政党のツイートデータを用いた
文章の右寄り左寄り分類システムの提案

伊能 沙知　野上 拓真　宮原 颯
アドバイザー教員　遠藤 靖典

1　序論

しばしば，政治に関する報道において右寄り左寄りといった言葉を耳にすることがあるが，用いる方も，それを受け取る方も，これらの言葉の意味を明確に意識しているかどうかは分からない．また，他者がある特定の意図を持って右寄り左寄りとラベル付けすることもあるだろう．政治に関する報道や思想は，時に国運を左右するほど重要なファクターとなることがあり，ある報道や思想に対して，何の根拠もなく右寄り左寄りの識別を行ったり，何の疑問もなくそれらの識別を受け入れることには大きなリスクが伴う．特に，外交問題をはじめとした政治に関する国民の関心が高まりを見せつつある昨今，この問題には細心の注意を払わなければならない．

右寄り左寄りという概念は，時代や場所によって多様に変化し，それらを超えた一義的な定義は困難である．しかし逆に，その時代や場所における右寄り左寄りを定義することはさほど困難ではない．それらの定義に即した識別規則を構成できれば，その識別規則に基づいて対象とする文章が右寄り左寄りのどちらに属するかを識別するアルゴリズムを構築することが可能であろう．そのアルゴリズムの識別結果の優位性を示すことができれば，上述のような政治思想リスクともいえる問題に，1 つの解決を示すことができると考えられる．

そこで本研究では，すでに広く普及しているツールであり，手軽に使用出来るウェブアプリケーションに着目し，政治思想への理解を支援するシステムの開発を目的とし，政党のツイッターデータを識別規則の構成に用いることで，入力文章を右寄り左寄りに識別するシステムを提案する．

2　手法

本演習で開発したシステムの概要は図 1 がリアルタイム処理を行う分類システムで，図 2 がバッチ処理で分類器を生成するサブシステムである．この実装したシステムの機能は，下記である．

(1) 右と左の教師データにおいて，ポジティブかネガ

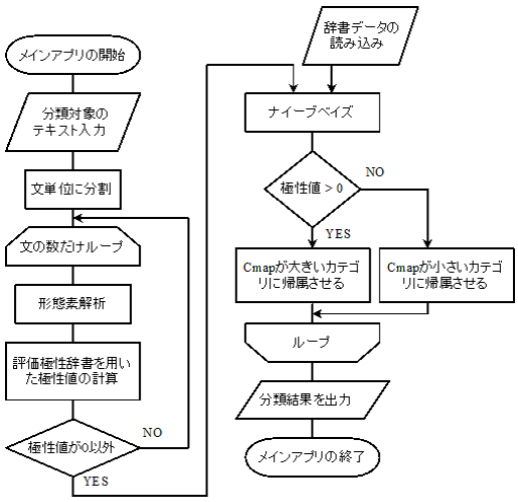

図 1: メインシステムの概要

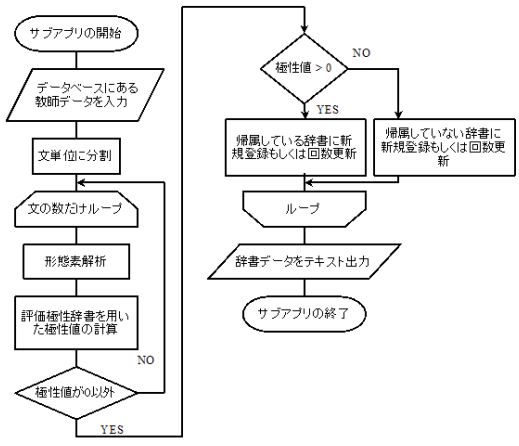

図 2: サブシステムの概要

ティブの両方の文の名詞のみを抽出．文章を区切るには

(1.1) 句点を基準とした．

(1.2) 句読点を基準とした．

そして，右と左の名詞のリストとその出現頻度を保持したリストをそれぞれ作成する．

(2) 同じ名詞がそれぞれのリストに存在する場合，

(2.1) 出現頻度が高い方のみを保持，低い方のリストから名詞を消去する．もし，出現頻度が同じ場合は，両方からその名詞を消去する．

(2.2 ）両方から名詞を消去する．

(3) 入力データにおいて，形態素解析，極性評価を行い，ポジティブな文とネガティブな文を対象とし，ナイーブベイズにより右と左の分類の評価値 c_{map} の計算を行う．そして，ポジティブな文なら c_{map} が高い方，ネガティブな文なら c_{map} が低い方のカテゴリーに分類する．

