ソーシャルネットワーキングサービス上の不要語抽出

根津　裕太† 　三浦　孝夫†

† 法政大学 理工学部創生科学科　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †yuta.nezu.6f@stu.hosei.ac.jp, ††miurat@hosei.ac.jp

あらまし　テキスト分類や情報検索では，効率向上を目標として，予め不要語除去を行う．既存の除去手法は，新聞記事などの形式が整った文書では有効な手段ではあるが，ソーシャルネットワーキングサービス (SNS) 上のテキストは表現の多様さから効果的とはいえない本研究では，不要語を統計量によって定義し，不要語条件を満たすための指標を提案する．本稿では，特徴量を算出するためのカテゴリ推定法として，SNS の時系列データを利用したカルバック・ライブラー情報量による分類法を利用する．また，SNS の未知語に対応するための事前処理を行い形態素解析の精度向上を図る．

キーワード　自然言語処理，不要語，ソーシャルネットワーキングサービス

1. 前　書　き

近年，インターネットの普及により，個人でも簡単に情報を発信することが可能になった．特に，ソーシャルネットワーキングサービス (Social Network Service, SNS) では，世界中の人々がリアルタイムで情報を発信しているため，そのデータ量は莫大な量である．そのため，利用者が欲しい情報を入手するための手段として，情報検索やテキスト分類の価値が高まっている．しかし，多様で莫大なデータ量を分析対象とするため，分析の処理時間や精度が問題となっている．

情報検索やテキスト分類において不要語の消去が欠かせない．不要語とは，検索・分類・特徴抽出に有効な働きをしない語のことであり，Google や SlothLib が一般公開している不要語リストが代表例である．

不要語抽出の研究は数多く提案されている [1] [2] が，SNS に限定した議論はあまりない [5]．本研究では，不要語を統計量によって定義し，不要語条件を満たすための指標を提案する．また，特徴量を算出するためのカテゴリ推定法として，SNS の時系列データを利用したカルバック・ライブラー情報量による分類法を利用する．さらに，SNS の未知語に対応するための事前処理を行い，形態素解析の精度向上を図る．

2 章では不要語の定義と性質について述べ，3 章では SNS の表現の特徴について述べる．4 章，5 章では提案手法に用いる用語の説明と不要語との対応付け，6 章では提案手法の概要を述べる．7 章で実験を行い，8 章では得られた結果について考察する．9 章で結論を述べる．

2. 不要語の定義と性質

不要語とは検索・分類・特徴抽出に有効な働きをしない語，カテゴリを一意に決定するために有用でない語と定義される [1]．例えば，助詞（「は」，「の」など）や助動詞（「です」，「れる」など）などといった「機能語」や，括弧や演算子，数字などといった「記号」などが不要語の代表である．これらの単語は共通して，

(1) 文書内に頻繁に出現する
(2) 文書のカテゴリに依らず多くの文書に出現する

といった特徴を有している．(2) の性質により，情報検索やテキスト分類をする際に，精度を低下させる原因となる．

既存の不要語抽出の方法 [1] として，これまで不要語リストの利用がある．一般的には Google や SlothLib が一般公開している不要語リストを用いることが多い [3] [4]．また，Feng らが提案した単語の出現頻度の平均値と分散を指標とした方法 [2] などが挙げられる．

3. SNS の表現の特徴

SNS ではインターネットを介して誰でも簡単にコミュニケーションを楽しむことができる．Twitter や Facebook などが普及しており，そこで蓄積されたテキストデータの量は莫大なものとなっている．

SNS の文書には，新聞記事などの形式の整った文書とは異なる以下の特徴を有している．

(1) 1 文書の語数が少なく，文章も短い
(2) SNS 特有の表現が頻出する
(3) 更新頻度が多い
(4) 時間経過によるトレンド（話題）が激しく変化する

不要語抽出の観点からは，文書が短い (1) ということは，特徴語の数が少ない．つまり，単語の除去で，文書の持つ情報がなくなる可能性がある．(2) については，「www」や「ワロタ」など，新聞記事などでは現れない単語や，顔文字やユニコードといった「意味を有した記号列」の出現．さらには，手軽に情報を発信できることによって，機能語の少ない崩れた表現が多い．このことから，従来の不要語リストや抽出法をそのまま適用できない．(3) により，文書集合が常に変動し続けることを意味ため，文書数などを用いた指標の算出（例えば DF）が困難になる．(4) により，不要語が時間によって変化する可能性が高いが，(3) の問題がある．

SNS（特に Twitter）の不要語抽出法には，Saif らが提案した，単語の共起関係と三角恒等式を適用させた感情分析の不要語抽出 [5] があるが，本稿ではテキスト分類を対象とした不要語抽出法を提案する．

4. 統 計 指 標

Yang らはテキスト分類の統計的学習における特徴選択法の比較研究を行っている [6]．特徴選択法とは特徴集合のうち，分類に有効な部分集合のみを選択する手法である．テキスト分類では，単語集合から特徴語を抽出することに相当する．特徴語とは不要語の対義語にあたる．すなわち，特徴選択によって得られた部分集合の余事象こそが不要語であるといえる．

