政党のツイートデータを用いた右寄り左寄り分類システムの提案

8班
伊能 沙知
野上 拓真
宮原 颯
アドバイザー教員　遠藤 靖典

目次

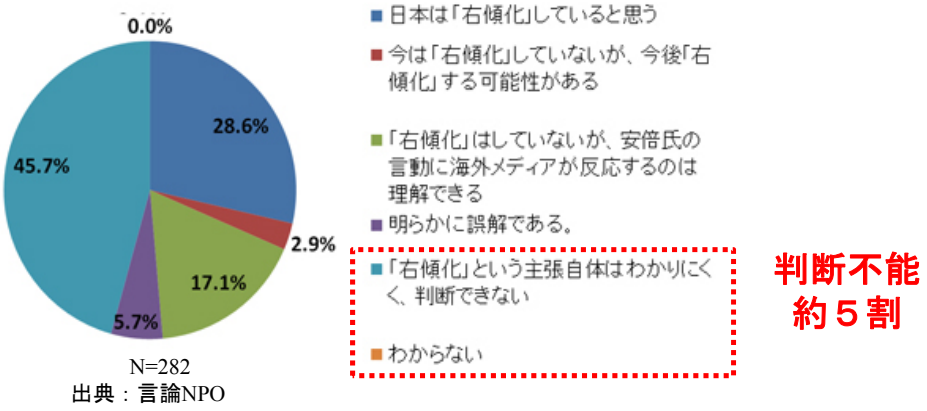
目次

- はじめに
  - 背景
  - 目的
- 開発したシステムについて
  - システムの概要
  - 利用した技法
  - システムの評価
- おわりに
  - まとめ
  - 今後の課題