(4) 帰属するカテゴリーを出力．

以下で，このシステムに利用している教師データの選定，形態素解析，評価極性辞書，ナイーブベイズについて述べる．

2.1 教師データの選定

テキスト分類をするにあたって，教師データの選定が必要になる．関連研究として，畑中ら [1] は新聞の社説を教師データとして文章の右翼度・左翼度を分類するシステムを提案しており，高い正解率で判定が可能であると報告している．

我々はこの教師データとして，情報サービス”Twitter” で発信されるツイートデータに着目した。情報サービスが広く普及している現在，思想や政策についての主張のためにこれらのサービスを利用する国会議員は多い．特に Twitter はユーザーの数が多く，1 つの投稿につき 140 文字という制限によって手軽に意見を発信出来るため，多くの国会議員が利用している．そこで我々は，各政党を右翼と左翼に分類し，その政党に所属する議員のツイートデータを教師データとして用いることにした．

2.2 右翼と左翼の定義

右翼と左翼については，現在はっきりとした定義が存在せず，思想の区分が曖昧であり，多くの議論がなされている．[2, 3, 4, 5] 例えば，片山 [6] は著書「近代日本の右翼思想」において，右翼は過去や現在に足場を置く団体であり，左翼は未来に期待する団体であると述べている．

我々は各政党を歴史的背景や政治思想などから右翼を自由民主党，左翼を社会民主党と日本共産党と定義した．以下に，その根拠となる各党の歴史的背景や政治思想を示す．

2.2.1 自由民主党

戦後，独立体制の基礎固めの時代において，政界では自由民主陣営，革新陣営を問わず大きく動揺を続けていた．二大政党による健全な議会政治の発展という強い要望が国民と政治家の間に芽生え，国民世論の強い要望と自由民主主義政党内部での反省も加わり，1955 年，自由党と日本民主党の「保守合同」により結成された保守政党として自由民主党が結成された．同年の「立党宣言」で政治の使命は民生の安定，公共福祉の増進，自主独立，平和の確立とし，立党の政治理念は議会民主政治と，個人の自由と人格の尊厳とした [1]．

平成 22（2010）年綱領 [2] では現状認識として，天皇の下に，また日米同盟を基軸とする外交政策で平和な日本を作り上げたとした．立党目的のうち「反共産・社会主義，反独裁・統制的統治」は達成されたが，独自の伝統・文化の喪失，経済成長の鈍化，財政悪化，少子化などの現実があり，もう 1 つの立党目的である「日本らしい日本の確立」が重要とした．常に進歩を目指す保守政党であり，自由主義，民主制，秩序の中の進歩，真実を語る，多様な組織との対話を掲げ，政策の基本的考えとしては，新憲法の制定，自主防衛，自助自立する個人の尊重，市場経済，公正な政策，財政の効率化と税制改正を掲げる．誇りと活力ある日本像を目指して，家族・地域社会・国への帰属意識，合意形成を怠らぬ民主制，国債残高減額，世界平和への義務を掲げる．現在の主な動きとしては，復興，経済再生，教育再生，外交再生等を指針として活動している．

2.2.2 日本共産党

日本共産党は第二次世界大戦後に公然活動を開始した，化学的社会主義を党是とする政党である．1951 年に「51 年綱領」と「軍事方針」を決定し，これに基づいて武装闘争を繰り広げた．1958 年には「51 年綱領」を廃止し，1961 年に現状規定や二段階革命方式等を規定した現綱領を採択する．その後党勢拡大を図り，1970 年代には党員，議席共に伸長させていったが，その後東欧の社会主義体制が一挙に瓦解し，共産主義イデオロギーの破綻が明らかとなったことで党勢は停滞し，徐々に減少していく．近年は大幅な規約改定を行い，マルクス・レーニン主義特有の用語や国民が警戒心を抱きそうな表現を削除，変更するなど，ソ