園田らは，日本語文書に対する連語抽出のため，いくつかの統計指標を用いたデータマイニング手法を提案しており，条件付き確率と統計フィルタの併用による効果は大きい [7]．

本研究では，Yang らが比較した統計指標のうち，情報利得 (Information gain, IG) と χ^2 統計量 (χ^2 statistic, CHI)，そして，文書内での単語の出現頻度である単語頻度　(Term Frequency, TF) について論じる．

4.1 単語頻度 (Term Frequency, TF)

TF とは文書あるいは文書集合中での語の出現頻度である．正規化した場合，以下のようになる．

$$tf(i,j) = \frac{f_{i,j}}{N_j} \tag{1}$$

$$tf(w_i) = \frac{f_{w_i}}{N} \tag{2}$$

ここで，$f_{i,j}$ は文書 j における単語 i の出現回数，N_j は文書 j での単語の総出現回数，f_{w_i} は文書集合 D での単語 i の出現回数，N は文書集合 D での単語の総出現回数である．また，式 (1) は文書内での TF，式 (2) は文書集合中での TF を示す．文書内では，多く出現する単語ほどその文書の特徴を示す．一方，多くの文書に出現する単語ほど TF が高くなるため，必ずしも文書を特定する特徴とは言えない．

頻出な単語は不要語の性質を有しているともいえるが，特徴語の性質を有することから，TF 指標のみで不要語と断定するには値しない．一方，文書集合中での出現頻度が高いということは，その単語を除去した際に，文書集合内での総単語数の削減率が高いことを示す．つまり，TF が高い単語を除去するほど，検索や分類の処理効率が向上する．このため，TF は不要語指標の 1 つであるとも考えられる．

4.2 情報利得 (Information gain, IG)

情報利得は，有無によって，カテゴリ予測のために得られる情報量のことであり，以下の式によって定義する．

$$G(w_i) = H(w_i) - H(w_i|c_K) \tag{3}$$

$$H(w_i) = -P(w_i)\log_2 P(w_i) - P(\overline{w_i})\log_2 P(\overline{w_i}) \tag{4}$$

$$H(w_i|c_K) = \sum_{k\in K} -P(w_i|c_k)\log_2 P(w_i|c_k) - P(\overline{w_i}|c_k)\log_2 P(w\overline{w_i}|c_k) \tag{5}$$

ここで，$\overline{w_i}$ は w_i の余事象，c_k はカテゴリを示す．

具体的には，文書集合全体での単語のエントロピーと各カテゴリでの単語のエントロピーの期待値との差をとった値である．例えば，単語「羽生結弦」がカテゴリ「フィギュアスケート」に 100 回出現し，他のカテゴリ「体操」と「バレーボール」には 1 回ずつ出現するとき，各カテゴリの総単語数 10000 に対して「羽生結弦」の情報利得は 0.0049 となる．それに対して,「世界大会」が全てカテゴリに 34 回ずつ出現するとき，単語「世界大会」の情報利得は 0 になる．つまり，カテゴリ毎の出現確率が等しい単語ほど 0 に近づき，偏りがあるほど値は大きくなる．そのため，情報利得が高い単語ほど，いずれかのカテゴリの特徴語である可能性が高い．

このように，情報利得が低いという条件は，2 章の (2) の性質を満たしていることになるため，不要語指標になりうる．

4.3 χ^2 統計量 (χ^2 statistic, CHI)

χ^2 統計量は，ある単語の出現確率が各カテゴリに対して独立であるかどうかを示す尺度であり，以下の式によって定義する．

$$\chi^2(w_i) = \sum_k \frac{(P(w_i|c_k) - E(w_i,c_k))^2}{E(w_i,c_k)} \tag{6}$$

$$E(w_i,c_k) = \frac{f_{w_i} \times N_{c_k}}{N_{c_k}} \tag{7}$$

ここで，N_{c_k} はカテゴリ c_k での単語の総出現回数である．$E(w_i,c_k)$ は単語 w_i がカテゴリに依らず独立であるという仮説下での，各カテゴリでの単語の出現回数である．$E(w_i,c_k)$ と実際の出現回数を比較し，差を足し合わせた値が統計量となる．

Yang らは，情報利得と χ^2 統計量には非常に強い相関があることを示し，例えば「羽生結弦」と「世界大会」の値はそれぞれ 192.18 と 0 となる．このことから，情報利得と同様に不要語指標の 1 つとなる．

5. カルバック・ライブラー情報量と時系列データ

4 章で説明した情報利得と χ^2 統計量は，事前に文書のカテゴリ分類を前提としている．

twitter など多くの SNS にはハッシュタグによるタグ付け機能存在するが，これらはツイートする人が自由にキーワードを与えるために，SNS 上に無数の類似したタグが存在することになる．また，1 ツイートに対し複数のハッシュタグが存在するため，ハッシュタグをカテゴリとして利用することは合理的ではない．本研究ではカルバック・ライブラー情報量 (Kullback-Leibler divergence, KLD) と時系列データを利用したカテゴリ分類を論じる．ここでは統計量の算出を目標としているため，カテゴリ分類を目的とするものでない．