背景

近年の政治報道の主張

日本は「右傾化」してきている！！



N=282
出典：言論NPO

判断不能 約５割

- どちらに偏っているかを認識していないのではないか？
- 政治思想は国の命運を左右する ➡ 思想がどちらに偏っているかを認識する必要

政治思想のリスク

目的

政治思想についての正しい理解を支援するツールの必要性

文章が右寄り・左寄りどちらの政治思想に近いかを分類するシステムの開発

ツールは多くの人が手軽に利用できることが望ましい

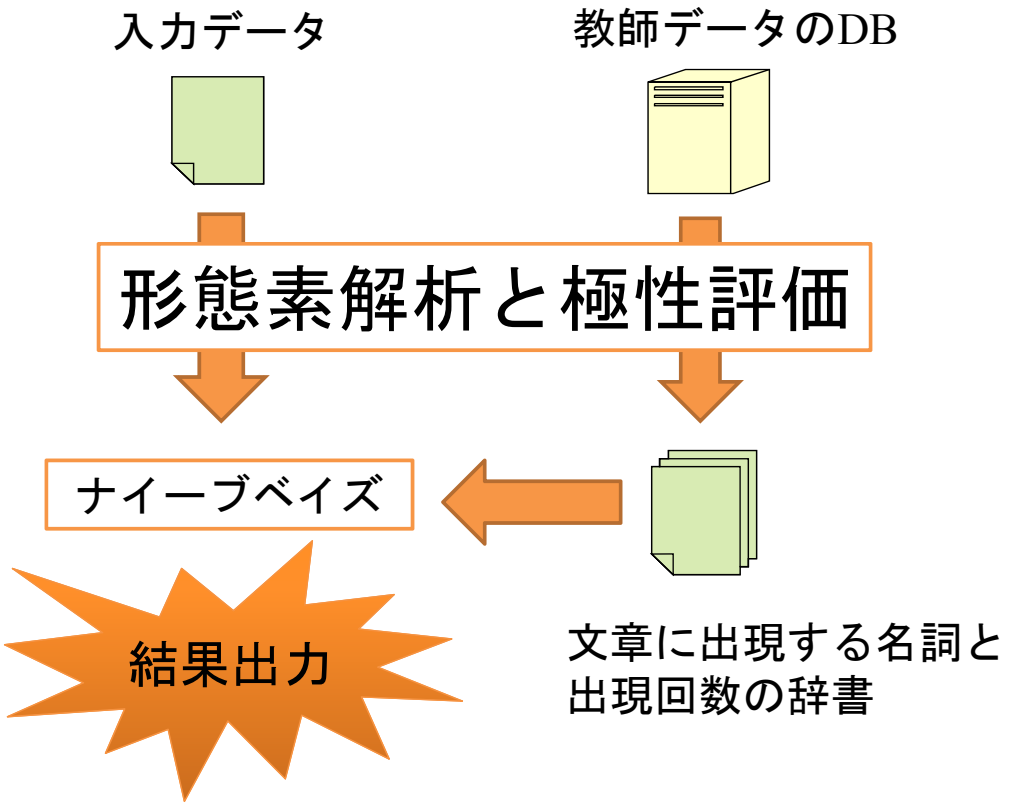
➡ ウェブアプリケーション

- ✓ 広く普及している
- ✓ 幅広い年齢層が利用できる
- ✓ 処理が早い

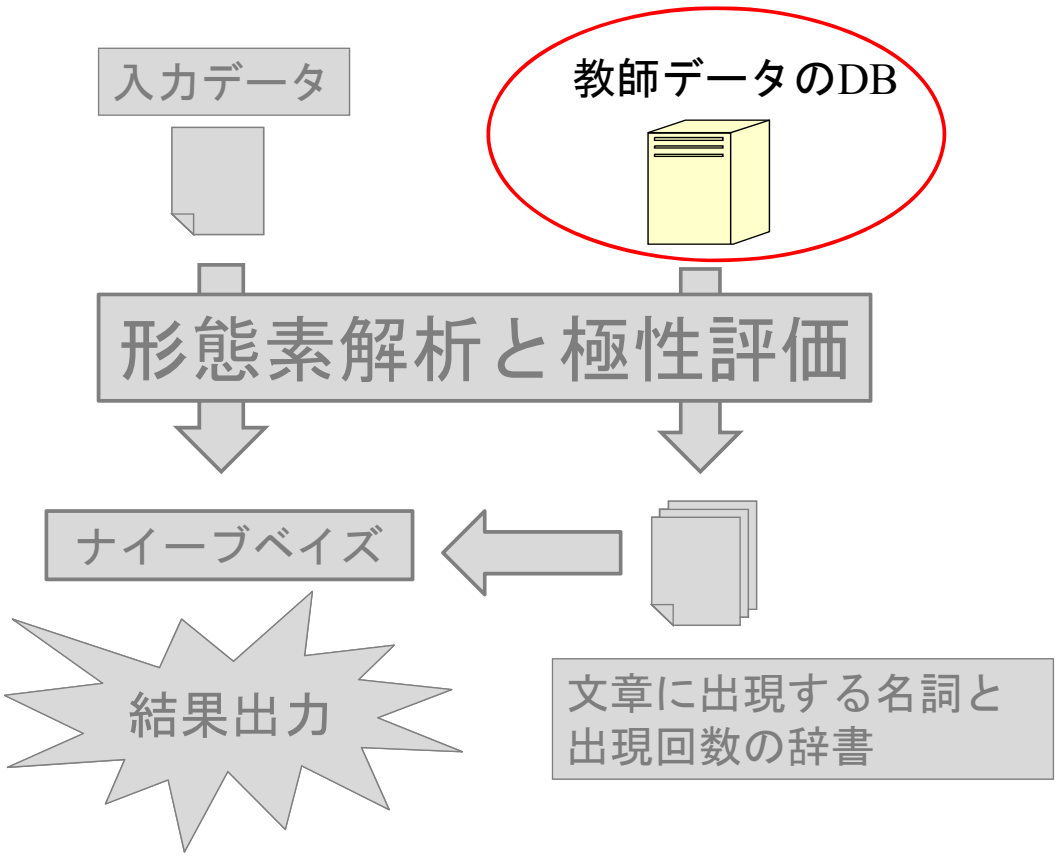
目次

- はじめに
  - 背景
  - 目的
- 開発したシステムについて
  - システムの概要
  - 利用した技法
  - システムの評価
- おわりに
  - まとめ
  - 今後の課題

