情報処理学会論文誌 Vol. 50 No. 9 2182–2192 (Sep. 2009)

ベイジアンフィルタにおける言語知識を用いないトークン抽出方式の提案と評価

藤田拓也[†1,*1] 松本章代[†2]
テュールスト マーティン ヤコブ[†2]

近年，社会問題ともなっているスパムメールに対抗するために，ベイズ理論を応用したスパムメールフィルタであるベイジアンフィルタが脚光を浴びている．しかし，社会環境のグローバル化により，多言語環境においても利用可能なスパムメールフィルタが求められている現在において，言語や文字コードの知識を用いないベイジアンフィルタは十分に検討されたとはいえない状況である．そこで本論文では，ベイジアンフィルタに最適な，言語知識を用いないトークン抽出方式の提案と評価を行う．具体的には，電子メールの構造に基づいたトークンへの属性付与や，適切なトークン長のバイト単位 N-gram によって，実用的な判別精度を持ったスパムメールフィルタが実現できることを明らかにする．また，言語の異なる複数のメールコーパスを用いた実験によって，言語や文字コードの知識を用いる既存手法との比較を行い，提案手法の有効性を示す．

Proposal and Evaluation of Improvements for Language-independent Tokenization in Bayesian Spam E-mail Filters

Takuya Fujita,[†1,*1] Akiyo Matsumoto[†2]
and Martin J. Dürst[†2]

Recently, Bayesian filters have attracted attention as a means to combat spam E-mail, which has become a social problem. However, not enough attention has been given to Bayesian filters that do not use knowledge about language or character encoding. This is an important requirement in today's multilingual society. This paper proposes and evaluates methods of language-independent token extraction optimized for Bayesian filters. We use byte-level N-gram tokens of appropriate length and assign attributes to these tokens based on E-mail structure. This leads to a spam filter with a discrimination accuracy high enough for use in practice. We also compare our proposed methods with existing methods that use knowledge about the language or character encoding using several E-mail corpora with different languages, and show the effectiveness of the newly proposed methods.

1. はじめに

近年，研究者の情報共有の場であったインターネットは急速に発展し，人々の生活に欠かせないものとなった．なかでも，既存のインフラに対して高い利便性を持った電子メールは広く一般に利用されるようになっている．しかし，インターネットの黎明期に設計された電子メールのプロトコルは，その当時のネットワーク状況を反映し，セキュリティよりも送信先に届くことを重視している．そのため，悪質な業者などが受信者の望まない電子メール（以下スパムメールと呼ぶ）を無差別かつ大量に送信するという問題が発生している．スパムメールが増加することにより，受信者がスパムメールと，スパムメールでない電子メール（以下ハムメールと呼ぶ）を選り分ける手間を強いられたり，重要な電子メールを見落としてしまうといった事態が発生している．

このような状況を受け，ベイズ理論を応用したスパムメールフィルタであるベイジアンフィルタが注目を集めている．ベイジアンフィルタとは，以前に受信したスパムメールおよびハムメールに現れたトークンを統計的に学習することで，その後受信した電子メールがスパムメールであるかハムメールであるかを判別する手法である．ベイジアンフィルタは，それまで用いられていたルールベースのスパムメールフィルタに対して判別精度が高いだけでなく，利用者が実際に受信した電子メールをもとに学習することから，個人の趣向に合わせたフィルタを低コストに構築できるというすぐれた特徴を持つ．

しかし，ベイジアンフィルタを用いるためには，一続きとなっている電子メールのテキストから何らかの方式でトークンを抽出する必要がある．トークンとして単語を選択した場合，英語であれば空白で区切られた部分を単語として取り出すことなど，日本語であれば形態素解析などが行われる．これらの処理には対象となる言語の知識が必要となる．そのた

†1 青山学院大学大学院理工学研究科理工学専攻
Graduate School of Science and Engineering, Aoyama Gakuin University
†2 青山学院大学理工学部情報テクノロジー学科
College of Science and Engineering, Aoyama Gakuin University
*1 現在，ソニー株式会社
Presently with Sony Corporation