5.1 カルバック・ライブラー情報量

KLD とは 2 つの確率分布の差異を測る尺度である．2 つの文書集合中の単語の出現確率分布をそれぞれ $P(w)$，$Q(w)$ としたとき，2 文書間の KLD は，

$$KLD(P(w)||Q(w)) = \sum_{w_i\in W} P(w_i)\log_2 \frac{P(w_i)}{Q(w_i)} \tag{8}$$

となる．$KLD(P||Q) = 0$ となるのは $P(w) = Q(w)$ のときに

限り，また非負値であるため，2 つの文書集合間での単語の分布に差があるほど，その値は大きくなる．

5.2 時系列データによるカテゴリ分類

新聞記事やニュースなどの話題は時間経過によって変化する．特に SNS では，リアルタイムで新しい文書が投稿されるため，話題が激しく変化し，それに連なって出現する単語の分布も変化する．話題の転換期の前後では単語分布の差が大きくなり，KLD も高くなる．本稿では一定時間毎に単語の出現頻度をとり，前期間での単語分布と現在の単語分布から KLD を算出する．値が閾値以下の場合，2 つの文書集合は同じカテゴリ（話題）であるとし，閾値を超えた場合，異なるカテゴリとみなす．

6. 提 案 手 法

本研究では，SNS の不要語を自動抽出するため，2 つの統計量を組み合わせた不要語指標を提案する．また，統計量算出の際に必要なカテゴリ分類法として，SNS の時系列データに従ったカルバック・ライブラー情報量を利用する方式を提案する．

低頻度の単語は，カテゴリ推定に有用ではなく，実験の処理効率にも悪影響を与えるため，実験対象外の不要語として，予め消去する．

あるカテゴリにおいて出現回数 0 の単語は，統計量の算出の際 $\log_2 0$ を生成する．これをゼロ頻度問題という．この問題を避けるため，カテゴリ c_k での単語 w_i の条件付き出現確率 $P(w_i|c_k)$ を MAP 推定する．

$$P(w_i|c_k) = \frac{f_{w_i,c_k} + 1}{N + W \times K} \tag{9}$$

ここで，f_{w_i,c_k} はカテゴリ c_k における単語 w_i の出現回数，W は文書集合での単語の種類数，K はカテゴリの数である．

ここで，不要語指標として，以下のものを提案する．

$$TFIG(w_i) = \log_2 \frac{tf(w_i)}{G(w_i)} \tag{10}$$

$$TFCHI(w_i) = \log_2 \frac{tf(w_i)}{\chi^2(w_i)} \tag{11}$$

指標の分子は，4.1 節で論じた TF であり，高いほど検索や分類の処理効率の向上が期待できる．指標の分母は，4.2 節と 4.3 節で論じた統計指標であり，低いほど分類精度の向上が期待できる．提案する指標は TF と統計指標の逆数の積であり，処理効率と分類精度を両立するものである．これらの値が高い単語ほど不要である。

7. 実　　験

本章では，提案手法の有用性を実験により示す．始めに結果の意義と手法の有効性を検討する．

7.1 実 験 環 境

本実験ではツイート文を対象として 2 つのコーパスを構成する．このため，統計解析ツール R のパッケージ (twitteR) を用い，キーワード検索によって指定期間内の指定キーワードを含む全ツイート文を収集する（ただし RT 文は除外する）．選定したキーワードは，@jptrend（1 時間毎にその時間のトレンドをつぶやく非公式アカウント）のツイートを参考にし，期間内に多くトレンド入りしたキーワードを採用する．詳細な情報を表 1, 表 2 に示す．

表 1　ツイートの取得期間とツイート数

コーパス	取得期間	ツイート数
コーパス 1	2018/7/30/13:00 ～ 2018/7/31/12:59	65545
コーパス 2	2018/12/14/4:00 ～ 2018/12/15/3:59	134182

表 2　取得キーワードとツイート数

コーパス 1		コーパス 2	
取得キーワード	ツイート数	取得キーワード	ツイート数
シャイニングマンデー	31034	ホームアローン	67930
あなた流のクレープ	699	#南極の日	1946
#月曜から夜ふかし	2802	LINE Pay	5525
#ijuin	2332	価格改定	1044
#musica_od	924	大太刀	20132
#ヤマノススメ	1718	笹寿司	892
アークザラッド	6656	流れ星	32878
#梅干しの日	1080	きんようび	3835
#tama954	897		
大阪桐蔭	17776		

収集したコーパスからテキスト情報と時間情報を抽出する．1 ツイートを 1 文書とし，文書に対して MeCab により形態素解析を行う．システム辞書には「mecab-ipadic-NEologd」を使用し，さらに付近 1ヶ月以内にトレンド入りしたキーワードを辞書に追加する．また，形態素解析を SNS に対応させるために以下の処理を行う．

