概要

事物に対するポジティブ/ネガティブという評価極性は議論が多いが，事物に対する違法性・危険性の視点からの評価極性は議論が少ない．ネガティブな事象を観測した際，⟨恐れ⟩，⟨怒り⟩，⟨悲しみ⟩，⟨嫌だ⟩などの感情が存在する．違法性が強ければ⟨怒り⟩に傾き，危険性が強ければ⟨恐れ⟩に傾くと予想される．違法性・危険性の強弱が分かれば感情分析の役に立つと考えられる．そこで，本研究では違法性や危険性を表す表現の自動収集を目的とする．

本研究では *SO-score* による算出方法を使用する [1]．*SO-score* とは，各単語について好評表現および不評表現との共起数の対比較により値を算出するものである．また，本研究では話題を指定する．これは一般に「ホテル」の評判分析ではホテルの文書集合から情報を抽出することと同じ考え方である．これにより，話題を指定し，その話題に依存した表現を抽出できる．

本研究の流れは以下の通りである．まずブログ記事を n 分割し n 個の文書集合にする．次に各文書集合から指定する話題を含む文を抽出し，これらを話題文書集合とする．さらに話題文書集合から危険性/安全性/違法性/合法性のある文を抽出する．具体的には文字列のマッチングによる粗い抽出の後に因果関係を学習したSVMでフィルタリングを行い，危険/安全/違法/合法文書集合を得る．話題文書集合から単語リストを作成し，その各単語について *SO-score* を算出する．*SO-score* は1単語につき n 通りが得られるので，単語ごとの *SO-score* の分布を求める．この分布により，安定して出現し，かつ，極性がはっきりしている単語を選び出す．

性能評価は，抽出した単語の精度と再現率で評価する．本研究では話題を「自転車」，「自動車」,「病院」とする．結果として，ヒストグラムを用いることでかなり単語の抽出数を絞り込むことができ，精度を上げることができた．ヒストグラムを用いない場合は抽出数が膨大なため，再現率は上がるが精度が下がった．以上により，危険性/違法性を表す単語を安定して自動収集する一つの方法を示すことができた．

目 次

図 目 次

表 目 次

第1章　はじめに

事物に対するポジティブ/ネガティブという評価極性は議論が多いが，事物に対する違法性・危険性の視点からの評価極性は議論が少ない．ネガティブな事象を観測した際，〈恐れ〉，〈怒り〉，〈悲しみ〉，〈嫌だ〉などの感情が存在する．違法性が強ければ〈怒り〉に傾き，危険性が強ければ〈恐れ〉に傾くと予想される．違法性・危険性の強弱が分かれば感情分析の役に立つと考えられる．そこで，本研究では違法性や危険性を表す表現の自動収集を目的とする．

本研究では *SO-score* による算出方法を使用する [1]．*SO-score* とは，各単語について好評表現および不評表現と共起数する確率により値を算出するものである．危険性については，危険だと思われる表現（例えば「危険」）と安全だと思われる表現（例えば「安全」）を *SO-score* に適応させる．違法性についても同様である．

また，本研究では話題を指定する．これは一般に「ホテル」の評判分析ではホテルの文書集合から情報を抽出することと同じ考え方である．これにより，話題を指定し，その話題に依存した表現を抽出できる．

ここで問題点は，コーパス全体から *SO-score* を算出すると，偶然良い値や悪い値が算出される可能性が高く，かつ，抽出される単語の数が多すぎることである．そこで，本研究ではコーパスを分割し，*SO-score* を多数算出して，統計をとる方法を提案する．この方法により，安定して出現し，違法性・危険性（もしくは合法性・安全性）のある単語を抽出することができると考えられる．

本論文の構成は以下の通りである．第 2 章では，先行研究について述べる．第 3 章では，提案手法の流れを述べる．第 4 章では，実装について述べる．第 5 章では，実験について述べる．第 6 章では考察を行う．最後に第 7 章でまとめを述べる．

第2章　先行研究

本章では，先行研究について述べる．まず，感情分析における極性について述べる．次に評価極性の算出方法として Turney[1] が提案した *SO-score* について述べる．

2.1　感情極性抽出

2.1.1　感情極性

文章内に含まれる感情（意見や態度を含む）の発見，特定，および分析は，様々な応用可能性を有する重要な課題である．感情分析における重要なリソースの 1 つとして，単語の感情極性が挙げられる．単語の感情極性とは，ある単語が良い印象を持つ（ポジティブ）か，それとも悪い印象を持つ（ネガティブ）かを示す二値変数である．単語の感情極性は，その後の感情・意見分析で基礎的な役割を果たす．

2.1.2　単語感情極性対応表

高村ら [2] は，日本語および英語の単語とその感情極性の対応表を示している．

感情極性値は，語彙ネットワークを利用して自動的に計算されたものである．もともと二値属性だが，-1 から$+1$ の実数値を割り当てている．-1 に近いほどネガティブ，$+1$ に近いほどポジティブと考えられる．表 2.1 に対応表の一部を示す．

表 2.1: 単語感情極性対応表の一部

優れる：すぐれる：動詞：1
良い：よい：形容詞：0.999995
喜ぶ：よろこぶ：動詞：0.999979
褒める：ほめる：動詞：0.999979
めでたい：めでたい：形容詞：0.999645
賢い：かしこい：形容詞：0.999486
善い：いい：形容詞：0.999314
適す：てきす：動詞：0.999295
天晴：あっぱれ：名詞：0.999267
祝う：いわう：動詞：0.999122
功績：こうせき：名詞：0.999104
賞：しょう：名詞：0.998943
・
・
・
苦しい：くるしい：形容詞：−0.999788
苦しむ：くるしむ：動詞：−0.999805
下手：へた：名詞：−0.999831
卑しい：いやしい：形容詞：−0.99986
ない：ない：形容詞：−0.99986
浸ける：つける：動詞：−0.999947
罵る：ののしる：動詞：−0.999961
ない：ない：助動詞：−0.999997
酷い：ひどい：形容詞：−0.999997
病気：びょうき：名詞：−0.999998
死ぬ：しぬ：動詞：−0.999999
悪い:わるい：形容詞：−1

2.2 *SO-score*

評価極性を分類する処理についての議論は多いが，その中でも，コーパスから得られる共起情報を用いて語句の評価極性値（評価極性の傾向を示す値）を判定する手法を Turney は提案した [1]．国際辞書などのエントリ情報を用いないため，見だし語単位やエントリ単位，複数語からなる句に対しても評価極性値を判定することができる．主に，好評表現と不評表現の出現比率を用い，好評表現の方が多い場合は好評，逆なら不評とした．Turney は *SO-score* を算出することで，これを示した．

2.2.1 *SO-score* の算出方法

ある語句 t の極性評価値 *SO-score(t)* は以下の式より算出する．ここで，*PMI* とは，2 つの語句間の共起を図る尺度を表す．

$$\textit{SO-score}(t) = \textit{PMI}(t, \text{“好評表現”}) - \textit{PMI}(t, \text{“不評表現”}) \tag{2.1}$$

$$\textit{PMI}(a, b) = \log_2 \frac{p(a, b)}{p(a)p(b)} \tag{2.2}$$

$p(a, b)$ はコーパス内において語句 a と語句 b が同一文で共起する確率，$p(a)$ は語句 a を含む文がコーパス内に出現する確率をそれぞれ表す．*SO-score(t)* で語句 t が “好評表現” と多く共起しやすければ，正に大きい値をとり，“不評表現” と多く共起しやすければ，逆に負に大きい値をとる．確率が 0 となる語句に関しては，log に 0 が入ってしまうのを避けるために，Turney は出現頻度に 0.01 を足している．また，*SO-score* を算出する際に，好評文と不評文での出現頻度が共に 4 より小さい評価表現は有効なデータとして扱わないこととしている．

