いことが分かる。

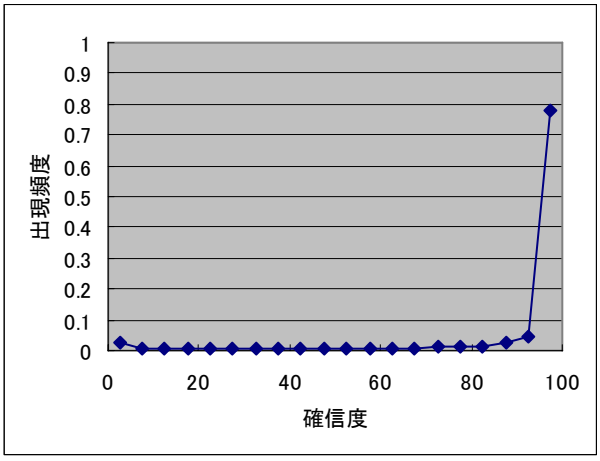

図１　社説の判定結果の分布

3.2 社説の判定実験 ― 朝日・産経・日経

1991～2005 年の 15 年分の毎日新聞と読売新聞の社説を学習データとした判定プログラムに、Web ページから収集した 2006 年と 2007 年の朝日新聞の社説を 1 年分ずつ、2007 年の産経新聞の社説を約 4 ヶ月分 200 件、そして記事データベースから収集した日本経済新聞の 1991～2005 年の社説 1 年分ずつをそれぞれテストデータとし、それらが毎日新聞に近いか読売新聞に近いかを判定させる実験を行った。また判定結果改善のため、以下の条件で素性データに変更を加え、これについても同様に実験を行い比較した。

- 条件１　末尾表現の排除

学習データ・テストデータから素性データの一つである末尾表現を排除し、単語と熟語のみとする。

- 条件２　数字を含む素性の排除

α値に毎日新聞と読売新聞で大きく差が出た素性のうち、単純な新聞社の表記方針を反映していると考えられるものを含む素性を、元の素性データから排除する。具体的には、数字全て（漢数字・アラビア数字と表記がことなるため）を排除する。（α値については 4 章参照）

それぞれの実験条件で判定を行った結果を表１、表２および図２に示す。

表１　朝日が「毎日」と判定された割合

	条件なし	条件1	条件2
朝日 2006	86.7%	86.9%	68.1%
朝日 2007	92.3%	90.9%	74.9%

表２　産経が「読売」と判定された割合

	条件なし	条件1	条件2
産経 2007	25.0%	36.5%	62.5%

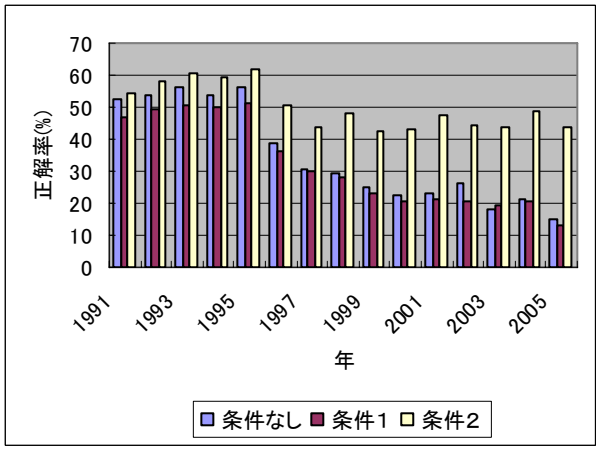

図２　日経が「毎日」と判定された割合

朝日新聞に関しては、どの条件についても同じ左翼系の新聞社である毎日新聞と判定されている。これは右翼度・左翼度判定システムとして望ましい出力である。

産経新聞については、75.0%の確率で毎日新聞と判定された。右翼系の産経新聞が毎日新聞と判定されるのは、右翼度・左翼度判定システムを作るという目的において好ましくない。条件１では改善はみられなかったが、条件２にて 62.5%で同じ右の読売新聞と判定され、結果に改善がみられたことになる。

日経新聞を毎日新聞・読売新聞に対して中道な位置付けであると考えるのなら、正解率が 50%付近をさまようことが理想である。実際 1995 年まではそのような結果だが、1996 年以降急激に読売新聞に近いと判定され始める。日経新聞をどう位置付けるのであれ、急激に変化するのは好ましくない。条件１では改善はみられなかったが、条件２では読売化はみられるものの 50%付近で留まっている。

3.3 学習データを 3 社にする実験

学習データに日経新聞も加えることで、読売新聞・毎日新聞の素性から中間的な思想を反映するものが薄れ、イデオロギーを反映する素性が濃縮されることを期待し、学習データを 3 社にして実験を行った。

まず、3 社について 1991~2005 年の 15 年分の社説を学習データとし、10 分割でのクロスバリデーションにて実験を行った。使用した素性については、3.2 節の条件 2 「数字を含む素性の排除」を適用している。結果、正解率は 83.3%となった。

また、この 3 社の判定プログラムに、2006 年と 2007 年の朝日新聞の社説を 1 年分ずつ、2007 年の産経新聞の社説約 4 ヶ月分 200 件をそれぞれテストデータとし、それらがどの新聞社にどれだけ近いかを判定させる実験を行った。結果を表 3 に示す。