(1) 「URL」「ユーザー名」の削除する
(2) ローマ字と数字を半角，カタカナを全角に統一する
(3) 動詞・形容詞の基本形化を行う
(4) 「サ変接続名刺+する」の場合の「する」を消去する
(5) Unicode を 1 単語で扱う
(6) 顔文字（「顔文字に使われる可能性の高い記号・文字」が 4 文字以上続いたもの）を 1 単語で扱う
(7) (5), (6) の条件を満たさない記号を消去する
(8) 3 文字以上連続する母音を統一する
　例 :「あああ」「あああああ」「あああああ」は全て「あああ」に統一する
(9) w（笑い表現）を統一する
　例 :「w」「ww」「www」は全て「w*」に統一する
(10) 連続する「っ」を処理する
　例 :「やっっば」を「やっば」に統一する
(11) 形容詞の話し言葉を基本形化する
　処理パターン 1 :「い」の前の言葉の母音が変形している場合
　例 :「すげぇ」「すげー」を「すごい」に統一する
　処理パターン 2 : 1 文字目の後に「っ」が挿入されている場合

例 :「こわい」を「こっわ」に変形する

(12) 人名を統合する（「姓」＋「名」は「姓名」で 1 単語）

(13) 以上の処理で拾わなかった 1 文字未知語を削除する

これらはすべて状況に依存しており，出現頻度を判断して導入した局所的規則である．

上記の処理によって抽出された単語について，各コーパスから TF，情報利得，χ^2 統計量の値を算出する．カルバック・ライブラー情報量の時間間隔は 30 分，閾値は 0.6 とする．また，コーパス 1 では 4 ツイート以下しか出現しない単語，コーパス 2 では 5 ツイート以下しか出現しない単語をあらかじめ除去する．

算出した統計量より，提案手法である TF/IG（式 (10), 以下 TF/IG），および TF/CHI（式 (11), 以下 TF/CHI）を指標とし，コーパス中の単語をランク付けする．そして，指標の上位 X%の単語を「不要語候補語」とする．また，比較のため TF, 1/IG，1/CHI の指標も同様に行う．

7.2 評 価 方 法

評価の基準としては以下の 2 項目を用いる．

(1) 単語の削減率

(2) 単純ベンズ分類器による正解率

(1) に関しては以下の式によって算出する．

$$R = \frac{N - N_d}{N} \tag{12}$$

ここで，N はコーパス中の全単語数，N_d は不要語候補語の総単語数である．(2) に関しては，多項分布に従った単純ベンズ処理を行う．不要語指標を算出したコーパスを分類の対象とし，コーパスを一様乱数に従って 1:9 に分けたのち，1 割で学習を行い，9 割でテストする．正解データは，ツイート抽出を行った際に用いたキーワードをそのまま正解カテゴリとする．例えば,「ホームアローン」というキーワードで収集したツイートの正解カテゴリは「ホームアローン」となる．各カテゴリ毎の正解率の平均値を評価基準とする．

7.3 実 験 結 果

TF，1/IG，1/CHI，TF/IG，TF/CHI それぞれでの上位 1%から 30%（1%毎区切り）を不要語候補語としたときの結果．元のコーパスの場合での精度，および，Google 及び SlothLib の不要語リストを不要語候補語とする．この結果を表 3, 表 4, 表 5, 表 6, 表 7 に示す．ここでは図 1（不要語リスト）をベースラインとする．

TF では他の指標に比べ，削減率が高いが正解率が低い．1/IG，1/CHI では閾値 X を上げるごとに，正解率の向上が確認できるが，削減率が他の指標に比べ低く，ともにベースラインを下回る．提案手法である TF/IG，TF/CHI では，閾値 X の値によっては，削減率・正解率ともにベースラインを上回る．

あそこ, あたり, あちら, あっち, あと, あなた, あれ, いくつ, いつ, いま, いや, いろいろ, うち, おおまか, おまえ, おれ, がい, かく, かたち, から, がら, くせ, ここ, こっち, こと, ごと, こちら, ごっちゃ, これ, これら, ごろ, さまざま, さらい, さん, しかた, すか, ずつ, すね, すべて, ぜんぶ, そう, そこ, そちら, そっち, そで, それ, それぞれ, それなり, たくさん, たち, たび, ため, だめ, ちゃ, ちゃん, てん, とおり, とき, どこ, ところ, どちら, どっか, どっち, どれ, なか, なかば, なに, など, なん, はじめ, はず, はるか, ひと, ひとつ, ふく, ぶり, べつ, へん, ぺん, ほう, ほか, まし, まとも, まま, みたい, みつ, みなさん, みんな, もと, もの, もん, やつ, よう, よそ, わけ, わたし, ハイ, 上, 中, 下, 字, 年, 月, 日, 時, 分, 秒, 週, 火, 水, 木, 金, 土, 国, 都, 道, 府, 県, 市, 区, 町, 村, 各, 第, 方, 何, 的, 度, 文, 者, 性, 体, 人, 他, 今, 部, 課, 係, 外, 類, 達, 気, 室, 口, 誰, 用, 界, 会, 首, 男, 女, 別, 話, 私, 屋, 店, 家, 場, 等, 見, 際, 観, 段, 略, 例, 系, 論, 形, 間, 地, 員, 線, 点, 書, 品, 力, 法, 感, 作, 元, 手, 数, 彼, 彼女, 子, 内, 楽, 喜, 怒, 哀, 輪, 頃, 化, 境, 俺, 奴, 高, 校, 婦, 伸, 紀, 誌, レ, 行, 列, 事, 士, 台, 集, 様, 所, 歴, 器, 名, 情, 連, 毎, 式, 簿, 回, 匹, 個, 席, 束, 歳, 目, 通, 面, 円, 玉, 枚, 前, 後, 左, 右, 次, 先, 春, 夏, 秋, 冬, 一, 二, 三, 四, 五, 六, 七, 八, 九, 十, 百, 千, 万, 億, 兆, 下記, 上記, 時間, 今回, 前回, 場合, 一つ, 自分, ヶ所, ヵ所, 箇所, ヶ月, ヵ月, カ月, 箇月, 名前, 本当, 確か, 時点, 全部, 関係, 近く, 方法, 我々, 違い, 多く, 扱い, 新た, その後, 半ば, 結局, 様々, 以前, 以後, 以降, 未満, 以上, 以下, 幾つ, 毎日, 自体, 向こう, 手段, 同じ, 感じ, 貴方, 我々, だれ, あり, おり, かた, です, ます, は, が, の, に, を, で, から, まで, も, と, より, え, それで, しかし