第3章　提案手法

本研究の提案手法の流れを示す．まず問題点を提示し，提案手法の手順を述べる．その後に各手順について詳しく説明する．

3.1　問題設定

本研究では *SO-score* による算出方法を使用する．ここで問題点は，コーパス全体からの *SO-score* の算出では偶然性が高く，かつ，単語の候補が多すぎることである．そこで，本研究では単語ごとに *SO-score* を多数算出し，統計的に安定化する方法を提案する．

3.2　手順

本研究ではブログ記事を使用する．まずブログ記事を n 分割し n 個の文書集合にする．次に各文書集合から指定する話題を含む文を抽出し，これらを「話題文書集合」とする．さらに話題文書集合から危険性/安全性/違法性/合法性のある文を抽出する．この際，不要な文を取り除くためのフィルタを用意し，各文書集合をフィルタにかける．得られた文書集合をそれぞれ,「危険文書集合」,「安全文書集合」,「違法文書集合」,「合法文書集合」とする．これにより話題文書集合と，危険/安全/違法/合法文書集合がそれぞれ n 個作成される．話題文書集合に出現する単語から単語リストを作成し，話題文書集合と，危険/安全/違法/合法文書集合における各単語の出現回数をカウントする．その各単語について *SO-score* を算出する．*SO-score* は 1 単語につき n 通りが得られるので，単語ごとにヒストグラムを作成し，単語ごとの *SO-score* の分布を求める．分布により，安定して出現し，かつ，評価極性がはっきりしている単語を選び出す．

以上の手順の流れを図 3.1 に示す（図では危険性を例に挙げる）．

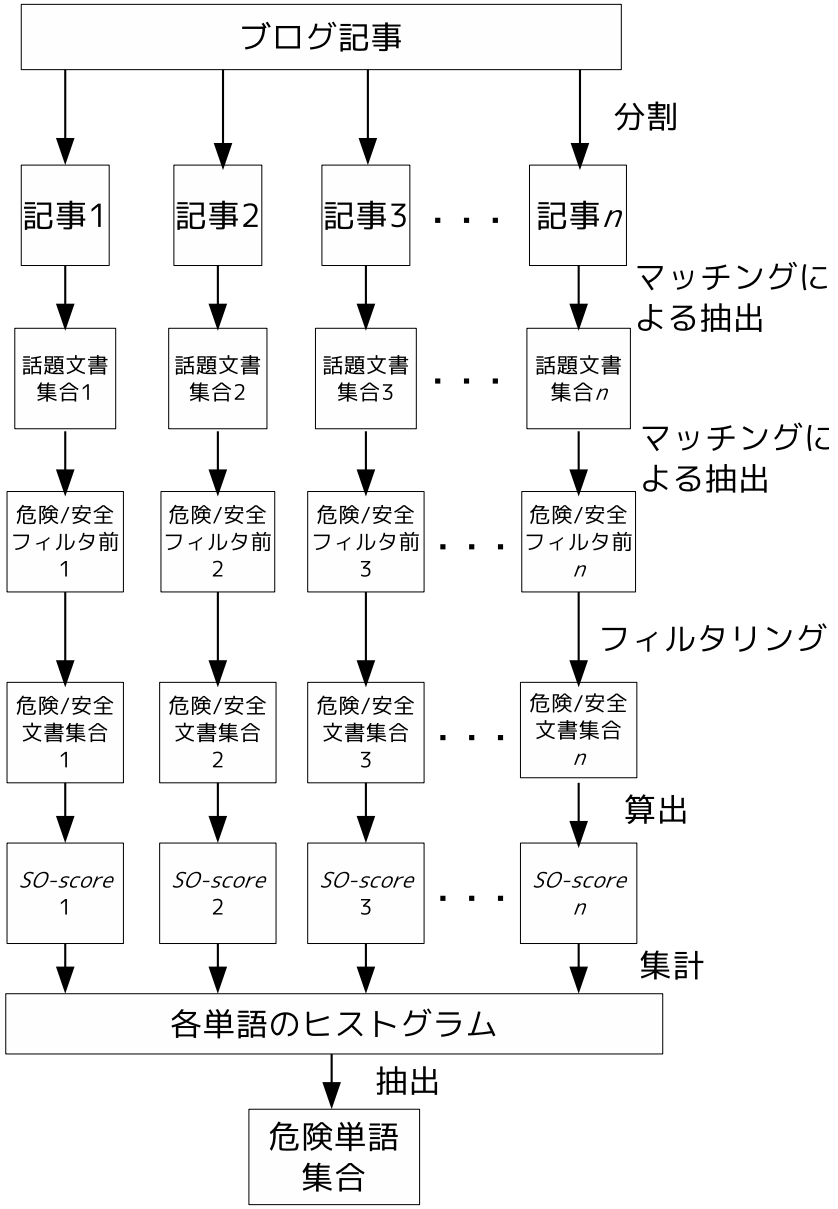

ブログ記事
分割
記事1
記事2
記事3
記事n
マッチングによる抽出
話題文書集合1
話題文書集合2
話題文書集合3
話題文書集合n
マッチングによる抽出
危険/安全フィルタ前1
危険/安全フィルタ前2
危険/安全フィルタ前3
危険/安全フィルタ前n
フィルタリング
危険/安全文書集合1
危険/安全文書集合2
危険/安全文書集合3
危険/安全文書集合n
算出
SO-score 1
SO-score 2
SO-score 3
SO-score n
集計
各単語のヒストグラム
抽出
危険単語集合

図 3.1: 提案手法の流れ

3.3 コーパスの分割

ブログ記事を分割することで文書集合に出現する単語にランダム性を持たせる．これは，時期ごとに分割すると「雪道」など時期に依存する単語が安定して出現しないためである．

分割方法を説明する．ブログ記事は日付ごとにまとめられている．分割方法としては，日付ごとに通し番号を付け，その番号を分割数 n で割った余りによってグループ分けを行う．

3.4 話題文書集合の作成

話題文書集合の作成について述べる．話題文書集合は，各話題（例えば「自転車」）を含む文をブログ記事から集めてきたものである．

3.4.1 マッチング

文の抽出はマッチングにより行う．マッチングとは，特定の文字列のパターンが各文において当てはまるかを調べることである．本研究では，指定した文字列が各文に含まれるかを調べることによってマッチングを行う．

3.4.2 文の抽出

分割した各ブログ記事から指定する話題を含む文を抽出する．指定する話題が含まれるかどうかはマッチングにより判定する．これにより n 個の話題文書集合が得られる．

3.5　各評価極性ごとの文の抽出

危険文書集合，安全文書集合，違法文書集合，合法文書集合の作成について述べる．まず，文の抽出について述べ，次にフィルタについて述べる．

3.5.1　粗い文の抽出

話題文書集合から危険性/安全性/違法性/合法性のある文を抽出する．各評価極性を表す単語（例えば「危険」）を用いてマッチングを行う．ここでで用いる単語については5章で述べる．

3.5.2　フィルタリング

文章には多くの表現が存在する．その中でも,「〜だから危険」という意味をもつ文は,「〜」の部分に危険性がありそうだが,「危険だから〜」という意味をもつ文はその部分においてむしろ危険を回避しているように思われる．そのため,「危険だから〜」という意味をもつ文を取り除き,「〜だから危険」という意味をもつ文を抽出するフィルタを作成した．