め，ある特定の言語に向けて設計されたスパムメールフィルタを他の言語に対して用いると，予期しない判別精度の低下を引き起こす可能性がある．

今日，社会環境のグローバル化により，個人が様々な言語で書かれた電子メールを受け取るケースが増加している．また，スパム送信者は電子メールにおいて一般的でない文字コードや短縮語を用いることなどによって，スパムメールフィルタをすり抜けようとする．そのため，フィルタにおいて言語や文字コードの知識を用いることは，利便性や判別精度の向上という観点において必ずしも適当であるとはいえない．

そこで本論文では，言語や文字コードに対する知識を用いないベイジアンフィルタにおいて，判別精度を向上させる手法を提案する．具体的には，トークン抽出時にトークンの出現箇所という属性に着目することで，N-gram のトークン長が短い場合の判別精度を向上させる．また，バイト単位の N-gram が適切なトークン長において，知識を用いるトークン抽出方式に対しても高い判別精度を持つことを明らかにする．

以下，2 章では，ベイジアンフィルタの概要と本論文において用いた方式について述べる．3 章では，テキストに対するトークン抽出方式について述べる．4 章では，電子メールの構造を用いた属性付与手法について述べる．5 章では，電子メールのヘッダと本文それぞれについて知識を用いる方式と用いない方式を比較した実験の概要について述べる．6 章では，実験結果と考察について述べる．また，実験結果から求めた最適なパラメータによる言語知識を用いないトークン抽出方式について述べる．7 章では，本論文のまとめと今後の課題について述べる．

2. ベイジアンフィルタ

ベイジアンフィルタとは，以前に受信したスパムメールおよびハムメールに現れたトークンを統計的に学習することで，その後に受信した電子メールのスパム確率を計算し，スパムメールをフィルタリングする手法である．利用者の主観によるラベリングに基づいて判別ルールを生成するため，個人によって異なるスパムメールの定義に合わせたフィルタリングを行うことが可能である．しかし，トークンの抽出までは統計的に行えないため，一般的には利用環境に合わせた言語や文字コードの知識を用いたトークン抽出の実装を必要とする．**図 1** にベイジアンフィルタにおける一般的な処理フローを示す．初めに，それまでに受信したスパムメールおよびハムメールそれぞれからトークンを抽出する．その後，助詞や助動詞など判別の際にノイズとなるトークンを除去することで有効なトークンを選別し，スパムメールおよびハムメールへの出現の偏りからトークン単位でのスパム確率を計算する．

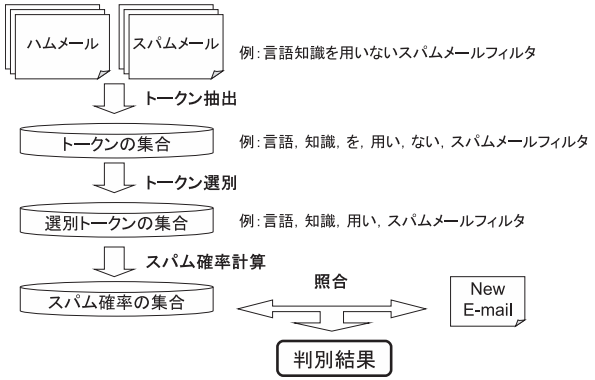


図 1　ベイジアンフィルタの処理フロー
Fig. 1　Bayesian filter processing flow.

最後に，新しく受信した電子メールから抽出したトークンのスパム確率の結合によって電子メールのスパム確率を計算し，確率の高いものをスパムメール，低いものをハムメールと判定する．次節より本論文において対象とするベイジアンフィルタ方式について解説する．

2.1 Paul Graham 方式

ベイズ理論を用いたスパムメールフィルタは以前から研究されていた[1)] が，2002 年に発表された Paul Graham 方式[2)] をきっかけとして，統計的な手法がスパムメールフィルタにおいて主流となった．Paul Graham 方式では，まずコーパスに存在する頻度が 5 以上のトークン w_i のトークンスパム確率 $P(w_i)$ を以下の式で計算する．

$$P(w_i) = \frac{\frac{b(w_i)}{n_{bad}}}{a\left(\frac{g(w_i)}{n_{good}}\right) + \left(\frac{b(w_i)}{n_{bad}}\right)} \tag{1}$$

a：バイアス（Paul Graham は 2 を使用）
$g(w_i)$：トークン w_i のハムメールにおける文書頻度
$b(w_i)$：トークン w_i のスパムメールにおける文書頻度
n_{good}：ハムメールの総数
n_{bad}：スパムメールの総数

