新聞社説・国会議事録に基づく言論のイデオロギー別分類

畑中允宏† 村田真樹‡ 掛谷英紀†
筑波大学システム情報工学研究科† 情報通信研究機構‡

概要 本研究では、自然言語処理技術によりイデオロギー別に文章を分類するシステムを構築する。これまで筆者らは、全国紙を発行する新聞社の社説を教師信号とする手法を提案してきたが、本研究ではイデオロギーの左右を峻別する指標として政党の左右に着目する。国会議事録内の発言を機械学習させ教師信号とし、それらへの類似性をもとに文章を分類することを試みる。また国会議事録中の発言と新聞社説を相互に学習・判定させることによって、政党の主義主張と新聞の論調との関連性を検証し、各新聞社がどの政党と類似した政治志向を有するかを検証する。

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。そうした政治的イデオロギー別に文章を分類したいという欲求も、知識人層には存在する。だが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、イデオロギーを測る客観的指標が得にくいことがある。

筆者らは、これまで全国紙を発行する新聞社の社説を教師信号とする手法を提案してきた[1][2]。しかし、新聞社の論調の違いは必ずしも全ての人にとってイメージできるものではない。

そこで本研究では、イデオロギーを測る客観的指標として政党の左右に着目する。主要な日本の政党は、左から右へ順に「共産党」「社民党」「民主党」「公明党」「自民党」と一般的に認識されている[3][4]。そこで、国会議事録に収められたこれらの党に属する議員の発言を教師信号とし、文章をイデオロギー別に分類することを試みる。そして、その分類の正当性を学習の結果得られたパラメータおよびクロスバリデーションの正答率で評価を行う。さらに、国会議事録を学習した結果得られた判別プログラムに各新聞社の社説を入力し、それぞれの新聞社がどの政党と類似した政治志向を有するかについて評価を行う。

本論文の構成は次の通りである。まず、2 章で実験に用いたシステムについて説明し、3 章で実験の説明と考察を行い、4 章でまとめを述べる。

2. システムの概要

本研究では、形態素解析ツールとして、ChaSen を用いる[5]。まず ChaSen を用い、電子化されている複数の文書データを形態素解析し、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する。

学習データを元に、機械学習のプログラムで文章の特徴を学習し、テストデータを元にシステムの精度を算出する。機械学習には最大エントロピー法を用いる[6]。システムの概要を図 1 に示す。最大エントロピー法のプログラムとしては maxent を利用する[7]。

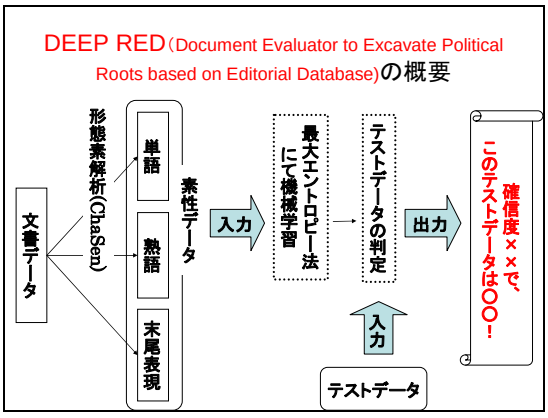


図 1　システムの概要

単語は名詞と動詞に限った。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とした。例えば、「アジア的優しさ」ならば、「アジア」：名詞、「的」：名詞、「優し」：形容詞、「さ」：名詞、

といったように ChaSen で分解されるが、それぞれが名詞、もしくは形容詞で、しかも名詞で終わっているため、熟語と判定され、素性となる。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。

また、新聞の社説をテストデータにする場合、漢数字を使う新聞社とアラビア数字を使う新聞社が混在するという問題がある。このような表記法の違いはイデオロギーとは全く関係ないが、数字の使用頻度が極めて高いこともあって、判定結果に大きな悪影響を与える。そこで、本研究の実験では学習データから数字を含む素性はすべて排除している。