SVM

フィルタとしてSVMを用いる．

SVM(Support vector machine)は，教師あり学習を用いる識別手法の一つで，線形入力素子を利用して2クラスのパターン識別器を構成する手法である．訓練サンプルから，各データ点との距離が最大となるマージン最大化超平面を求めるという基準で線形入力素子のパラメータを学習する．

素性

素性はデータ点を表す要素の種類である．本研究で用いる素性は「危険」あるいは「危＋⟨ひらがな一文字⟩」より前に出現する単語列には「L：」という文字列を，後に出現する単語列には「R：」という文字列を付与したものである．単語列にはunigramとbigramを使用しており，unigramについては助詞を取り除いている．この素性を与えることで,前と後に出現する単語の傾向をつかみ,「〜だから危険」と「危険だから〜」という文の

分類ができると考えた．なお，危険性以外の場合でも同じトレーニングデータを使用できるように，素性を与える際に「危険」および「危＋⟨ひらがな一文字⟩」を含む単語は除いている．また，句読点は素性には含めない．

SVM のトレーニングデータは「自転車」の話題についての 806 文で，正例が 423 文，負例が 383 文である．トレーニングデータの例を以下に示す．トレーニングデータは左から，id，正解値，素性となっている．正解値は +1 の場合は危険性があるとする．

トレーニングデータの例

原文： でも雨の中音楽聴きながら自転車乗ってなおかつ片手で傘さしてたら危険やん

B1 +1 L1:でも L1:雨 L2:でも雨 L2:雨の L1:中音 L2:の中音 L1:楽 L2:中音楽 L1:聴き L2:楽聴き L2:聴きながら L1:自転 L2:ながら自転 L1:車 L2:自転車 L1:乗っ L2:車乗っ L2:乗って L1:なおかつ L2:てなおかつ L1:片手 L2:なおかつ片手 L2:片手で L1:傘 L2:で傘 L1:さしてた L2:傘さしてた L1:ら L2:さしてたら R1:ん R2:やん

原文： 大阪は雨で自転車は危ないので、職場に置いて帰りました

B2 -1 L1:大阪 L2:大阪は L1:雨 L2:は雨 L2:雨で L1:自転車 L2:で自転車 L2:自転車は R1:職場 R2:ので職場 R2:職場に R1:置い R2:に置い R2:置いて R1:帰り R2:て帰り R2:帰りまし R2:ました

安全性/違法性/合法性についても同様のトレーニングデータを用いて，フィルタにかける．フィルタにかけることで得られる文書集合をそれぞれ，危険文書集合，安全文書集合，違法文書集合，合法文書集合とする．

3.6 単語リストの作成

話題文書集合に出現する名詞を抽出し，単語リストを作成する．名詞は一般名詞と用言性名詞のみとする．これは，数詞や代名詞なども名詞ではあるが，本研究では必要ないからである．

作成した単語リストにおける各単語について，出現回数を，話題文書集合，および，危険/安全/違法/合法文書集合からカウントする．

3.7 *SO-score*の算出

2.2 節で述べた *SO-score* による算出方法を本研究では使用する．

SO-score の計算式における好評表現と不評表現を極性に合わせて設定する．危険性の視点からの場合，好評表現を安全性のある表現（安全表現）とし，不評表現を危険性のある表現（危険表現）とする．これを適応させた *SO-score* の計算式を以下に示す．

$$SO\text{-}score(t) = PMI(t, \text{“安全表現”}) - PMI(t, \text{“危険表現”}) \tag{3.1}$$

違法性の視点からの場合，好評表現を合法性のある表現（合法表現）とし，不評表現を違法性のある表現（違法表現）とする．これを適応させた *SO-score* の計算式を以下に示す．

$$SO\text{-}score(t) = PMI(t, \text{“合法表現”}) - PMI(t, \text{“違法表現”}) \tag{3.2}$$

PMI の値は話題文書集合ならびに危険/安全/違法/合法文書集合から求める．

3.7.1 式の拡張

SO-score を求める際に，評価極性を表す単語として複数の単語を用いる場合がある．その場合には，*PMI* の算出式を拡張することで *SO-score* を算出する．その式を以下に示す．

$$SO\text{-}score(t) = PMI(t, P) - PMI(t, N) \tag{3.3}$$

$$PMI(a, X) = \log_2 \frac{p(a, X)}{p(a)p(X)} \tag{3.4}$$

$$p(X) = \frac{1}{N} \left| \bigcup_{x \in X} s(x) \right| \tag{3.5}$$

$$p(a, X) = \frac{1}{N} \left| \bigcup_{x \in X} s(a, x) \right| \tag{3.6}$$

ここで，t は単語とし，P, N は単語の集合とする．$s(x)$ はコーパスにおいて単語 x を含む文の集合を返す関数とし，$s(x, y)$ はコーパスにおいて単語 x, y の両方を含む文の集合を返す関数である．$P(X)$ は集合 X の要素のいずれかが出現する確率を求める関数である．$|\ |$ は集合の要素数を表す．N はコーパスの総文数を表す．

例えば，$t =$ “片手”，$X =$ {“危険”，“危ない”，“危うい”} とした場合，

$$p(X) = \frac{1}{N} |s(\text{“危険”}) \cup s(\text{“危ない”}) \cup s(\text{“危うい”})|$$

$$p(\text{片手}, X) = \frac{1}{N} |s(\text{“片手”，“危険”}) \cup s(\text{“片手”，“危ない”}) \cup s(\text{“片手”，“危うい”})|$$

となる．

3.8 統計

分割したブログ記事から，各単語について多数の *SO-score* が得られた．その統計を算出する．

3.8.1 ヒストグラムについて

ヒストグラムとは，縦軸に度数，横軸に階級をとった統計グラフの一種で，データの分布状況を視覚的に認識することができる．主に統計学や数学，画像処理等で用いられる．

3.8.2 ヒストグラムの設定

本研究ではヒストグラムを以下のように設定する．

階級は *SO-score* の値によるものとし，階級の幅を w とする．この際，$-w/2$ 以上 $w/2$ 未満の階級が存在する．度数の総和は最大で n（コーパスの分割数）となる．

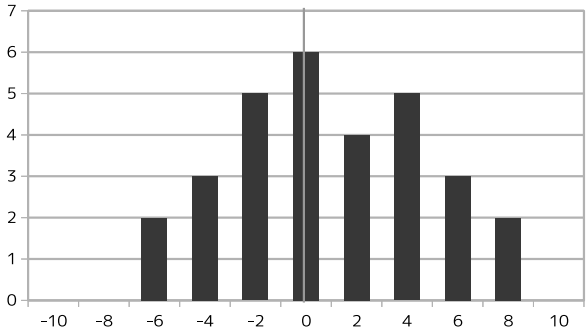
例えば，階級の幅を $w = 2$ とし（以降，例を挙げる際には階級の幅を $w = 2$ とする），*SO-score* の値が -2.531 の場合，-3 以上 -1 未満の階級に分布されることとなり，その階級の度数に 1 が加算される．これにより，各単語についての *SO-score* の分布状況を知ることができる．