システムの概念図



システムの概念図



教師データの選定

関連研究：畑中ら（2008）

右：読売新聞　左：毎日新聞
として，社説を教師データに用いて文章の右翼度・左翼度を分類

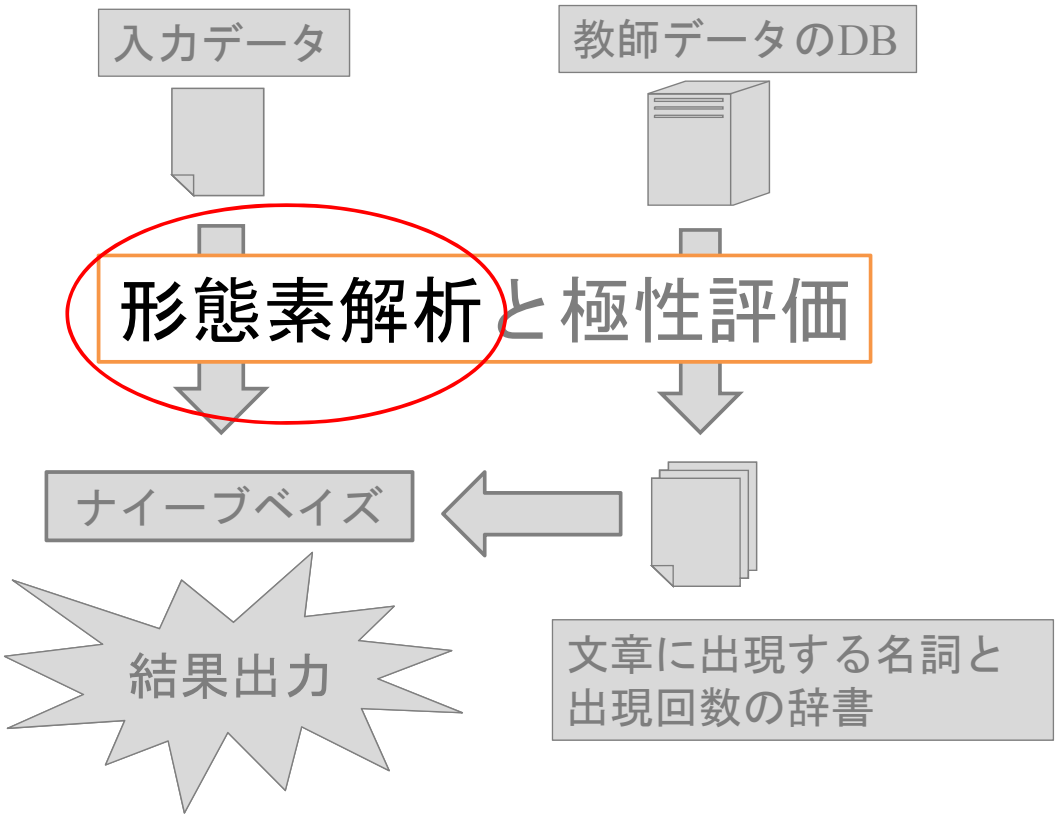
情報サービス“Twitter”に着目

- 140文字までの投稿制限があり，手軽に情報を投稿・閲覧することが可能
- SNSの中でもユーザー数が多い

➡ ネット選挙解禁により，国会議員も積極的に利用している

政党の歴史的背景，政治思想から
右：自由民主党　　左：社会民主党，日本共産党