[1] https://www.jimin.jp/index.html
[2] http://www.jimin.jp/aboutus/pdf/kouryou.pdf

フトイメージを強調したものとしているが，基本路線に変更はない．

日本を「高度に発達した資本主義国でありながら，国土や軍事等の重要な部分をアメリカに握られた事実上の従属国となっている」と現状認識し，「民主主義革命」によって対米従属から抜け出して真の主権を回復した後，「社会主義的変革」によって一切の強制のない，真に平等で自由な人間関係からなる共同社会を目指すとしている．

2.2.3 社会民主党

社会民主党は社会民主主義を掲げる政党として「平和・自由・平等・強制」を日本における民主主義利の理念としている．

1945 年，戦前の社会大衆党などを母体に，前身である日本社会党が結成される．社会大衆党は更に，左派の労働農民党，中間派の日本労農党，右派の社会民衆党などが合同したものであった．3 派による派閥対立を経て，1955 年に社会党統一を果たし，野党第一党の地位を得て保守勢力に対する革新勢力の中心として存続した．1960 年代には内部の派閥抗争や当時の社会主義に幻滅を与える数々の事件により，党勢は停滞，微減した．1986 年には従来の平和革命による社会主義建設を否定し，自由主義経済を認める「日本社会党の新宣言」が決定されるが党勢は退潮し，1993 年には 55 年体制以来最低の議席数となる．1994 年には自社さ政権である村山政権が発足するが，村山首相は就任直後に，安保条約肯定，原発肯定など，旧来の党路線の変更を一方的に宣言し，党の求心力を大きく低下させる．村山内閣総辞職後，イメージの一新のために社会民主党に改称するが，同時に民主党の結成を機に約半数の党所属国会議員が社民党を去る．その後も社民党の衰退に歯止めはかからず，現在は野党第一党としての立場を失っている．

2.3 形態素解析

このテーマにおける分類を実装するには，自然言語処理，つまり日本語を処理させる必要がある．ここで代表的で最も基本的な技術として形態素解析がある．形態素解析とは，自然言語で書かれた文を形態素（意味を持つ最小の言語単位）に分割する技術のことをいう [3][4]．代表的な形態素解析エンジンとして，MeCab[5] がある．Mecab はオープンソースの形態素解析エンジンで，文を入力すると，単語で分けた後，品詞，活用形，読み等多くの情報を出力してくれる．このようなツールを用いて自然言語の情報を抽出した後，テキスト分類を行う．例として「憲法は改正すべきだ」という文の形態素解析の結果を表 1 に示す．

表 1: 形態素解析の例

元の文章	憲法は改正すべきだ					
分かち書き	憲法	は	改正	す	べき	だ
品詞	名詞	助詞	名詞	動詞	助動詞	助動詞 記号
基本形	憲法	は	改正	する	べし	だ
活用形	*	*	*	文語基本形	体言接続	基本形

2.4 評価極性辞書

評価極性辞書とは，名詞や用言などの語句に対して極性を付与したものをまとめた辞書の事である．多くの場合は，語句に対してポジティブとネガティブの 2 つの極性を付与し，その極性を利用して語句あるいは文章の分類分けを行う．[7]

例えば，同じ原子力発電に関係する文章でも「原子力発電は安全であり，推進していくべきだ」という文章と「原子力発電は危険であり，直ちに廃止すべきだ」という文章では，その内容は正反対のものとなる．両者の文章を比較すると，前者の文の「安全」や「推進」といった単語にはポジティブなニュアンスが含まれており，後者の文の「危険」や「廃止」といった単語にはネガティブなニュアンスが含まれている．つまり，その文章に含まれるポジティブな語句とネガティブな語句の量を比較することによって，その文章が原子力発電に対して肯定的な文章なのか，否定的な文章なのかを判断する．これを応用して，文章の類似度に加えて，極性も考慮した上で分類を行う事が出来る。