3.9 評価極性の決定

作成したヒストグラムに対して，単語の選択の視点から 2 つの条件を設け，さらに極性の決定の視点から 1 つの条件を設ける

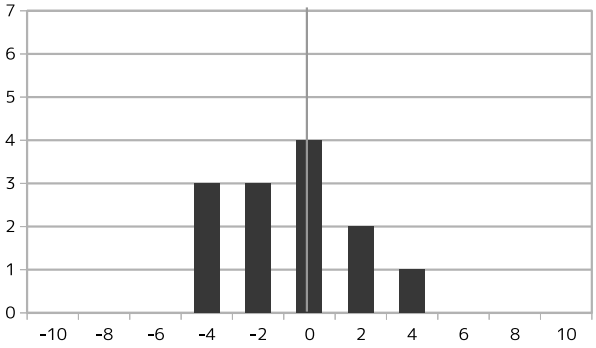
3.9.1 単語の選択の視点からの条件

条件 1： 度数の総和が n であるヒストグラム．n はコーパスの分割数である．n 分割した文書集合全てで出現している単語のヒストグラムは，度数の総和が n となる．例を図 3.2 に示す．



度数の総和＝2+3+5+6+4+5+3+2＝30

条件を満たす



度数の総和＝3+3+4+2+1＝13

条件を満たさない

図 3.2: 条件 1 の例

条件 2： $-w/2$ 以上 $w/2$ 未満の階級の度数が最大でないヒストグラム．この階級の度数が最大である場合は，極性が好評あるいは不評に傾いていないということになる．例を図 3.3 に示す．

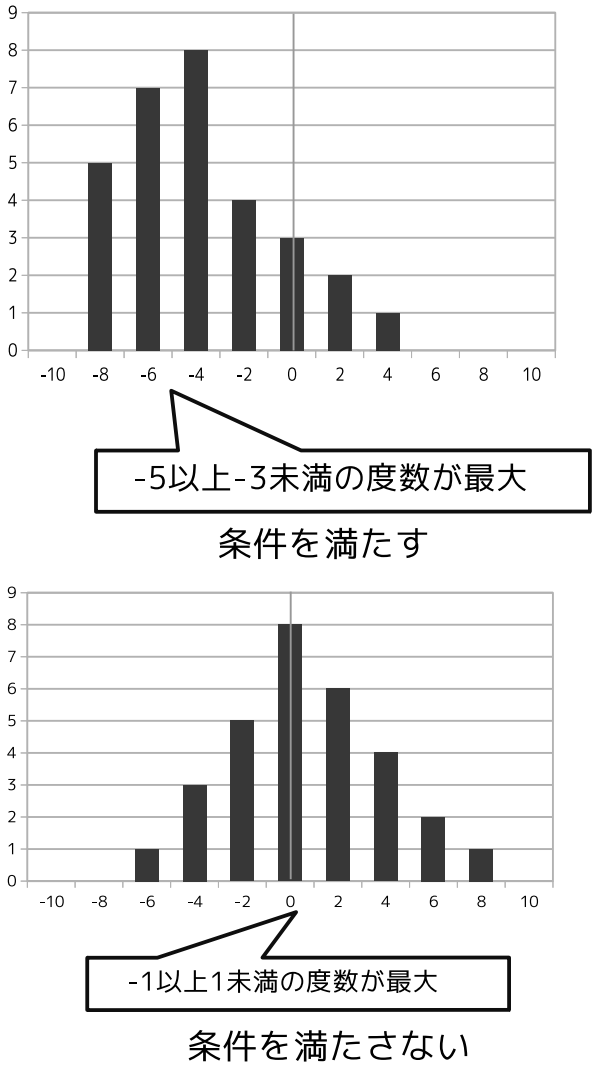


図 3.3: 条件 2 の例

3.9.2　極性の決定の視点からの条件

条件 3： 最大度数が負の階級であれば危険性あるいは違法性があり，最大度数が正の階級であれば安全性あるいは合法性があるとする．例を図 3.4 に示す．

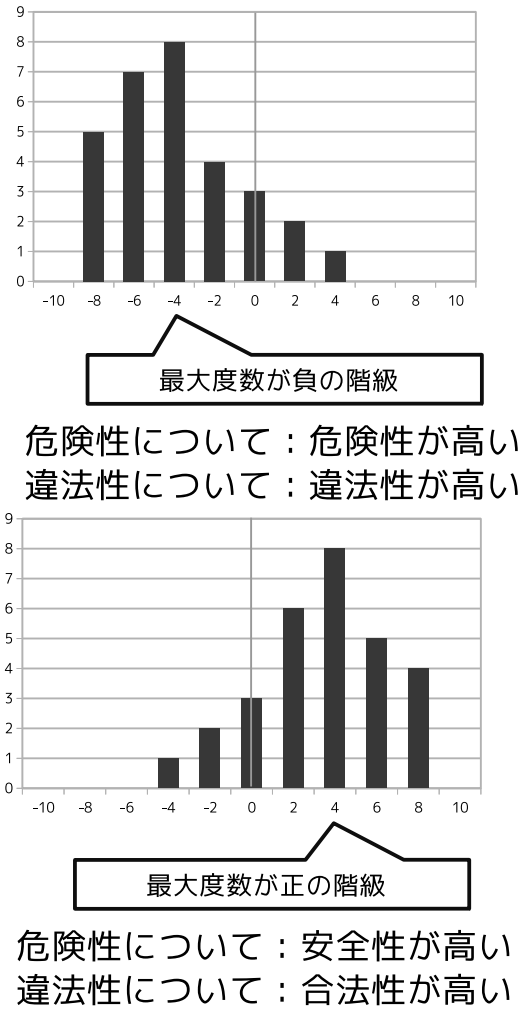


図 3.4: 条件 3 の例

条件 1，2 により，安定して出現し，極性がはっきりしている単語を選び出すことができる．条件 3 により，選択した単語の極性を決定することができる．

第4章　実装

本章では，手法の実装について説明する．

4.1　システム環境

本システムは次の環境下で実装する．OS には VineLinux5.2 を，プログラミング言語には Ruby を用い，ツールは MorphAnalyzer（形態素解析），SVM を実装している TinySVM[3] を使用する．

4.2　システムの概要

4.2.1　記事ファイルのリストの作成

研究室には，ブログ記事ファイルが用意してある．特定のディレクトリから，それぞれのブログ記事ファイルまでのパスをまとめたリストを作成する．作成したリストの例を表 4.1 に示す．記事のファイル名は 1.mar とする．

作成したリストに通し番号を付与する．

4.2.2　「domain」フォルダの作成

話題文書集合の抽出と単語リスト作成を行う．

domain_get.rb

話題文書集合を作成するプログラム．実行すると，domain1，domain2，…，domain30 のようにファイルが生成される（分割数 30 とする）．各ファイルには話題を含む文が抽出されている．そして，作成されたそれぞれのファイルを形態素解析にかける．

表 4.1: 記事ファイルのリスト

./IBlog2008/1/20080731/1.mar
./IBlog2008/1/20080801/1.mar
./IBlog2008/1/20080802/1.mar
./IBlog2008/1/20080803/1.mar
./IBlog2008/1/20080804/1.mar
./IBlog2008/1/20080805/1.mar
・
・
・
./IBlog2013/1/20130429/1.mar
./IBlog2013/1/20130430/1.mar
./IBlog2013/1/20130431/1.mar

count_word.rb

形態素解析したdomain1〜domain30から一般名詞と用言性名詞を抽出し，各単語の話題文書集合における出現回数をカウントする．実行すると，domain1.wrd，domain2.wrd，･･･，domain30.wrdのようにファイルが生成される．各ファイルには抽出した一般名詞と用言性名詞の出現回数が出力されている．

count_sentence.rb

domain1〜domain30における文の数をカウントする．実行すると，domain_sentenceというファイルが生成され，各ファイルにおける文の数が出力されている．