判別対象である電子メールのスパム確率 P は，最も特徴的な（スパム確率が 0.5 から最も遠い）15 個のトークンのスパム確率を以下のように結合することで求められる．この結

合方式を，ベイズの結合確率と呼ぶ．

$$P = \frac{\prod_{i=1}^{15} P(w_i)}{\prod_{i=1}^{15} P(w_i) + \prod_{i=1}^{15}(1 - P(w_i))} \tag{2}$$

ここで，特徴的な 15 個のトークンのみを用いるのは，スパム送信者が関連性のない文書を含む長いスパムメールを送信してくる場合への対処である．つまり，スパム確率の高いトークンを中立なトークンによって打ち消させないようにしている．最終的にスパム確率 P は 0 から 1 の範囲の数値となり，0.9 を超えた電子メールをスパムメールとして判定する．

2.2 Gary Robinson 方式

Gary Robinson 方式[3] は，低頻度トークンに対するトークンスパム確率計算に関して Paul Graham 方式を改良したものである．Gary Robinson 方式では，まず Paul Graham 方式のトークンスパム確率 $P(w_i)$ をバイアスをかけずに計算し，その後にスパムの事前確率を含めた最終的なトークンのスパム確率 $F(w_i)$ を計算する．Paul Graham 方式では頻度が 5 未満のトークンはスパム確率が信頼できないことから判別に用いなかったが，Gary Robinson 方式では最終的なトークンのスパム確率を，コーパスから求めたスパム確率，スパムの事前確率それぞれの信頼度による加重平均によって求めることで，すべてのトークンを扱えるようにしている．

$$P(w_i) = \frac{\frac{b(w_i)}{n_{bad}}}{\left(\frac{g(w_i)}{n_{good}}\right) + \left(\frac{b(w_i)}{n_{bad}}\right)} \tag{3}$$

$$F(w_i) = \frac{s \cdot x + n \cdot P(w_i)}{s + n} \tag{4}$$

x：スパムの事前確率（本論文では 0.5 を使用）

n：トークン w_i の受信した電子メール全体における文書頻度

s：スパムの事前確率の信頼度（本論文では 0.001 を使用）

判別対象の電子メールのスパム確率 S は，含まれるすべてのトークンから以下のように求められる．

$$P = 1 - \left\{\prod_{i=1}^{m}(1 - F(w_i))\right\}^{\frac{1}{m}} \tag{5}$$

$$Q = 1 - \left\{\prod_{i=1}^{m} F(w_i)\right\}^{\frac{1}{m}} \tag{6}$$

$$S = \frac{P - Q}{P + Q} \tag{7}$$

最終的にスパム確率 S は -1 から 1 の範囲の数値となり，1 に近い電子メールをスパムメールとして判定し，-1 に近い電子メールをハムメールとして判定する．なお，本論文で行った実験はすべて Gary Robinson 方式を用いている．

3. トークン抽出方式

ベイジアンフィルタでは，トークンを単位としてスパム確率を計算する．しかし，電子メールのテキストは一続きになっているため，その中からトークンを抽出することが必要となる．このトークン抽出方式はベイジアンフィルタの性能に大きく影響するため，様々な工夫が試みられている．トークン抽出方式は対象となる言語に応じて様々な方式が存在し，それぞれ処理速度，メモリ消費量，そしてベイジアンフィルタに用いた場合の判別精度などで一長一短である．次節より，本論文において検討または比較の対象としたトークン抽出方式について述べる．

3.1 知識を用いる方式

英語のような分かち書きする言語においては，トークンは一般に単語が用いられる．これは，一般的なテキストにおいて実用上最小の意味単位を構成するのが単語であり，分かち書きする言語において最も自然にトークンとして抽出できる単位だからである．

“A Plan for Spam” [2] においては，トークンを構成する文字をアルファベット，数字，ダッシュ，アポストロフィ，$マークであると定義し，それらの文字以外によって区切られるとしている．また，それらのうち数字のみで構成されるもの，および HTML のコメントはトークンでないと定義し，アルファベットの大文字小文字は区別しない．本論文ではこの方式を Paul#1 方式と呼ぶ．