表 3　朝日・産経の判定結果

	読売	毎日	日経
朝日 2006	26.1%	55.7%	18.1%
朝日 2007	22.0%	60.6%	17.4%
産経 2007	31.5%	26.0%	43.0%

朝日新聞でテストすると最も高く産経新聞でテストすると最も低い毎日新聞は、左翼の教師信号として理想的である。一方読売新聞は、日経新聞が中道というより右側に傾いていると判定されているため、日経新聞と互いに素性を食い合ってしまっているようにもみえる。

4. 考察

最大エントロピー法においては、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。3.3 節の 3 社で学習したときの新聞社ごとに α が高い素性上位 90 件を付録として示す。

α 値の高い素性を眺めると、思想を反映しているもの、新聞社の表記法の違いが影響しているものとが見受けられる。前者は「国際社会」「市場経済化」「庶民」「キム」等であり、後者は「こたえる」「応える」「小泉首相」「小泉純一郎首相」等である。またこの付録の α 表では数字を含む素性を排除しているが、そうしなければここに「3」「三」等の数字の表記法の違いが影響する素性が多く入ってくる。

だが同じ表記法の違いであっても、一つの社説の中で、漢字の使用の有無等の違いが影響する素性はそういくつも存在しないだろう。だが数字はほぼ必ず、しかも複数で存在する。他の表記法の違いが影響する素性とは判定への悪影響が比にならないと思われる。

5. まとめ

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。読売新聞・毎日新聞の社説の判定では高い正解率が得られ、他の新聞の判定については、3.2 節、3.3 節の実験より、右翼・左翼の判定ができる可能性を示唆する良好な結果が得られた。思想を反映しているとはいえない素性を排除し、右翼系・左翼系のその他の文書を更に学習データに加えれば、新聞社の方針を反映する素性は薄れ、思想を反映した素性のみが残り、より正確な右翼度・左翼度判定が可能になると考えられる。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]Eric Sven Ristad、Maximum Entropy Modeling for Natural Language 、 ACL/EACL Tutorial Program, Madrid、1997
[3]内山将夫氏、maxent
http://www2.nict.go.jp/x/x161/members/mutiyama/software.html

付録

読売のα上位の素性		毎日のα上位の素性		日経のα上位の素性	
読売新聞	国交	応え	された。	し	サダム
わが国	疑問だ。	毎日新聞	文部省	同国	産業界
反面	開発途上国	キム	行う	様々	ロ
小泉首相	さ	さまざま	誰	同社	今日
はたん	社会資本	純一郎	財政再建	する	農家
アメリカ	急ぐ必要がある。	応える	多発	われわれ	前提
昭和	かも知れない。	露	論理	議論	経済発展
破綻	問題だ。	富市	年度予算	点である。	るはずである。
来月	も知れない。	たのだ。	世界的	いる	くなっている。
て欲しい。	阻止	なのだ。	正常	ねらい	促す
先月	貿易一般協定	龍太郎	たち	とき	公算
るものだ。	とした。	やる	金権	なか	けである。
欲しい。	あたっ	であろう。	いえ	インフラ	やり方
社会党	諮問機関	護熙	名	半面	み
ためだ。	村山首相	るのだ。	庶民	つくる	ねらっ
読売新聞社	ぐ必要がある。	日	喜朗	事柄	メド
キロ	措置だ。	アップ	おか	役所	信頼感
ものだ。	乗せる	位置付け	行い	グローバル	分権
か年	メートル	小泉純一郎首相	年度	こたえ	氏
のことである。	生かす	そのこと	分かる。	かである。	強める
橋本首相	ゴミ	村山富市首相	同時多発テロ	事である。	温暖化ガス
言える	迅速	恵三	分かる	日本経済新聞社	事例
市場経済化	訳	であった。	破たん	である。	うである。
平成	きょう	在り方	部	このほど	主導
国際社会	作成	いのか。	分かり	巡る	踏み出す
こたえる	制度だ。	ひとつ	行わ	払しょく	違っ
てもいる。	のだろう。	橋本龍太郎首相	イラク戦争	かた	あげる
課題だ。	知れ	このこと	と思う。	おり	かねばならない。
狙い	ってはならない。	いわ	政府組織	信認	法である。
以前	至る	のである。	したのである。	不幸	照らし
するべきだ。	予定だ。	市民	きぐ	つくっ	正日
べきだ。	か年計画	賃金	サイド	対外	選択肢
年前	断念	なのである。	余	要である。	情報化
あと	あり方	分から	外	ことし	中東和平
三者	欠か	毎日新聞社	ここ	義務づけ	大事である。
取りまとめ	半年	アピール	たのである。	人員	規律
中学	効果的	おき	より	我々	張してきた。
論議	安保	いのだ。	樹立	需要	低下
見せ	言えるだろう。	モラル	はない。	ものである。	日本電信電話
重要だ。	手順	もつ	プラス	話し合い	中国
摩擦	曖昧	思う	多国	加える	肝心
妨げる	新党	行っ	るのか。	発展途上国	わけ
速やか	もなっている。	言い換えれ	ねばなるまい。	仕組み	成長率
としている。	市場開放	ことだ。	ダウン	たくさん	旨
図	抜本	その意味	ジョンイル	意志	本位