2.5 ナイーブベイズ

ナイーブベイズ [8] は，テキスト分類において利用される技法の 1 つである．実装が簡単で高速という特徴から，精度評価のベースラインとしてよく使われる．ナイーブベイズによるテキスト分類を行うには，まず人間が教師データを与えて学習を行う必要がある．また，テキストを分類するにあたっては，単語の集合として与えられる場合が多く，単語が文章内のどこに出てくるかは考慮されない．このようなテキスト表現を bag-of-words といい，この処理を行うに当たっては形態素解析を利用するのが一般的である．

ナイーブベイズには多項モデルとベルヌーイモデルの 2 つがあるが，今回は分類器に利用した多項モデル，つまり多項ナイーブベイズについて詳しく述べる．まず、文書 d がカテゴリ c に属す事後確率 $P(c|d)$ は，ベイズの定理より，

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \propto P(c)P(d|c) \qquad (1)$$

と表せる．$P(c)$ は文章がカテゴリ c に含まれる事前確率，$P(d)$ が文章が生成する事前確率，$P(d|c)$ が

[3] http://gengoro.zoo.co.jp.

[4] http://www.sist.ac.jp/~kanakubo/reserch/natural_language_processing.html.

[5] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

あるカテゴリ c が与えられたときに文章 d が生成される事後確率を示す．この式から，カテゴリを予測したい未知の文章は，事後確率 $P(c|d)$ が最も高いカテゴリへ分類する．この確率を求めるには，$P(d)$ はどのカテゴリにおいても共通であるため無視し，$P(c)$ と $P(d|c)$ のみを計算すればよい．

まず、$P(c)$ は，教師データの各カテゴリ文章数の総文章数に占める割合を計算する事で求める．例えば，教師データ 100 文章中で右寄りの文章が 60，左寄りの文章が 40 であったとすれば

$$P(c = \text{右}) = \frac{60}{100} = 0.6$$
$$P(c = \text{左}) = \frac{40}{100} = 0.4$$

となる．次に $P(d|c)$ は，文章 d は bag-of-words で n_d 個の単語の集合 $[w_1, w_2, ..., w_{n_d}]$ として表し，単語間の独立性を仮定すると下記のように計算できる．

$$\begin{aligned} P(d|c) &= P(w_1 \wedge \cdots \wedge w_{n_d}|c) \\ &= \prod_{1 \leq i \leq n_d} P(w_i|c) \end{aligned} \tag{2}$$

同時確率をそれぞれの確率の積で表せるのが確率論的独立性の定義であるため，上の式で第 2 式から第 3 式へは単語の出現確率の間に独立性を仮定しないと成り立たない．ところが，本来は単語の出現に独立性は成り立たず，例えば,「リスク」と「マネジメント」は共起しやすい,「テキスト」と「分類」は共起しやすいといったような関連性が生じてくる．ナイーブベイズではこれを無視して，単語の出現は独立と無理矢理仮定して文書の確率を単語の確率の積で表して単純化を行うといった特徴を持つ．

(2) 式から，$P(w_i|c)$ が必要になる．これは，教師データのカテゴリ c に単語 w_i が出てきた回数をカテゴリ c の全単語数で割れば求める事が出来る．そこで $T(c, w)$ をカテゴリ c に単語 w が出てきた回数，V を教師データ中の全単語集合とすると

$$P(w_i|c) = \frac{T(c, w_i)}{\sum_{w' \in V} T(c, w')} \tag{3}$$

となる．分母は V のすべての単語に関して足し合わせているが，実際は対象カテゴリ c に出てくる単語に絞っても結果は同じとなる．これは，そのカテゴリに出てこなかった単語は $T(c, w) = 0$ となるためである．

以上から，最終的に分類されるカテゴリである最大事後確率（MAP）カテゴリ c_{map} を求めると，

$$\begin{aligned} c_{map} &= \arg\max_c \hat{P}(c|d) \\ &= \arg\max_c \hat{P}(c) \prod_{1 \leq i \leq n_d} \hat{P}(w_i|c) \end{aligned}$$