4.2.3 「danger」フォルダの作成

危険文書集合の作成を行う．

danger_get.rb

「domain」で生成したdomain1〜domain30のファイルから，危険性を含む文を抽出する．実行すると，danger1，danger2，･･･，danger30のようにファイルが生成される．これらの各ファイルを形態素解析にかける．

id_give.rb

danger1〜danger30における各文にidを付与する．実行すると，danger1.id，danger2.id，…，danger30.idのようにファイルが生成される．

test_make.rb

形態素解析したdanger1〜danger30における各文を素性の形に変換する．実行すると，danger1_test，danger2_test，…，danger30_testのようにファイルが生成される．

SVMの実行

SVMの2値分類を行うプログラムを使用する．この際，テストデータをdanger1_test〜danger30_testとする．実行すると，01_results.datというファイルが生成され，idと，各文に対する正解値と推定値が出力されている．

result_svm.rb

01_results.datの推定結果と，danger1.id〜danger30.idを照らし合わせ，危険性があると推定した文のみを抽出する．実行すると，danger1.svm，danger2.svm，…，danger30.svmのようにファイルが生成される．これらのファイルを形態素解析にかける．

count_sentence.rb

danger1.svm〜danger30.svmにおける文の数をカウントする．実行すると，danger_sentenceというファイルが生成され，各ファイルにおける文の数が出力されている．

その他の極性の文書集合の作成

「safty」,「illegal」,「lawfull」というフォルダをそれぞれ作成する．これらのフォルダ内での動作は「danger」と同様であり，作成されるファイル名はsafty1やsafty1.idのようにフォルダ名に依存する．

4.2.4 「dan_so_score」フォルダの作成

危険性についての*SO-score*を算出する．

count_word.rb

「domain」フォルダで生成した domain1.wrd〜domain30.wrd を参照し，各単語の危険/安全文書集合における出現回数を，形態素解析した danger1.svm〜danger30.svm および safty1.svm〜safty30.svm からカウントする．実行すると，word_data1，word_data2，…，word_data30 のようにファイルが生成され，各単語の話題文書集合および危険/安全文書集合における出現回数が出力されている．

count_sentence.rb

「domain」フォルダの domain_sentence,「danger」フォルダの danger_sentence,「safty」フォルダの safty_sentence のそれぞれのファイルから各文書集合の文の数をまとめる．実行すると，data1，data2，…，data30 のようにファイルが生成され，それぞれの話題文書集合および危険/安全文書集合の文の数が出力される．

so_score.rb

word_data1〜word_data30 および data1〜data30 から *SO-score* を算出する．実行すると，so_score1，so_score2，・・・，so_score30 のようにファイルが生成される．各単語の *SO-score* の値が出力される．

4.2.5 「dan_total」フォルダの作成

多数算出した *SO-score* の統計をとる．

score_total.rb

「so_score」フォルダの so_score1〜so_score30 から統計をとり，各単語についてヒストグラムのデータを得る．そして 3.9 節で示した条件 1 を満たす単語を抽出する．実行すると，score_total というファイルが生成され，抽出した単語，度数が最大である階級値が出力される．

select_word.rb

score_total における単語のうち，3.9 節で示した条件 2 を満たす単語のみを抽出する．実行すると，select_word というファイルが生成され，抽出した単語，度数が最大である階級値が出力される．

4.2.6 「ill_so_score」および「ill_total」フォルダの作成

それぞれ「dan_so_score」および「dan_total」内での動作と同様である．

4.2.7 シェルスクリプト

以上の流れをまとめて処理するために，シェルスクリプトを作成する（「shell」フォルダに作成）．プログラムの実行，形態素解析，SVM の実行などを手動で行うのは時間を要するので，シェルスクリプトを用いる．シェルスクリプトを用いた場合の総実行時間は約 3 時間程度である．

4.3 実行の様子

例として,「片手」の危険性についての *SO-score* は次のように算出される．「domain」フォルダにおける count_word.rb で s(“片手”) を算出する．「dan_so_score」フォルダにおける count_word.rb で s(“片手”, “危険表現”) および s(“片手”, “安全表現”) を算出し，word_data1〜word_data30 にそれらの数が出力される．word_data ファイルの例を図 4.1 に示す．

「dan_so_score」フォルダにおける count_sentence.rb で s(“危険表現”)，s(“安全表現”)，コーパスの総文数 N をまとめ，data1〜data30 にまとめる．data ファイルの例を図 4.2 に示す．

「dan_so_score」フォルダの so_score.rb により，word_data ファイルと data ファイルから *SO-score* が算出され，so_score1〜so_score30 に出力する．so_score ファイルの例を図 4.3 に示す．

ファイル	編集	オプション	バッファ	ツ
片手	55	9	1	
財団	3	0	0	
対物	2	0	0	
空っ風	2	0	0	
管理	36	0	1	
アルバム		8	0	0
安値	116	0	1	
宗教	3	0	0	
鋼鉄	1	0	0	
要領	5	1	0	
局面	1	0	0	
症候	2	0	0	
本会議	1	0	0	
無防備	1	0	0	
半紙	1	0	0	
持ち腐れ		2	0	0
遊歩道	39	2	0	

図 4.1: word_data ファイルの例

ファイル	編集	オプション	バッファ
308	242	45611	

図 4.2: data ファイルの例

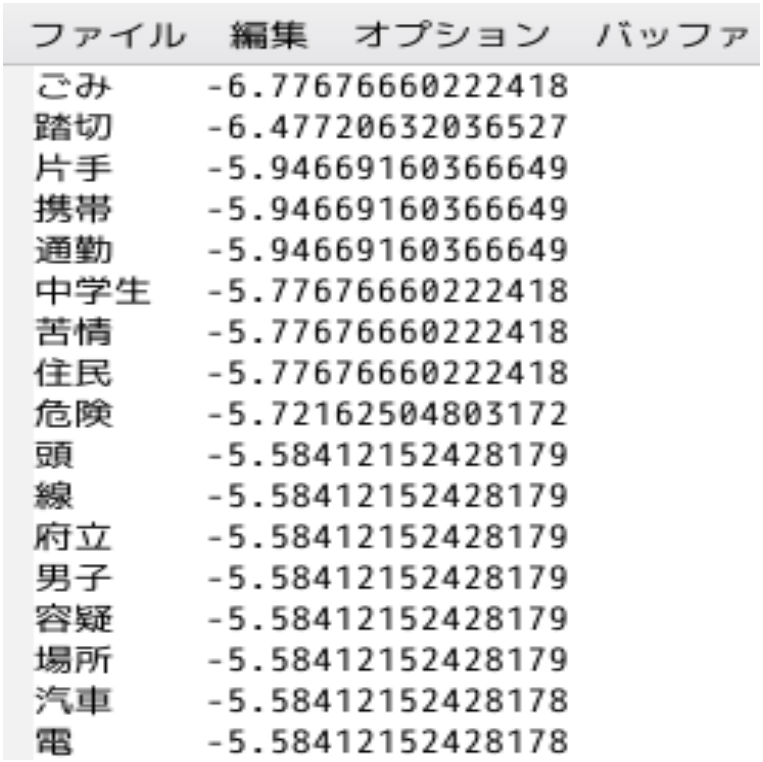


図 4.3: so_score ファイルの例

第5章　実験

まず，提案手法を用いる方法で危険性/違法性のある単語を自動抽出する．さらに，比較手法として，フィルタを用いない場合，ヒストグラムを用いない場合，両方用いない場合について危険性/違法性のある単語を自動抽出する．これらの抽出した単語を精度と再現率により評価を行い，提案手法と比較手法との差を検出する．これにより，本研究で用いたフィルタとヒストグラムの性能評価を行う．