と定義し，所属する議員のツイートデータを教師データに用いる

システムの概念図



形態素解析

自然言語で書かれた文を形態素に分割する技術

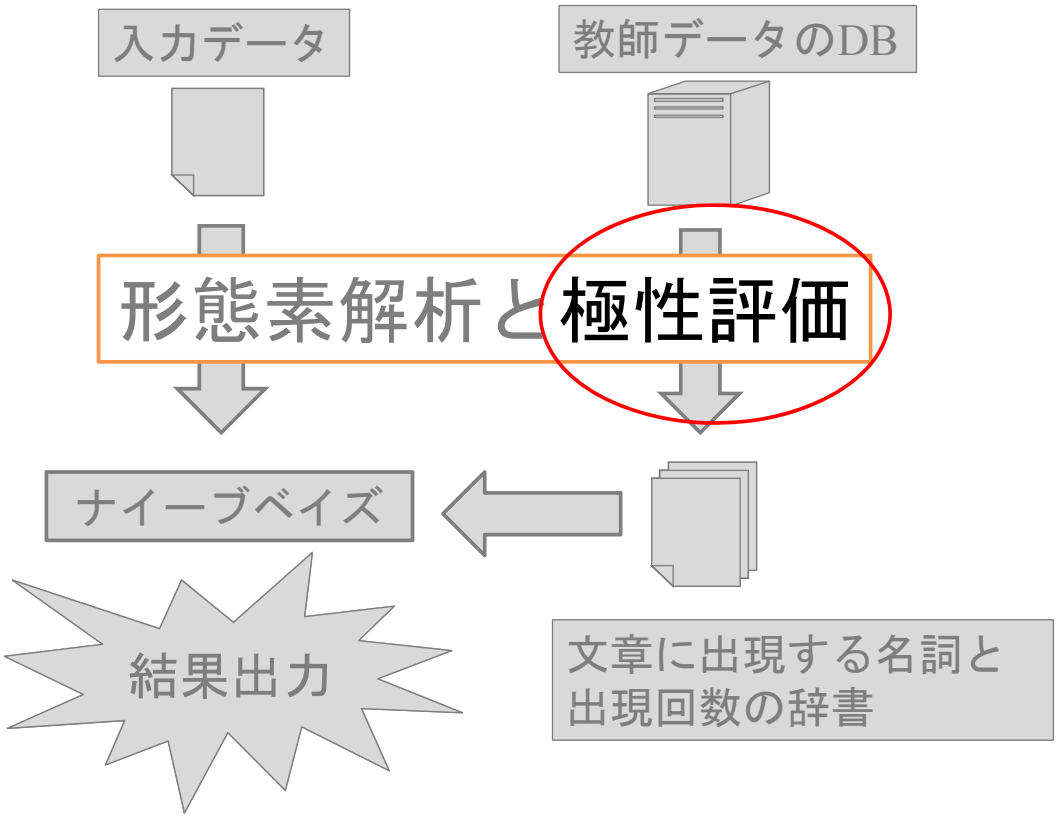
名詞，用言等意味を持つ最小の言語単位

代表的な形態素解析エンジン：MeCab（オープンソース）

形態素解析の例（Mecab）

分かち書き	憲法	は	改正	す	べき	だ
品詞	名詞	助詞	名詞	動詞	動助詞	助動詞
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