3. 国会議事録

3.1 国会議事録の分析

本研究では 1999 年の第 145 回から 2008 年の第 169 回までの衆議院・参議院の議事録を国会会議録検索システム[8]より入手し、学習データに用いる。

議事録の構造は図 2 のようになっている。学習においては、それらの発言をある一定の単位でまとめて整理しなければならない。本研究では号 1 つ、つまりある日の委員会 1 日分の中の、発言者○○さんの発言すべてを 1 つの単位（新聞社説を学習するときの社説 1 つ分に対応）とする。

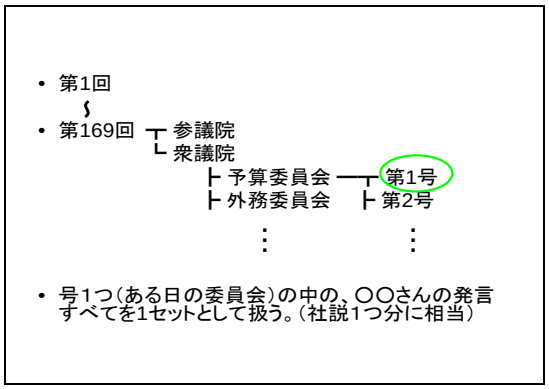
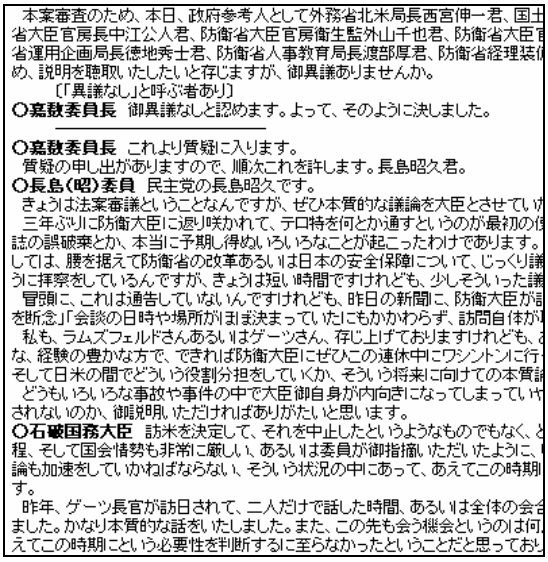


図 2　国会議事録の構造

3.2 学習データの作成

国会議事録検索システム内のデータには、会議での発言者とその所属政党を結びつけるデータベース的な情報は用意されていない。だが発言者の政党タグ付けは、政党別の左右を学習するには必須である。そこで、議事録内での「質問者」がその日の最初の登場時に自らの所属と名前を述べることに注目する。（図 3）


本案審査のため、本日、政府参考人として外務省北米局長西宮伸一君、国土
省大臣官房長中江公人君、防衛省大臣官房衛生監外山千也君、防衛省大臣官
省運用企画局長徳地秀士君、防衛省人事教育局長渡部厚君、防衛省経理装備
め、説明を聴取いたしたいと存じますが、御異議ありませんか。
〔「異議なし」と呼ぶ者あり〕
○嘉数委員長 御異議なしと認めます。よって、そのように決しました。