5.1　実験条件

5.1.1　リソース

コーパスは 2008 年 7 月 31 日から 2013 年 4 月 31 日までのブログ記事を使用する．

5.1.2　話題の選択

本研究では話題を,「自転車」,「自動車」,「病院」とする．これらの話題について，危険性および違法性のある単語を抽出する．

5.1.3　各評価極性を表す単語

3.5 節の各評価極性ごとの文の抽出において，各評価極性を表す単語が必要である．本研究で使用した単語を以下に示す．

危険性　「危険」「危＋⟨ ひらがな一文字 ⟩」

安全性　「安全」

違法性　「違法」「異常」「不当」「違反」

合法性　「合法」「正常」「正当」「当然」

5.1.4 各値の設定

コーパスの分割数 n を，$n = 30$ とする．ヒストグラムの階級の幅 w を，$w = 2$ とする．

5.2 評価方法

本研究では精度と再現率による人手評価を行う．

5.2.1 精度

各単語について，自動で抽出した危険性/違法性のある単語のうち，人の判断で危険性/違法性があるとされた割合で評価する．被験者 5 名に，危険/安全及び違法/合法の両単語を示し，判定してもらう．算出式を以下に示す．

精度 =（危険性/違法性があると判定した数の和）/（単語の評価数 × 被験者の数）

5.2.2 再現率

一般に危険性/違法性のありそうとされる単語が抽出できた割合で評価する．被験者 5 名に，危険性/違法性のありそうな単語を 10 個ずつ連想してもらう．その単語のうち，いくつの単語が自動で抽出できているのかを評価する．算出式を以下に示す．

再現率 =（連想してもらった単語のうち抽出できた数の和）/（10 × 被験者の数）

5.3 実験結果

まず，提案手法を用いて危険性のある単語，および，違法性のある単語の抽出を行う．次に，比較手法を用いて危険性のある単語，および，違法性のある単語の抽出を行う．そして，それらの評価を行う．

それぞれの場合において抽出した単語を示すが，ヒストグラムを使用しない場合と両方使用しない場合については，単語の数が膨大であるため割愛する．

5.3.1 提案手法を用いた場合の単語の抽出

危険性

各話題について自動抽出した危険および安全な単語を以下に示す．

「自転車」に関する危険な単語

片手　メール　お婆さん　結構　不注意　当然　道　携帯　目　傘　スピード
ブレーキ　一　状態　練習　運転　灯火　非常

「自転車」に関する安全な単語

子ども　日　ヘルメット　ルール　交通　則　利用者　高齢者　対策　整備
確認　認証　教室　推進　組立　教育　基準　指導　環境　マーク　協会

「自動車」に関する危険な単語

致死　過失致死　タクシー　バイク　違反　気　自転車　運転　人

「自動車」に関する安全な単語

システム　衝突　開発　技術　協会　乗用車　制度　リコール　生産　製品
性能　対策　問題　確保　センター　バス　仮称　一貫　コンテナ　原因
関係　世界　向上　法律　海陸　環境　構造　基準　品質　運送

「病院」に関する危険な単語

状態　生命　医者　後　母　命　救急　診察

「病院」に関する安全な単語

看護　対策　医療　地域　管理　機関　安心　委員会

違法性

各話題について抽出した違法および合法な単語を以下に示す.

「自転車」に関する違法な単語

切符　罰金　傘　信号　マナー　輪　公道　交通　事故　道路　ルール　ブレーキ　自動車

「自転車」に関する合法な単語

携帯　片手　バイク　走行　事　不注意　お婆さん　メール

「自動車」に関する違法な単語

容疑　過失　傷害　ひき逃げ　逮捕　酒気　乗用車　男性　事件　現行犯　タクシー　義務　交通　判決　法令　危険　県警　罪　同県　運転手　過失致死　運転　道路　事故

「自動車」に関する合法な単語

人　販売　事　保険　メーカー　影響　関連　生産

「病院」に関する違法な単語

レントゲン　交通事故　骨　無し　異常　検診　健康診断　業務　脳　目　容疑
予防　事態　相談　気　体調　為　整体　行き　症状　検査　CT　事件　原因
内科　診断

「病院」に関する合法な単語

血圧　医療　場所　仕事　医者　救急　もの　施設　状況　患者　声　診療
生活　関係　人　インフルエンザ

5.3.2　フィルタを使用しない場合の単語の抽出

危険性

各話題についてフィルタを使用せずに抽出した危険および安全な単語を以下に示す.

「自転車」に関する危険な単語

片手　不注意　当然　結構　メール　行為　お婆さん　トラック　灯火　運転
人　目　雨　非常　ブレーキ　所　家　電話　車　雪　傘　状態　携帯　一

「自転車」に関する安全な単語

日　警察　交通　音　高齢者　幼児　環境　前　ルール　対策　全国　交通事故
マナー　ヘルメット　コース　利用　　整備　確認　推進　快適　防止　運命
教育　基準　開運　トラ　指導　ストラップ　対応　則　警視庁　組立　教室
認証　安心　協会　マーク

「自動車」に関する危険な単語

致死　過失致死　タクシー　バイク　免許　違反　気　自転車　前　運転
人　脇道　少年　お話　追記　個所　ガード　犯行　入り口　セルフ　静
止　列　時期

「自動車」に関する安全な単語

システム　環境　衝突　管理　対策　規制　問題　米　防止　技術　バス
確保　シートベルト　機能　協会　仮称　テスト　検査　一貫　コンテナ
原因　実験　制度　日　生産　事業　リコール　電子　海陸　法律　調査
乗用車　製品　向上　構造　性能　センター　品質　運送　開発　基準

「病院」に関する危険な単語

非常　状態　母　精神　先生　命　入院　そう

「病院」に関する安全な単語

地域　医療　機関　対策　看護　施設　管理　確保　委員会　安心

違法性

各話題についてフィルタを使用せずに抽出した違法および合法な単語を以下に示す.

「自転車」に関する違法な単語

行為　マナー　駐車　罰金　容疑　切符　輪　公道　傘　スピード　交通
歩行者　歩道　自動車　事故　ルール　駐輪　道路

「自転車」に関する合法な単語

バイク　通勤　お婆さん　メール　家　ライト　片手　不注意

「自動車」に関する違法な単語

過失　ひき逃げ　容疑　同県　県警　酒気　乗用車　男性　事件　タクシー
傷害　容疑者　義務　法令　駐車　交通　判決　危険　運転者　罪　行為
現行犯　処分　過失致死　運転　道路　事故

「自動車」に関する合法な単語

影響　もの　電気自動車　日産　産業　軽自動車　事　保険　状況　メーカー
関連　生産

「病院」に関する違法な単語

レントゲン　容疑　骨　交通事故　無し　不明　結果　打撲　首　整体　業務
健康診断　ストレス　肺　肩　頭痛　事態　相談　気　目　勤務　脳　息子
検査　行き　症状　CT　原因　体　事件　内科　診断

「病院」に関する合法な単語

血圧　家　仕事　医者　場所　必要　感染　環境　医療　診療所　患者　救急
状況　診療　もの　写真　人　施設

5.3.3 抽出した単語の評価

抽出した単語について精度と再現率で評価を行う．提案手法の評価と比較手法の評価を同時に示す．ヒストグラムを使用しない場合および両方使用しない場合は単語の抽出数が膨大であるため，精度を算出する際に 20 件をサンプリングし，危険率 5%の誤差を示している.「自転車」についての評価を表 5.1 に,「自動車」についての評価を表 5.2 に,「病院」についての評価を表 5.3 に示す.