システムの概念図



評価極性辞書

名詞や用言などに対して極性を付与したものをまとめた辞書

ポジティブ or ネガティブ

この極性辞書を利用して文章の極性評価を行う事が可能！

例えば・・・

原子力発電は安全であり，推進していくべきだ　肯定的

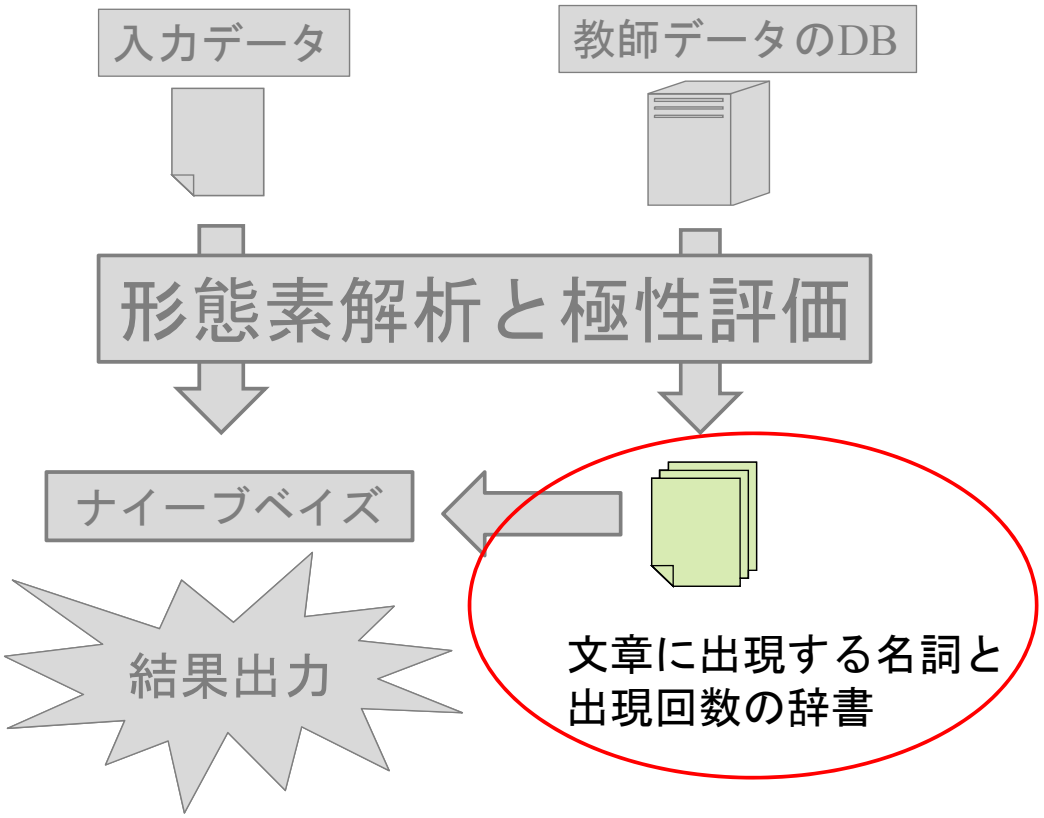
正反対

原子力発電は危険であり，直ちに廃止すべきだ　否定的

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数を比較

➡ 肯定的 or 否定的な文かを評価

システムの概念図



教師辞書の作成方法

ツイートデータの文章を句切る

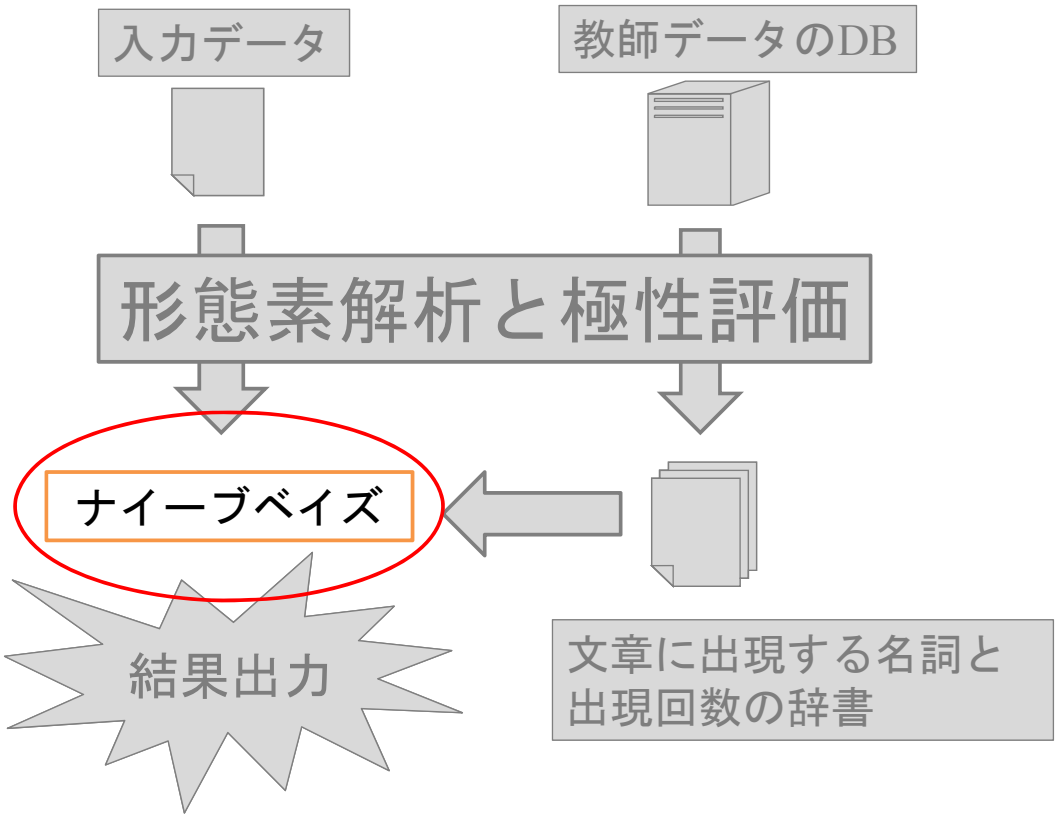