○嘉数委員長 これより質疑に入ります。
質疑の申し出がありますので、順次これを許します。長島昭久君。
○長島(昭)委員 民主党の長島昭久です。
きょうは法案審議ということなんですが、ぜひ本質的な議論を大臣とさせていた
三年ぶりに防衛大臣に返り咲かれて、テロ特を何とか通すというのが最初の使
誌の誤破棄とか、本当に予期し得ぬいろいろなことが起こったわけであります。
しては、腰を据えて防衛省の改革あるいは日本の安全保障について、じっくり議
うに拝察をしているんですが、きょうは短い時間ですけれども、少しそういった議
冒頭に、これは通告していないんですけれども、昨日の新聞に、防衛大臣が訪
を断念」「会談の日時や場所がほぼ決まっていたにもかかわらず、訪問自体が
私も、ラムズフェルドさんあるいはゲーツさん、存じ上げておりますけれども、と
な、経験の豊かな方で、できれば防衛大臣にぜひこの連休中にワシントンに行っ
そして日米の間でどういう役割分担をしていくか、そういう将来に向けての本質論
どうもいろいろな事故や事件の中で大臣御自身が内向きになってしまっていや
されないのか、御説明いただければありがたいと思います。
○石破国務大臣 訪米を決定して、それを中止したというようなものでもなく、と
程、そして国会情勢も非常に厳しい、あるいは委員が御指摘いただいたように、
論も加速をしていかねばならない、そういう状況の中にあって、あえてこの時期
す。
昨年、ゲーツ長官が訪日されて、二人だけで話した時間、あるいは全体の会合
ました。かなり本質的な話をいたしました。また、この先も会う機会というのは何
えてこの時期にという必要性を判断するに至らなかったということだと思っており

図 3　安全保障委員会の一部

この登場時の 1 行目の発言を ChaSen で形態素解析し、その中の単語に「日本共産党」「社民党（社会民主党）」「民主党」「公明党」「自民党（自由民主党）」のいずれか 1 つだけ存在した場合に、その党名で発言者と発言をタグ付けし抽出する。

また、質問者の名前に付く肩書きは、「君」「委員」「議員」「分科員」等になる。これ以外の「大臣」「参考人」「委員長」「政務官」「政務次官」「事務総長」「事務次長」「官房（副）長官」「政府委員」が含まれる場合は排除し、学習には議事録内での「質問者」に限ることにする。最終的に抽出された発言の件数は 35,456 件になる。

抽出した発言の件数を政党別に見てみると、「共産 6,016」「社民 4,058」「民主 13,649」「公明 5,041」「自民 6,692」となった。

また、発言 1 件に含まれる文字数について、500 文字間隔での頻度分布を図 4 に示す。

学習する際に、党によって発言の件数に差が大きいと、判定結果が件数の多い政党に近づいてしまう。そこで本研究の実験では、学習に採用する発言の文字数を 500~10,000 文字に限定し、一番件数の少ない社民党にあわせて、各党 3980 件の発言をランダムに選び、これらを学習データに採用する。

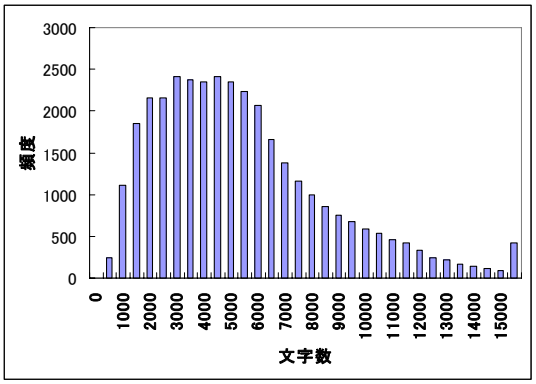


図４　発言文字数の分布（5 党の合計）

3.3　判定精度の検証実験

「共産」「社民」「民主」「公明」「自民」それぞれ 3,980 件の発言を学習し、10 分割のクロスバリデーションにて判定精度の検証を行ったところ、結果は 92.1%という異常に高い正解率となった。

判定システムがどういった素性を手がかりにして政党別に発言を分類しているかを見てみるため、政党別に判定に大きな影響を与えた上位の素性をいくつかピックアップしてみる。（表１）