図 1　Google と SlothLib が公開している不要語リスト

表 3　元のコーパスおよび Google+SlothLib の不要語リストでの削除率と正解率

	コーパス 1		コーパス 2	
	削減率	正解率	削減率	正解率
元のコーパス	-	78.7799	-	85.0075
Google+SlothLib	26.65522	85.56741	24.23754	88.82252

表 4　コーパス 1 での削減率

上位 X%	TF	1/IG	1/CHI	TF/IG	TF/CHI
1	55.86149833	0.217286655	0.26623976	19.13465971	20.1215584
2	64.53556123	0.417780664	0.485561887	21.79176509	22.41685859
3	69.39525104	0.632319431	0.745084589	23.74653079	23.8737478
4	72.87916521	0.803604414	0.973057473	26.18573887	26.98700249
5	75.5149999	1.059157946	1.142510531	27.96616638	28.25540698
6	77.5376486	1.202455187	1.344225823	29.74058925	30.68341996
7	79.19370876	1.33425201	1.904896634	31.28958357	31.95548829
8	80.56459927	1.956088759	2.050432894	32.31556332	32.87949089
9	81.77041299	3.514751576	3.785469375	33.19600681	33.43130739
10	82.82020776	6.350571408	6.61742181	33.78588665	34.30401609
11	83.74583873	6.461301094	6.778224112	34.60658903	34.91903401
12	84.59218806	6.609076374	7.247400448	35.19463694	35.45456673
13	85.35091031	6.768555619	7.379095497	35.67673852	36.18306223
14	86.04928082	7.241395805	7.697646892	36.3277843	36.83705943
15	86.6925918	7.361895757	7.810615598	36.93313372	37.6516554
16	87.28969756	7.530839948	8.618901598	37.55303677	38.1456645
17	87.84772226	7.644113974	8.745507967	38.25456226	38.72434923
18	88.36055947	7.766445852	8.92381533	38.81380824	39.27107706
19	88.84591781	7.926128644	8.99800829	39.15128952	39.82503081
20	89.30206712	8.076142943	9.225065211	39.5829114	40.29756568
21	89.72687017	8.204479464	9.365716339	40.0340738	40.86251098
22	90.13019898	8.345537685	9.462808361	40.59220027	41.37402514
23	90.50574359	8.465935864	9.524076074	41.08051005	41.93785093
24	90.86510621	8.653809946	9.64976648	41.59182066	42.36591073
25	91.20584425	8.914146837	9.832958976	42.05275333	42.76964663
26	91.52632934	9.699228454	9.985721163	42.50025189	43.18651133
27	91.83287145	9.842729243	10.2624436	42.8484194	43.83145069
28	92.11966948	9.967401913	10.4143916	43.2221321	44.20078712
29	92.39659547	10.09421183	10.56277752	43.66148876	44.78852971
30	92.66151217	10.26386844	10.71218118	43.95408789	45.36324528

8. 考　　察

8.1 ベースラインとの比較

提案手法にて，上位 X%にランクインした単語のうち，ベースライン（図 1）の単語の出現数とベースラインの総単語数 (323 語) に対する出現率を表 8 に示す．