- 句点で句切る
- 句読点で句切る

極性を持つ文から名詞を抽出

各カテゴリに対して名詞とその出現頻度を保持した辞書を作成

- 同じ名詞は出現頻度が高い方を保持
- 同じ名詞は両方削除

システムの概念図



ナイーブベイズ①

ベイズの定理

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

※　c：カテゴリ　d：文章

$P(c)$：カテゴリ（右・左）の事前確率

右の教師データが60文，左の教師データが40文であれば

$$P(c = 右) = \frac{60}{100} = 0.6$$

$$P(c = 左) = \frac{40}{100} = 0.4$$

ナイーブベイズ②

ベイズの定理

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

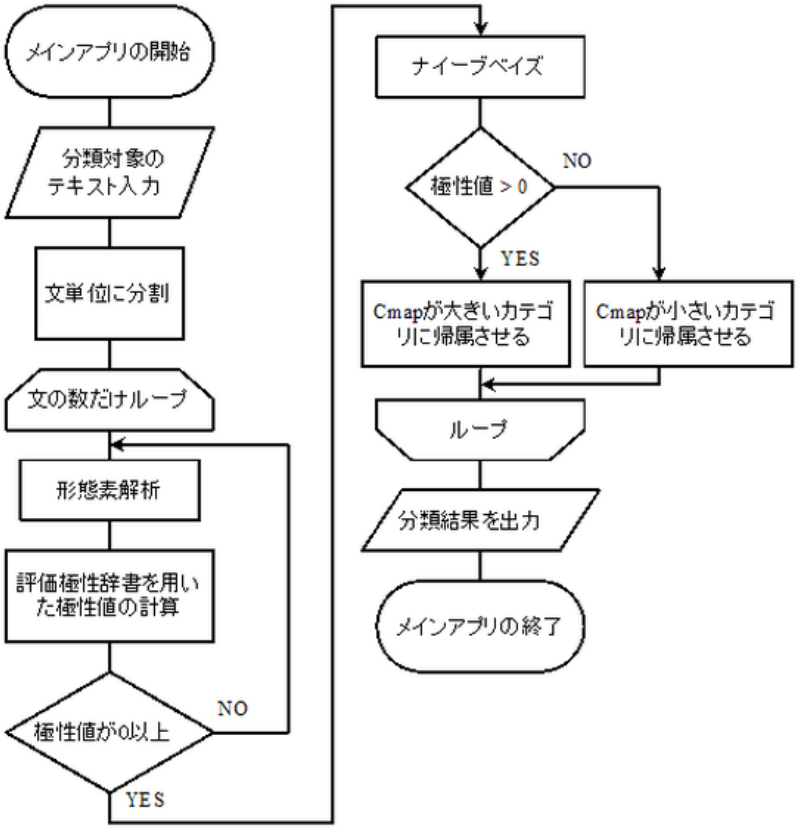
$P(d|c)$文章の事後確率

単語間の独立性を仮定した場合・・・

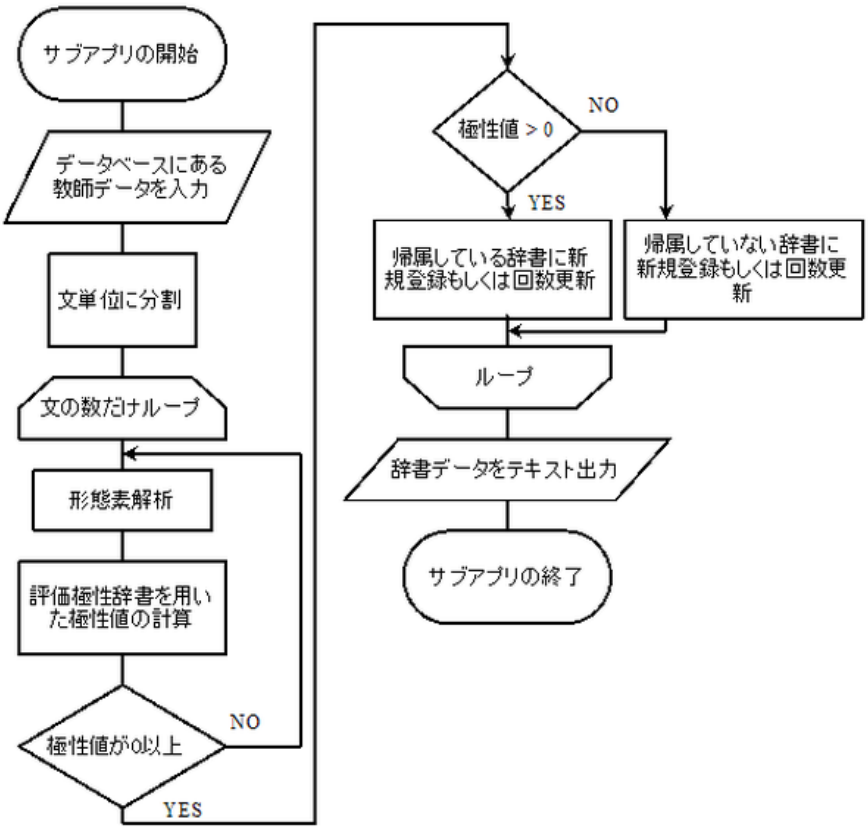
$P(\omega_1 \wedge \cdots \wedge \omega_{n_d}|c) = \prod_i P(\omega_i|c)$：単語の事後確率の積

$$P(\omega_i|c) = \frac{\text{カテゴリ}(c)\text{に単語}(\omega_i)\text{が出てきた回数}}{\text{カテゴリの全単語数}}$$