表１　判定に影響を与えた素性の一部抜粋

共産党	社民党	民主党	公明党	自民党
日本共産党	社会民主党	新緑風会	改革クラブ	自由民主党
日本国憲法	護憲連合	無所属クラブ	神崎	地元
社会保障	福島瑞穂です。	EU	連立	我が国
破壊	辻元清美です。	分権	与党	国際社会
経団連	労働	慣習	外国人	宗教
労働者	原発	秩序	児童	国益
日弁連	差別	労働時間	PKO	日本国民
賞金	沖縄	安全保障上	伝統	法治国家
男女共同参画	イラク戦争	農業政策	革命	経済
ASEAN	非核	均等待遇	防衛	国防
侵略戦争	朝日新聞	国連中心主義	庶民	新しい憲法

思想を反映している素性も多くみられるが、発言者の所属や名前を含む素性もかなり存在し、これが判定を単純化させ悪影響を及ぼしていると考えられる。

そこで党名・所属・人名を含む素性をできる限り排除して判定精度の再検証を行った。末尾表現は、素性に採用する際に形態素解析していないため上で挙げた素性の排除が難しいことや、新聞の社説をテストする際、議事録の「ですます」調と社説の「である」調の違いのために有効な素性として機能しないと考えられるため排除した。

再度同様の条件にて判定精度を見てみたところ、正解率は 76.8%となった。各政党別に、どの政党の発言と分類されたかの割合を図５に示す。図の上から順に、共産党の発言をテストした場合、次が社民党の発言の場合を意味し、グラフの部分が学習結果によってどの政党の発言と判断されたかの件数の割合を示している。

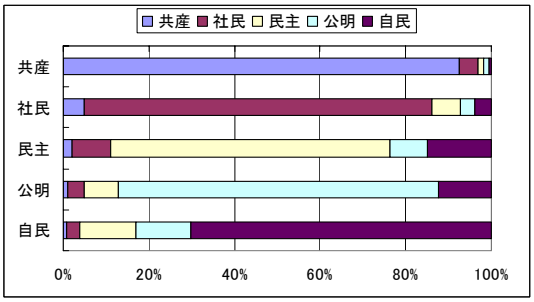


図５　クロスバリデーションの結果　党別正解率

共産党や社民党のような独自色の強い政党は正解率が高い。特に共産党は 93%の正解率を誇っている。対照的に民主党は 65%という一番低い正解率であった。これは、民主党が旧社会党系の議員から旧自由党系の議員までを抱えており、党としてイデオロギー的統一感が弱いことを考えると、妥当の結果と言えるかもしれない。

全体を俯瞰すると、正解である自党から離れた位置にいる党ほど、判定されている割合が少なくなっている傾向が見られる。前提としていた各政党の左右の定義に沿っている結果だと言えよう。

3.4　新聞の社説と政党の関連性

5 政党の議事録の発言を学習したシステムに、新聞の社説をテストにかける実験を行った。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、日本経済新聞が中道、読売新聞・産経新聞が保守・右派と位置付けられているおり[9]、これによって各新聞社の思想がどの政党とより関連性があるかを検証することができる。

使用する新聞の社説は、記事データベースから取得した毎日・日経・読売の 1999~2005 年分、および朝日は 2006~2007 年分[10]、産経は 2007/6～2008/8[11]である。結果を図６に示す。図の一番上の朝日から産経まで、下にいくほど右側の新聞社のテスト結果になっており、グラフの部分が、学習結果によってどの政党の発言と判断されたかの件数の割合を示している。

右寄りの新聞社ほど、自民党・公明党に近いと

判定されている割合が多い傾向が見られる。新聞の社説や議事録の政党別発言がうまく思想を反映している結果と考えられる。

ただし、どの新聞社も右派政党より左派政党のほうの割合が大きい、これは、マスメディアが全体的に左傾していることを示唆しており、多くの人の直感的印象に沿う結果と言えよう。また、読売新聞が保守・右翼系でありながら共産党と判定された割合が５社で一番大きいのも興味深い結果と言える。