表 5　コーパス 1 での正解率

上位 X%	TF	1/IG	1/CHI	TF/IG	TF/CHI
1	76.90582871	78.83424433	78.82452565	84.42814812	85.03722684
2	66.05098347	78.96864229	78.96561528	85.36113134	85.61713279
3	64.80334174	79.01377015	78.92822799	85.83484909	85.80300818
4	62.36276533	79.12759322	79.02520412	86.38548064	86.73345862
5	61.94955106	79.20846865	79.12220758	86.81766886	86.84081664
6	56.85976232	79.36150652	79.19609344	87.14833688	87.35235747
7	55.54633748	79.43248008	79.41077961	87.32295161	87.51161872
8	54.73478381	79.69997671	79.49508102	87.49340265	87.5792157
9	53.6088944	80.19564776	79.9413385	87.60449808	87.60468324
10	51.99666527	80.78893596	80.69829289	87.59887425	87.65927459
11	49.41332655	80.85364559	80.73372231	87.69362523	87.7114975
12	45.57589672	80.94112965	80.82467652	87.7466111	87.75763509
13	44.80884469	81.05251392	80.89343918	87.74014256	87.81751752
14	42.90746856	81.18733532	81.02169046	87.79488947	87.86530147
15	42.26746456	81.27159073	81.08497731	87.30659792	87.43336061
16	40.83059651	81.38506153	81.3061352	87.31501943	87.30132174
17	40.22136979	81.44785449	81.48185998	86.85413441	87.35548765
18	39.23772874	81.54072827	81.61713396	86.93869208	87.42901402
19	37.85092083	81.65964976	81.73269813	86.9570681	87.39637233
20	37.20876208	81.80930618	81.78389209	86.86696658	87.37688085
21	36.63740937	81.90688992	81.95722909	86.84793443	87.38427494
22	35.48136749	81.92539486	82.05079057	86.82524474	87.3736467
23	34.63560053	82.08560327	82.09102962	86.93238815	87.3784002
24	33.23884925	82.16439626	82.25141092	86.87384853	87.38796379
25	32.82579437	82.31797921	82.30301487	86.87143811	87.42140238
26	31.76064133	82.53965185	82.46648531	86.89673616	87.41774264
27	31.10283625	82.62972648	82.47557977	86.91328143	86.91197699
28	30.73940005	82.68269504	82.63570396	86.92760442	86.85731953
29	29.83983691	82.81810047	82.7263909	86.94371044	86.93257087
30	29.21152876	82.87221207	82.69835408	86.99025875	86.91427377

表 6　コーパス 2 での削減率

上位 X%	TF	1/IG	1/CHI	TF/IG	TF/CHI
1	59.20558633	0.115413571	0.116841469	8.047437735	8.029564388
2	67.79592084	0.225854117	0.226838874	10.31331535	10.31331535
3	72.97205209	0.306554989	0.300400255	19.15811118	19.17263635
4	76.51028555	0.397989716	0.397989716	21.61385002	20.85317414
5	79.09596313	0.538613077	0.535511092	25.23229197	22.9849278
6	81.09118017	0.673278656	0.677513113	27.39019093	27.50309337
7	82.69160795	0.733447335	0.739602069	30.88819704	30.87795556
8	84.01615593	0.802921972	0.79927837	34.01608207	31.18342732
9	85.14311234	0.946795033	0.960384686	34.27699356	34.14296807
10	86.14958367	1.049308282	1.075355116	34.61338669	34.6843877
11	87.03522527	1.181462729	1.138281116	36.29786352	35.76545441
12	87.80963851	1.322775421	1.256993624	37.18793652	37.57814666
13	88.49153379	1.414259386	1.40421486	39.6754338	38.21365987
14	89.10828737	1.50623573	1.535876929	41.20505752	41.29644301
15	89.67161786	1.623865006	1.609832212	42.07326891	42.23550769
16	90.18792618	1.8516394	1.694619827	42.56141317	42.41192699
17	90.66134832	1.933768169	1.944058885	43.52573688	43.36876655
18	91.1031105	2.023725761	2.033868762	44.57731011	44.69459472
19	91.50961837	2.093052684	2.162231894	44.94935146	45.49008177
20	91.88456476	2.376662824	2.381241947	46.02327868	46.25543525
21	92.23060853	2.556331818	2.451455151	46.89183474	46.94476545
22	92.55572619	2.707935225	2.686615226	47.31242464	48.49433051
23	92.86031167	2.801142516	2.775391109	48.40333892	49.34063104
24	93.14697455	2.89055849	2.847573828	49.80533807	50.52992259
25	93.41669962	3.009074047	2.916112946	50.65336194	51.06405502
26	93.6686498	3.183622302	3.044968456	51.87618448	51.24889399
27	93.90755196	3.251373614	3.247188394	52.60264004	52.22665965
28	94.13375074	3.359007601	3.370283073	53.10077958	52.5678781
29	94.3492649	3.488650917	3.519129159	53.50295452	53.36897827
30	94.55365131	3.6056401	3.621297743	54.02704242	53.88277546

上位 30%で見たとき，ベースラインの半数程度が不要語の条件を満たす．表 3 より，ベースラインが分類精度の向上に効果的である．しかし，残り半数の単語は不要語の条件を満たしておらず，ベースラインの全ての単語が不要語として機能しているとは言いがたい．このことから，ベースラインのリストは，さまざまなコーパスに広く対応しているが，各コーパスにとって最適な不要語とは言えない．

8.2　SNS 特有の不要語

提案手法にて，コーパス 1，コーパス 2 両方で上位 5%にランクインした単語を図 2，図 3 に示す．

と, の, 高い, ん, そして, れる, たち, 上がる, お, で, t, 元気, 全て, 多分, 頃, が, あと, あり, 貰う, かも, くる, 非常, 民, アレ, でも, ない, なる, はぁ, 的, ど, ある, うん, 始める, あげる, 無い, する, その, 実質, どー, 微妙, もう, みたい, 帰る, や, ら, しれる, ツイート, 勢い, 別, こと, カッコ, まさに, <U+0E51>, あたり, マジ, たぶん, 色々, できる, ちゃ, ちる, 大変, もはや, はい, 上, 突然, 変更, いいね, まつ