フローチャート①



フローチャート②

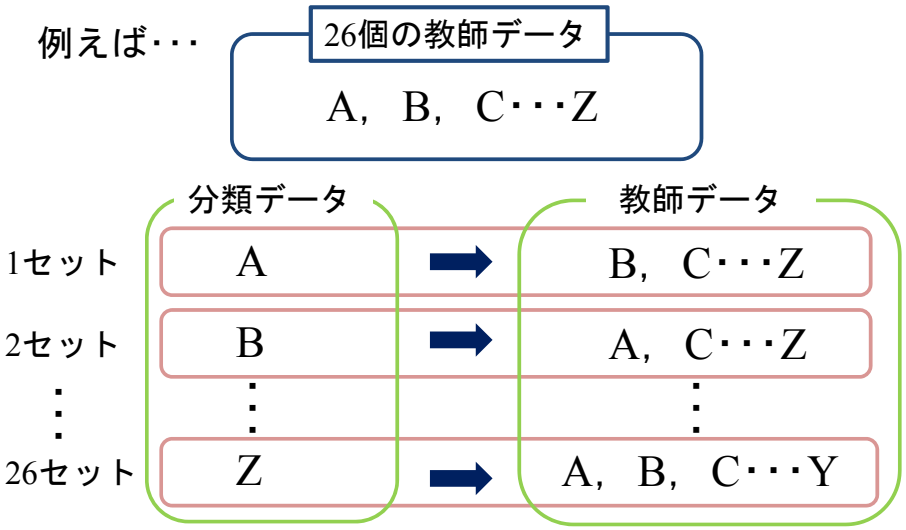


システムの評価 ①

Leave-one-out Cross Validation

教師データを1つ抜き出して分類データとし，残りの教師データで訓練された分類器で分類

➡ これを全ての組み合わせで行う



システムの評価 ②

文の区切り	被った名詞の扱い	CV
句点	多い方を保持	53.38 %
句点	両方消去	57.43 %
句読点	多い方を保持	52.03 %
句読点	両方消去	55.41 %

考察

- **句点＞句読点**

文章が短くなり，極性を持つ単語が含まれない場合が増える．

➡ 極性評価が出来ず教師データとして扱われない文が増え，教師辞書の精度が低下．

- **多い方を保持＞両方消去**

「私たち」等の日常的によく使われる，右左の分類には関連しない単語が多く消去，分類精度が向上．

システムの評価 ③

合計40の検証データを最も評価結果の良かったシステムで分類

- **右寄りと思われる検証データ20文**
  - ➡ 右に分類：8つ，左に分類：8つ，分類不可：4つ
- **左寄りと思われる検証データ20文**
  - ➡ 右に分類：8つ，左に分類：7つ，分類不可：5つ

※分類不可：文章中に極性，あるいは教師辞書中の単語がない場合

考察

- **教師辞書の精度**
  思想に関係の無さそうな語句によって分類されてしまっている.
- **文章の構造が捉えられない**
  語句の極性の数のみで文章の極性を判断してしまう.
- **打消しの表現の判定**
  「～ない」といった文意を反転させる語句を判断出来ていない

システムの評価 ④

誤分類の例-1

・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。

➡ 「圏」など**思想には関係のなさそうな語句**で分類されてしまっている.

誤分類の例-2

・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

➡ 批判的な文章であるが，「意欲」や「高まり」等の**ポジティブと判断される語句の方が多い**ため，肯定的な文と判断されてしまっている.

目次

- はじめに
  - 背景
  - 目的
- 開発したシステムについて
  - システムの概要
  - 利用した技法
  - システムの評価
- おわりに
  - まとめ
  - 今後の課題

まとめ

- 政治思想への理解を支援するツールとして，政党のツイートデータを用いたテキスト分類システムを開発
- Web アプリケーションとして実装
- システムの構築に利用した技法
  - 形態素解析
  - 評価極性辞書
  - ナイーブベイズ
- システムの評価
  - Leave-one-out Cross Validation
  - 任意データの分類評価

今後の課題

- 分類精度の向上
  - 教師データの拡充
  - 誤分類の原因を考慮した分類手法の改善
- 異なったアプリケーション等への応用
  - スマートフォンアプリ等への実装

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 Sch utze, 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.