となる．P に対して $\hat{P}$ と書いているのは，真の値ではなく推定値だからである．$\hat{P}(w|c)$ というのは非常に小さい数な上に，文章中には多くの単語が含まれるため積部分がアンダーフローを起こす可能性がある．そこで，事後確率の大小関係は対数を取っても変化しないことを利用し

$$\begin{aligned} c_{map} &= \arg\max_c \log \hat{P}(c|d) \\ &= \arg\max_c \left\{ \log \hat{P}(c) + \sum_{1 \leq i \leq n_d} \log \hat{P}(w_i|c) \right\} \end{aligned}$$

というように表す事が出来る．ナイーブベイズのほとんどの実装がこの式である．

ナイーブベイズにおいてはアンダーフロー以外に，ゼロ頻度問題という問題がある．ゼロ頻度問題とは，未知の文章のカテゴリを予測する際，教師データのボキャブラリに含まれない単語を 1 つでも含んでいると単語の条件付確率 $\hat{P}(w|c)$ が 0 となり，単語の条件付き確率の積で表される $\hat{P}(d|c)$ も 0 となる，つまり新しい文章が生成される確率が 0 となってしまう現象である．この問題を解決するには，単語の出現回数に 1 を加えるラプラススムージングが用いられる．これを踏まえると，最終的な条件付き確率の式は

$$\begin{aligned} \hat{P}(w_i|c) &= \frac{T(c, w_i + 1)}{\sum_{w' \in V} (T(c, w') + 1)} \\ &= \frac{T(c, w_i + 1)}{\sum_{w' \in V} T(c, w') + |V|} \end{aligned}$$

となる．

3 数値例

システムの機能について述べたように，(1) について (1.1)，(1.2) の 2 パターン，(2) について (2.1)，(2.2) の 2 パターンがあることから，組み合わせると計 4 パターンが考えられる．したがって，今回それぞれについて評価を行う．本システムの汎化能力を評価する手法として Leave-one-out Cross Validation[9] を用いた．まず Cross Validation とは，データをテスト用と学習用と分け，はじめに学習用データで学習した後，残しておいたテスト用データで性能評価を行う，という試行をデータの分け方を変更しながら繰り返し実行し，結果を平均するといった方法である．その中でも，Leave-one-out Cross Validation はテスト用データを一つのみとし，他を学習用データとする手法である．

本システムで利用した教師データは，右派ツイートが 102 個，左派ツイートが 111 個である．その結果，それぞれの手法について得られた評価を表 1 に示す．それぞれの手法の評価では，(2.2) よりも (2.1) を用い

た方が良い評価を受け，(1.1) よりも (1.2) を用いた方が良い結果となった．したがって最も良い結果が (1.1) と (2.2) の組み合わせとなった．(1.2) よりも (1.1) の方が精度が良い理由としては，評価極性辞書を用いていることによる弊害が影響していることが考えられる．今回評価極性辞書で極性評価を出来ない文については分類対象外としているため，句読点で文を区切ると，一つの文章が短くなり，極性を示す単語が含まれていない場合が増え極性評価が出来ず，教師データとして扱われなかった結果，分類に用いられる辞書の精度が低下したことが考えられる．また，(2.1) よりも (2.2) の方が精度が良い理由としては，本システムの教師データにはツイートを用いており，「私たち」等の日常的によく使われる単語が非常に多く登場している．そこで，単語が右左両方の辞書で登場した場合，出現回数が多いものを保持するよりも．両方の辞書から消すことにより，日常的によく使われ，分類にはほぼ関連しない単語が多く消去されたことで，分類精度の向上に繋がったということが考えられる．

これら理由から，(1.1) と (2.2) を組み合わせたシステムが最も分類精度が良かったと考えられる．

表 2: Leave-one-out Cross Validation によるシステムの評価

(1)	(2)	CV
(1.1)	(2.1)	53.38%
(1.1)	(2.2)	57.43%
(1.2)	(2.1)	52.03%
(1.2)	(2.2)	55.41%

これらの考察から，1 ツイートを一つの文と見なした上，単語が右左両方の辞書で登場した場合，両方の出現回数の差を出現回数が多い方の値とし，少ない方の辞書から消すという方法にシステムを変更したところ，Leave-one-out Cross Validation の成功率は 62.43% となった．