図 2　両方のコーパスで TF/IG 上位 5%の単語

「ツイート」「マジ」などの SNS 特有の表現のほか，顔文字やユニコードもコーパスを跨いで，不要語として抽出されている．

8.3　TF/IG，TF/CHI の比較

TF/IG，TF/CHI の 2 つの指標間のケンドール順位相関によると，コーパス 1 では 0.8986，コーパス 2 では 0.9462 となる．すなわち，2 つの指標には強い相関がある．また，2 つの指標上位 30%以上にランクインした単語のうち，指標間での順位の差が大きい語 top50 を表 9，表 10 に示す．

と, の, ん, そして, れる, たち, お, あと, で, t, 高い, 元気, 全て, 多分, 上がる, が, 貰う, かも, くる, 始める, 非常, 民, でも, ない, なる, はぁ, 的, もう, 頃, 帰る, ど, ある, あげる, カッコ, まさに, <U+0E51>, 無い, する, その, アレ, 実質, どー, みたい, や, ら, あり, しれる, うん, ツイート, 学校, 勢い, 別, こと, (^^), あたり, マジ, たぶん, 色々, できる, ちゃ, ちる, 大変, もはや, 反応, はい, 上, 突然, いいね, まつ, 微妙, 案件, 頼む

図 3　両方のコーパスで TF/CHI 上位 5%の単語

単語によっては順位に 1000 以上の差があるものも存在する．このように，それぞれの指標で抽出される不要語には差が生じている．上位 1%から 30%の各閾値で，指標間の差の平均は，コーパス 1 では TF/CHI の方が削減率 0.67%，正解率 0.23%上回る．一方，コーパス 2 では TF/IG の方が削減率 0.17%，正解率 0.05%上回る．これより，どちらが不要語指標としてよりよいのかについては一概に言えない．

9.　結　　論

本稿では単語頻度，情報利得，χ^2 統計量といった統計量を，SNS の時系列データを用いたカルバック・ライブラー情報量によるカテゴリ分類から算出し，組み合わせた不要語指標を提案した．実験では提案手法と Google と SlithLib が公開してい

表 7　コーパス 2 での正解率

上位 X%	TF	1/IG	1/CHI	TF/IG	TF/CHI
1	87.57940127	85.05436796	85.05436796	86.35149187	86.35512031
2	83.12747268	85.11533939	85.11533939	86.56857785	86.56857785
3	74.18709514	85.1778296	85.17823846	88.11134578	88.12732029
4	71.48860137	85.20219445	85.20219445	88.53165113	88.3174657
5	69.35105135	85.23736772	85.23716329	89.27633381	88.77577812
6	60.79242669	85.24647795	85.24898146	89.64811468	89.67587939
7	60.00647285	85.27965393	85.29231183	90.48051675	90.47198672
8	57.02326666	85.36133946	85.36154388	91.27306434	90.48073555
9	53.02842618	85.47200453	85.46567027	91.34358898	91.31374936
10	47.33893979	85.56789436	85.57251743	91.38793129	91.38461008
11	46.16799212	85.66607369	85.67439731	91.77489172	91.57621904
12	44.72489004	85.7560464	85.72580309	91.84780447	91.94545639
13	44.05095495	85.79814078	85.79375089	92.63567248	92.06022007
14	42.6585451	85.84124225	85.83816735	93.1595112	93.20731609
15	40.48154558	85.88231129	85.88593974	93.33291057	93.37580482
16	39.4788214	86.01977981	85.98696326	93.40920076	93.43989394
17	38.35806644	86.05515874	86.05125177	93.81104202	93.61293017
18	37.46658502	86.10889803	86.12324168	94.08269546	94.13186657
19	37.05738613	86.12146299	86.10464017	94.1323025	94.30047558
20	36.13985343	86.1370522	86.1119204	94.37573391	94.34935803
21	35.91574268	86.2227364	86.25023581	94.42886568	94.43655348
22	35.26110598	86.29126673	86.29717988	94.50421609	94.81273288
23	34.46229186	86.37913851	86.37346457	94.74023365	95.06709336
24	33.12731653	86.44030622	86.45235135	95.12131463	95.33087312
25	32.60478674	86.47248074	86.50280578	95.28186692	95.35549024
26	31.99355093	86.51314284	86.51083376	95.38522048	95.37046366
27	30.39461267	86.60911281	86.61714515	95.51262018	95.47861958
28	29.95114646	86.68719279	86.68732103	95.54086018	95.44579039
29	28.8048148	86.76891477	86.79018771	95.57107542	95.55960047
30	28.30743993	86.78154504	86.79714192	95.59829431	95.64159557

表 8　上位 X%でのベースラインの単語の出現数

	コーパス 1			
	TF/IG		TF/CHI	
上位 X%	出現数	出現率	出現数	出現率
5	57	17.64706	58	17.95666
10	102	31.57895	103	31.88854
15	119	36.84211	120	37.1517
20	129	39.93808	132	40.86687
25	146	45.20124	144	44.58204
30	159	49.22601	159	49.22601
	コーパス 2			
	TF/IG		TF/CHI	
上位 X%	出現数	出現率	出現数	出現率
5	41	12.6935	40	12.3839
10	68	21.05263	67	20.74303
15	102	31.57895	102	31.57895
20	117	36.22291	118	36.53251
25	146	45.20124	143	44.27245
30	160	49.5356	164	50.77399