表 5.1: 「自転車」についての評価

	危険性の評価			違法性の評価		
手法	精度	再現率	抽出数	精度	再現率	抽出数
提案手法	0.59	0.22	18	0.42	0.14	13
フィルタ無し	0.50	0.22	24	0.33	0.14	18
ヒストグラム無し	0.20±0.18	0.52	3,228	0.10±0.13	0.58	2,551
両方無し	0.10±0.13	0.50	3,266	0.10±0.13	0.58	2,677

表 5.2: 「自動車」についての評価

	危険性の評価			違法性の評価		
手法	精度	再現率	抽出数	精度	再現率	抽出数
提案手法	0.67	0.08	9	0.57	0.18	23
フィルタ無し	0.50	0.22	24	0.48	0.22	27
ヒストグラム無し	0.10±0.13	0.60	3,016	0.10±0.13	0.68	2,889
両方無し	0.10±0.13	0.66	3,253	0.20±0.18	0.68	3,184

表 5.3: 「病院」についての評価

	危険性の評価			違法性の評価		
手法	精度	再現率	抽出数	精度	再現率	抽出数
提案手法	0.28	0.06	8	0.12	0.06	25
フィルタ無し	0.25	0.02	8	0.09	0.04	32
ヒストグラム無し	0.10±0.13	0.54	3,927	0.00±0.00	0.40	4,478
両方無し	0.10±0.13	0.54	4,095	0.00±0.00	0.42	4,679

第6章　考察

まず，提案手法をベースに考察する．次に，結果に対しての誤り分析を行い，追加実験を行う．

6.1　実験結果の考察

6.1.1　フィルタの性能

表 5.1 の「自転車」について考察すると，危険性の場合も違法性の場合も単語の抽出数が増えている．再現率には変化が無く，精度が下がっている．

表 5.2 の「自動車」については，危険性と違法性のどちらにおいても，単語の抽出数が増え，精度が下がり，再現率が上がっている．

表 5.3 の「病院」については，危険性の場合は単語の抽出数に変化が無いが，精度と再現率が下がっている．違法性の場合は単語の抽出数が増え，精度と再現率が下がっている．

これらのことから，フィルタを用いることで，全体的に単語の抽出数を絞り込むことができ，精度が上げることができた．以上によりフィルタを用いることで効果はあったといえる．

6.1.2　ヒストグラムの性能

表 5.1，5.2，5.3 から，ヒストグラムを用いない場合，どの場合においても単語の抽出数が膨大であり，全体的に精度が下がっていることがわかる．再現率が上がっているが，単語の数が増えることで網羅性が上がっているからだと考えられる．これらのことから，ヒストグラムを用いることで単語の抽出数を大幅に絞り込み，精度を上げることができた．

6.2　誤り分析

6.2.1　分析 1

提案手法の再現率が低い原因を分析した.

失敗例

各話題について抽出できなかった単語の例を以下に示す.

抽出できなかった「自転車」に関する危険な単語の例

事故　転倒　故障　パンク　強風　夜道　衝突

抽出できなかった「自転車」に関する違法な単語の例

無灯火　盗難　飲酒　二人乗り

抽出できなかった「自動車」に関する危険な単語の例

居眠り　ひき逃げ　事故　故障　雪　酒　雨

抽出できなかった「自動車」に関する違法な単語の例

駐車　信号　スピード　定員　シートベルト　無灯火

抽出できなかった「病院」に関する危険な単語の例

ミス　不衛生　感染　停電　事故

抽出できなかった「病院」に関する違法な単語の例

ミス　保険　無断　点検　過労

分析

抽出できなかった「自転車」に関する危険な単語として「事故」がある．この「事故」を危険性のある表現として用いて,「自転車」に関する危険な単語を再抽出した．その結果を以下に示す.

「事故」を用いた場合の「自転車」に関する危険な単語

同士　増加　ニュース　怪我　男性　接触　責任　相手　死亡　保険 賠償
人身　衝突事故　加害者　うち　都内　転倒　多発　被害者　過失　高校生
新聞　記事　状態　乗用車　ケース　損害　国道　電車　交通事故　現場
目　交差点　自動車　バス　原因　防止　横断歩道　携帯　女性

「事故」を用いた場合の「自転車」に関する危険な単語

環境　交通　学校　場所　ヘルメット　日　方法　道　教育　教室　子ども
対応　対策　利用　推進　幼児　指導　確認　基準　整備　確保　則
マーク　安心　認証　協会　組立

提案手法で抽出した単語と比較すると，危険な単語の種類が大きく違うことがわかる．「自転車」と「危険」及び「危＋⟨ひらがな一文字⟩」が共起する単語と,「自転車」と「事故」が共起する単語が違うため,「自転車」に関する危険な単語として「事故」が抽出されなかったことがわかる．

提案手法では抽出できなかったが,「事故」を用いることで抽出できている危険な単語も多い．このように，各評価極性を表す表現として，話題によって特定の単語を用いることで，性能が向上することが予想される．

また，失敗例に「無灯火」という単語があるが，これは「無」と「灯火」の 2 単語として抽出されるためであると考えられる．実際に提案手法において「自転車」に関する危険な単語として「灯火」が抽出されている．このような表現を 1 単語として抽出するために，名詞が 2 単語以上連続した場合は一つの表現としてまとめる方法が考えられる．

6.2.2　分析 2

提案手法で抽出した「病院」に関する違法な単語に,「骨」,「脳」,「目」などの身体的な単語が存在する．これは，評価極性を表す単語として用いた「異常」と「正常」が問題だと予想される．よって，違法性を表す表現として「異常」を，合法性を表す表現として「正常」を用いずに違法性のある単語を抽出した．その結果を以下に示す．

「異常」「正常」を用いずに抽出した「病院」に関する違法な単語

医師　総合　精神　事件

「異常」「正常」を用いずに抽出した「病院」に関する合法な単語

大学　中　検査　動物　人　施設　先生

身体的な単語を抽出しないことに関しては狙い通りだったが，単語の抽出数が大幅に減ってしまい，よくない結果となった．本研究の提案手法における「病院」という話題は難しかったといえる．

6.3 追加実験

6.3.1 実験 1

提案手法ではヒストグラムの階級の幅を 2 としたが，幅を 1 に変えて実験を行う．捨てる単語は −0.5 以上 0.5 未満の階級の度数が最大である単語とする．話題「自転車」とし，危険性について実験を行う．新しく抽出された単語には下線をひく．その結果を以下に示す．

階級の幅を変えた「自転車」に関する危険な単語

携帯　灯火　不注意　結構　片手　メール　お婆さん　目　非常　当然
ハンドル　練習　傘　状態　ブレーキ　運転　一　通勤　通行　気　電話
歩道　道　走行　自転　車道　人　後ろ　道路　そう　スピード　車

階級の幅を変えた「自転車」に関する安全な単語

ヘルメット　ルール　交通　学校　警察　利用者　子ども　確認　対策　整備
教育　高齢者　環境　指導　推進　教室　則　組立　基準　認証　協会
利用　安心　マーク

単語の抽出数が少し増えたが,「ハンドル」,「気」,「後ろ」,「そう」など危険性とは関係のない単語が増えてしまっている．この結果から，提案手法で設定した階級の幅は，適切であったといえる．