また，班員が自ら選出した右寄りと左寄りそれぞれに代表されると思われる個人や団体が示した文章 20 文ずつを本システムにおいて分類し，その結果について班員自身で解釈するという方法を行った．その結果，右寄りの文章 20 文のうち，右寄りが 8 文，左寄りが 8 文，分類不可が 4 文で，左寄りの文章 20 文のうち，右寄りが 8 文，左寄りが 7 文，分類不可が 5 文であった．

選出した右寄りと左寄りの文章について，正しく分類出来なかったものについて例文を示し，その考察を行う．

選出した右寄り文章

・安倍首相が「アジア太平洋の大きな自由経済圏の第一歩にしなければならない」と述べたのはもっともだ

この文章では，本システムで作成した教師辞書で右寄りと左寄りに含まれる単語がそれぞれ存在したことから，右と左の評価値としてはほぼ同じくらいの値であったものの強制的に左に分類されてしまっていた．

選出した左寄りの文章

・安倍政権が意欲を燃やす憲法改正に対して各界からの批判の声は日増しに高まりを見せています

この文章は憲法改正を批判した文章であるが,「意欲」や「高まり」等のポジティブとされる単語から全体が肯定的な文章と捉えられ，結果として憲法改正に対して肯定的な文章と判断され，右に分類されてしまったと考えられる．この対策としては，現在の文章の極性を評価するだけでなく，文章を構造的に捉え分析を行うアルゴリズムの構築等が考えられる．

その他の改善点として,「ない」という単語における打ち消しの表現を考慮する必要性も考えられる．打消しの表現で用いられる「ない」は評価極性辞書では極性を与えられておらず、文意を反転させる効果を持つ．例えば「自衛隊は平和のために維持していくべきだ。」という文章と,「自衛隊は平和のために維持すべきではない。」という文章では，文中に使用される単語の極性は同じであるが，文末の「ない」という表現によって文意は真逆となっているため，このような表現方法を考慮するアルゴリズムが必要であることが伺える．

4 結論

本演習では，政治思想への理解を支援するツールの開発を目的として，政党のツイッターデータを用いたテキスト分類システムを提案した．システムの構築に利用した技術として，形態素解析，評価極性辞書，ナイーブベイズを用いた．また，分類アルゴリズムについては 4 パターンを提示し，それぞれ Leave-one-out Cross Validation を用いて汎用性のあるシステムであるかの評価を行った．また，評価データにおいて実際に文章を分類した結果，評価極性辞書に該当する単語が存在しないため分類できなかった文が多いことや,「ない」による文意の反転，文章を構造的に捉える必要性など多くの問題点があることが明確に示せた．

今後の課題として，分類精度の向上や異なったアプリケーション等への応用が挙げられる．分類精度の向上の方策としては，教師データの拡充や評価辞書の

作成，分類手法の変更，実験での問題点の改善等が考えられる．また，アプリケーション等への応用については，現時点ではWebアプリケーションへの応用が実現しており，今後はより手軽に利用できるスマートフォンアプリ等への実装が考えられる．

参考文献

[1] 畑中允宏, 金丸敏幸, 村田真樹, 掛谷英紀 , 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, 2008.

[2] 岡崎 正道, 日本の左翼と右翼の源流, 岩手大学人文社会科学部, 言語と文化・文学の諸相, pp.105-119, 2008.

[3] 堀幸雄, 最新右翼辞典, 柏書房, 2006.

[4] アントニオ ネグリチ, 未来派左翼〈上・下〉―グローバル民主主義の可能性をさぐる, NHK ブックス, 2008.

[5] 磯田 光一, 鈴木邦男「腹腹時計と狼」極左・極右近接の精神状況 (思想と潮流), 朝日新聞社, 朝日ジャ-ナル , 17, 54, pp.57-59, 1975.

[6] 片山杜秀, 近代日本の右翼思想, 講談社, 2007.

[7] 東 宏一, 国会議員のツイッター分類とその応用, 筑波大学大学院 システム情報工学研究科 知能機能システム専攻 修士論文, 2012.

[8] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, 情報検索の基礎, 共立出版, 2012.

[9] カーネル多変量解析-非線形データ解析の新しい展開, 赤穂昭太郎, 岩波書店, 2008.