る従来の不要語リストを比較し，上位 10%を閾値としたとき，単語の削減率が平均 8.90%，単純ベイズによる分類精度が平均 2.31%向上した．

今後の課題として，カルバック・ライブラー情報量や上位 X%などの閾値の推定法の検討，日本語以外での言語や twitter 以外での SNS による検討などが挙げられる．また，既存の不要語リストに代わる，SNS 専用の固定不要語リストの作成なども視野に入れたい．

表 9　コーパス 1 での指標間の順位の差

TF/IG 特有		TF/CHI 特有	
単語	指標間の順位の差	単語	指標間の順位の差
月曜から夜ふかし	1083	甘い	795
関	960	$U+0001F493$	741
etc	847	名付ける	699
ご	831	集	652
女性	766	シャイニー	648
連載	725	実行	635
文章	719	挑む	635
準備	670	某	634
ゆう	664	ハイライト	627
万が一	663	方々	620
自宅	663	稲妻	614
返せる	654	義務	606
#マツコ	654	昼	578
ケース	654	社会人	576
届け	643	チューズデイ	574
今回	629	5 日	564
住む	606	Google 検索	564
いかん	600	公務員	563
ギリギリ	587	ドラゴン	562
スカディ	577	火星大接近	559
夫	574	中二病	558
ワケ	574	賛成	552
行為	574	プレス	551
Amazon	574	定休	546
セガ	574	is	543
足る	572	ハッピー	533
表情	570	LINE NEWS	532
途端	569	1 日	531
#山田健太	569	#linenews	530
$U+035C$	564	木曜	521
お題	564	うわ	519
急上昇	564	つけ	518
命令	564	ブラック	518
魔法少女	564	だけど…	518
あぁ	560	そのうち	518
2 人	559	アンケート	517
今週	554	使える	516
$U+0001F60F$	548	誕生	514
第一	544	ごく	507
TOP	544	層	505
活かす	544	ガンダム	501
広い	535	オール	498
見当たる	534	菅野完	495
発揮	523	照らす	495
落ち込む	522	考える	495
$U+25BF$	521	人達	494
イイ	519	社会	493
夢	519	マックスター	493
はいる	517	罰則	491
PC	517	輝かしい	489

文　献

[1] 徳永　健伸, ”情報検索と言語処理（言語と計算）”，東京大学出版会, 1999.

[2] Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han, Lu Sheng Wang, ”Automatic Construction of Chinese Stop Word List”, Proceedings of the 5th WSEAS International Conference on Applied Computer Science, 2006.

[3] https://sites.google.com/site/kevinbouge/stopwords-lists

[4] http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/N

[5] Hassan Saif, Miriam Fernandez and Harith Alani, ”Automatic Stopword Generation using Contextual Semantics for

表 10　コーパス 2 での指標間の順位の差

TF/IG 特有		TF/CHI 特有	
単語	指標間の順位の差	単語	指標間の順位の差
うい	385	燃える	401
食べる	384	七	398
スト	374	花火	398
図	370	$U+0001F4AE$	398
拝む	370	自然	382
主演	364	願い	357
頭痛	361	叶える	356
やから	361	チラッと	349
全体的	361	非	349
この間	346	宇宙	346
武器	344	$U+0E34$	334
しむ	343	お気	334
えええ*	343	the	332
策	343	行動	327
許せる	343	風邪引	326
つまむ	343	ロマンティック	326
感慨深い	336	寄り	326
$U+0001F644$	332	決意	326
おい	331	ほん	319
てっきり	330	レッスン	303
オリジナル	326	おかえりなさい	303
煽る	326	塵	303
昼寝	326	トコ	303
だぁ	325	以内	303
による	318	現場	303
神様	316	真夜中	297
バージョン	315	光	270
最初	311	たく	263
うつ	310	ピーク	255
有名	308	そしたら	252
もと	307	‘」∠) :	250
悪	307	カネ	250
お預け	307	(▽ ;)	249
ぁぁぁあ	307	入る	248
並み	307	ブランケット	247
画	307	録	247
なり	304	さぶ	246
溶ける	291	1 枚	239
or	291	1990 年	239
入り	288	15 日	238
正月	286	青い	238
使い	286	雨	236
自身	281	次に	234
はじめ	279	あまりに	233
により	274	それでも	232
そむ	272	いつの間にか	231
限る	270	絶望	228
日本	265	想像	227
したう	264	込む	227
どうでしょう	264	有る	227

Sentiment Analysis of Twitter”, The 13th International Semantic Web Conference (ISCW), 2014

[6] Yiming Yang, Jan O. Pedersen, ”A Comparative Study on Feature Selection in Text Categorization”, Proc. International Conference on Machine Learning (ICML), 1997, pp.412-420

[7] 園田　匠, 三浦　孝夫, ”条件付き連語を用いた名詞句の文脈抽出”, 第六回データ工学と情報マネジメントに関するフォーラム (DEIM), 2014