6.3.2 実験 2

3.9 節で設けたヒストグラムの条件のうち，条件 1 についての検証を行った．

提案手法では，n 分割したコーパスに対して，n 個の *SO-score* を算出している単語を対象とした．そこで，追加実験として，条件の値を $n-1$，$n-2$ …，と緩くしていった場合の単語の抽出数を調べる．その他の条件は変更しない．$n=30$ とした場合の結果を図 6.1 に示す．縦軸を単語の抽出数，横軸を度数の総和の条件値とする．

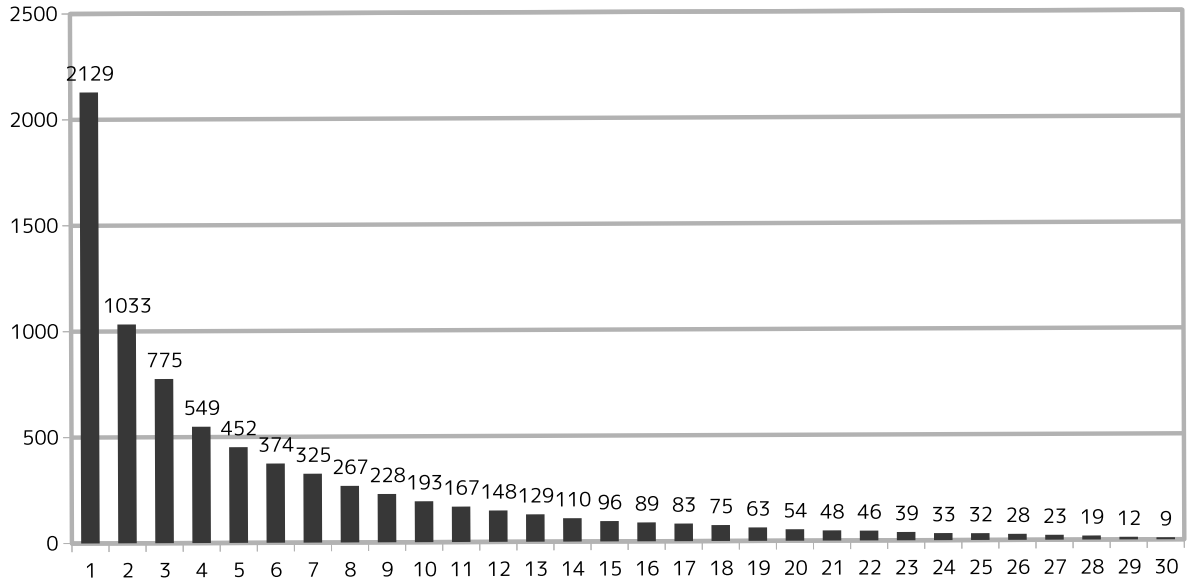


図 6.1: 条件 1 の変更による単語の抽出数の変化

条件を緩くすると，対数的に単語の抽出数が増えた．条件が 30 の時，9 個の単語しか抽出できていなかったので，条件 1 の値を調節することで，抽出された単語の評価も変わると予想できる．

6.3.3 実験3

抽出した各単語について分散値を算出し，*SO-score* の値と分散値の関係を調べた．話題は「自転車」とし，危険性について抽出した単語を対象とした．その結果例を表6.1に示す.

表 6.1: 各単語の分散値

単語	最大度数の階級	分散値
携帯	-7以上-5未満	4.02873215947458
不注意	-7以上-5未満	1.57605792000423
お婆さん	-7以上-5未満	0.829433999636594
結構	-7以上-5未満	5.17675858915692
メール	-7以上-5未満	4.46497032930076
片手	-7以上-5未満	3.31970570186753
当然	-7以上-5未満	3.05842616998923
目	-5以上-3未満	4.13728934906708
危険	-5以上-3未満	1.70440236508101
灯火	-3以上-1未満	4.80729606209156
ー	-3以上-1未満	5.90068939096196
スピード	-3以上-1未満	3.38926698084143
傘	-3以上-1未満	3.87467340090654
	・	
	・	
	・	
協会	7以上9未満	1.31428999091063
マーク	7以上9未満	2.07368269280907
安全運転	7以上9未満	4.11033672493546
利用	7以上9未満	3.35675192632065
認証	7以上9未満	0.138224926287827
安心	7以上9未満	1.71385772791478

分散値が4以内の単語を選び出すと,「不注意」,「お婆さん」,「片手」,「危険」,「スピード」,「傘」となる.「目」や「一」など不要な単語を捨てることができているので，分散値に閾値を与えることで精度を上げることができる可能性がある.

また，6.3.2節で示す方法で単語の抽出数を増やし，分散値によって有用な単語を選び出すことで再現率を上げることができるかもしれない.

6.3.4 実験4

「怖い」,「恐ろしい」,「安心」などの感情表現語はいずれの話題にも適応できる可能性がある.「自転車」と「病院」について好評表現を「安心」とし，不評表現を「怖い」,「恐ろしい」として再実験を行う．抽出した単語を以下に示す.

「自転車」に関するな不評な単語

後ろ　家　スピード　車　信号　通勤　一　人　そう　気

「自転車」に関する好評な単語

走行　ロード　交通　危険　推進　ルール　確認　子ども　教室　基準　対策
則　組立　環境　利用　保険　整備　協会　マーク　認証

「病院」に関する不評な単語

話　注射　いっぱい　感染　一　ま　嫌い　検診　病気　行

「病院」に関する好評な単語

熱　総合　妊婦　一緒　医師　医療　行き　患者　情報　小児科　看護　血
言葉　経験　担当　職員　担当　センター　ひと　機関　出産　かかりつけ
設備　診療　関係　内科　介護　体制　管理　対応　連絡　市民　女性　ペット
保険　健康　スタッフ　全国　対策　環境　求人　施設　地域

精度や再現率が上がるとはいえないが，抽出される単語の種類が変わることが分かった.「病院」に関しては「感染」や「病気」などを得ることができたので，悪くない結果といえる.

第7章　おわりに

本研究では危険性や違法性を表す表現の自動収集を目的とした．Tuney[1] が提案した *SO-score* の算出方法をコーパス全体から行うと，偶然良い値や悪い値が算出される可能性が高く，単語の候補が多すぎるという問題点があった．そこで，本研究ではコーパスを分割し，*SO-score* を多数算出することで，統計的な安定化を図る方法を提案した．提案手法を用いた場合，単語の抽出数が絞られ，再現率が下がるが精度を上げることができた．

再現率が低い原因は，各評価極性を表す表現として用いた単語が適切ではなかったことが挙げられる．「自転車」における「事故」など，極性を表す表現を各話題における特定の単語を用いることで，一般的に危険性あるいは違法性があるとされる単語を抽出できると予想される．

本研究では危険性・違法性に着目し，極性の詳細化を図った．他の極性についても，本研究の提案手法を適応させることができるのか試す必要がある．

謝辞

本研究を進めるに当たり，終始に渡り研究の進め方や本論文の書き方など，細部に渡る御指導を頂きました徳久雅人講師に心から御礼申し上げます.

また，本研究を進めるに当たり，種々の御助言を頂きました村田真樹教授，および，村上仁一准教授に心から御礼申し上げます.

その他様々な場面で御助力を頂いた計算機工学 C 講座の学生の皆様に感謝の意を表します.

参考文献

[1] Turney, P. D: “Thumbs Up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews.” In Proc. of ACL2002, pp.417-424, 2002.

[2] 高村大也，乾孝司，奥村学：“スピンモデルによる単語の感情極性抽出”，情報処理学会論文誌ジャーナル，Vol.47 No.2 pp. 627–637， 2006

[3] TinySVM : http://chasen.org/taku/software/TinySVM/