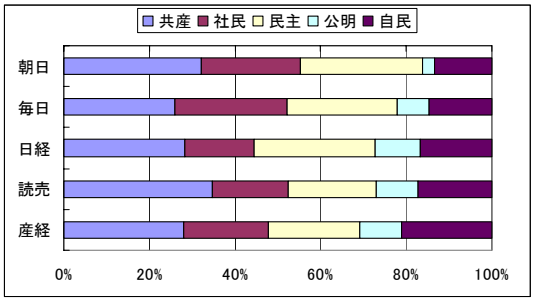


図６　新聞社の判定結果の比較

3.5　しんぶん赤旗

日本共産党が発行する機関紙「しんぶん赤旗」をテストデータとする実験を行った。最も左寄りの文章である赤旗の判定結果を見ることで機械学習の精度が確認できると期待される。

2005 年 2 月から 2008 年 9 月までの Web 上のしんぶん赤旗[12]の、社説に該当する「主張」（計 1314 件）を、国会議事録を学習したシステムに入力し判定を行った結果を表２に示す。正解率は非常に高いものとなり、本システムの有効性が示された。

表２　しんぶん赤旗のテスト結果

共産	社民	民主	公明	自民
1226	61	17	0	10

3.6　主観評価

判定システムの示す結果がどの程度妥当であるかを調べるため、政治的話題に明るい大学教員Ｋ氏に協力を得て主観評価実験を行った。Ｋ氏は自らを自民党と民主党の中間的な立場と自己分析している。Ｋ氏が過去に書き溜めた時事問題や社会問題に関するエッセイをテストデータとして計 108 件入力したところ、40 件が民主党、38 件が自民党と判定された。7 割以上を自民・民主両党が占め、両党ではその数が拮抗していることから、Ｋ氏の自己分析に近い判定結果が得られたと言える。

次に、それぞれのエッセイについて、出された判定結果がどの程度妥当であるかをＫ氏本人に 4 段階で評価してもらったところ、「正当 39」「許容 36」「やや不当 18」「不当 15」との主観評価が得られた。「正当」、「許容」と評価されたものが全体の７割程度となっている。「やや不当」「不当」と評価された文章についてＫ氏の見解を聞いたところ、ある党の主張を引用しながら批判している文章が、批判対象の党と判定される傾向があるとのコメントが得られた。こうした判定を回避する新たな対策を盛り込むことが今後の課題となる。

4.　おわりに

本研究では、政治的イデオロギーを測る客観的指標として国会の議事録に着目し、これらとの類似性をもとに文書のイデオロギー別分類を試みた。また、学習結果に基づき政党と新聞社の関連性を検証し、この手法によって言論のイデオロギー別分類が可能であることを示唆する良好な結果が得られた。

参考文献

[1] 畑中允宏，木村弦，金丸敏幸，村田真樹，掛谷英紀 (2007): 新聞の社説を教師信号とする文章の政治志向判定，第 3 回メディア情報検証学術研究会講演論文集

[2] 畑中允宏，金丸敏幸，村田真樹，掛谷英紀 (2008): 新聞の社説を教師信号とする文章の右翼度・左翼度判定 第二報，言語処理学会第 14 回年次大会講演論文

[3] Wikipedia 日本語版 － 「日本の政党一覧」等

[4] Wikipedia 英語版 - List of political parties in Japan

[5] 奈良先端科学技術大学院大学 松本研究室 ChaSen http://cl.aist-nara.ac.jp/

[6] Ristad, E. S. (1998). "Maximum Entropy Modeling Toolkit, Release 1.6 beta."1997 http://www.mnemonic.com/

[7] 内山将夫氏. maxent　http://www2.nict.go.jp/x/x161/members/mutiyama/software.html

[8] 国会会議録検索システム http://kokkai.ndl.go.jp/

[9] Wikipedia 英語版 － Japanese Media

[10] asahi.com　http://asahi.com/

[11] MSN 産経ニュース　http://sankei.jp.msn.com/

[12] しんぶん赤旗　http://www.jcp.or.jp/akahata/