“Better Bayesian Filtering” [4] においては，前述の Paul#1 方式に加えて，エクスクラメーションマークもトークンを構成する文字列であると定義している．また，数字に挟まれたカンマとピリオドをトークンとして抽出することで，IP アドレスや長い数字列にも対応する．そして，“$20-25” のような幅のある数字は，“$20” と “$25” という 2 つのトークンとして抽出し，アルファベットの大文字小文字を区別する．本論文ではこの方式を Paul#2

方式と呼ぶ．

ここで，英文からトークン抽出する場合，“is” などのように非常に一般的な単語は判別の際のノイズとなることが知られている．そのような語をストップワードと呼び，トークンとして用いないという処理が一般に行われる．本論文では，予備実験の結果や，ストップワードの除去がベイジアンフィルタの判別精度向上に貢献するという報告[5]に従って，ヘッダを除く英文において Paul#1 方式および Paul#2 方式を用いる際にはストップワードの除去を行っている．

日本語のテキストの場合には形態素解析や，文字単位での N-gram によるトークン抽出方式が一般に利用されている．日本語は分かち書きされていないため，単語に分離する際に辞書を用いた形態素解析が必要となる．また，より簡便な方法としては文字単位での N-gram を用いる方法も一般的である．本論文では，ベイジアンフィルタにおいて形態素解析と文字単位での N-gram の判別精度はほとんど変化しないという報告[6]に従って，形態素解析器 Mecab[7] を用いる．この方式を Mecab 方式と呼ぶ．

3.2　知識を用いない方式

言語や文字コードに対する知識を用いない場合，形態素解析やストップワードの除去を行うことはできない．ここで，バイト単位の N-gram を用いれば，あらゆる文字コードや言語に対して適用可能なトークン抽出方式が実現可能である．しかし，意味のある単位でトークンを抽出する既存方式に対して判別精度が低下する恐れがある．英文の電子メール本文を対象とした場合，単語単位に比べて文字単位 N-gram のほうが判別精度において優れるという報告[8]もあるものの，日本語や中国語に対しても有効であるかは未知数である．

また，バイト単位の N-gram を用いた場合，テキストの正規化が行えないという問題がある．これは同時に，文字コードそのものも判別の手がかりにすることができることも意味する．スパムメールの場合，スパム送信者がわざと間違った単語や短縮語を用いて送信してくることも考えられる．日本語の電子メールでは ISO-2022-JP を用いることが一般的であるが，スパム送信者は意図的にシフト JIS などを用いることで形態素解析や文字単位の N-gram を失敗させようとしてくる場合がある．また，電子メールは様々な言語や文字コードで送られてくる可能性があるので，言語や文字コードに依存しないバイト単位の N-gram であればロバストに対応できる．

本論文ではトークン長 1 から 6 の場合のバイト単位 N-gram を検討した．

4. 電子メールの構造に対する手法

テキストからのトークン抽出方式に関しては前章のとおりである．しかし，電子メールはプレーンテキストではなく構造を持ったテキストである．そのため，電子メールの構造をどのように扱うかに関しては検討を必要とする．電子メールの構造は RFC 5322 によって規定されている．この形式から大きく逸脱した電子メールは，メールクライアントが正しく解釈できないため表示されない．スパムメールが画面に表示されなければ，スパム送信者は目的を果たすことができない．そのため，RFC 5322 に基づく電子メールの構造に対する知識を用いることは，どのような言語で書かれたスパムメールに対しても有効なはずである．

電子メールの構造に関する知識を用いる場合，たとえば本文からトークンとして単語を抽出するのではなく，ヘッダから何らかの有用なトークンを抽出する方式が考えられる．このような方式としては，ヘッダから MTA が記録したタイムスタンプを抽出する方式[9]や，発信元 IP アドレスなどを抽出する方式[10]があげられる．

しかし，本論文ではより幅広い環境で有効なスパムメールフィルタを実現するために，可能な限り知識を用いないシンプルな方式をとる．ヘッダにおいて，どのフィールドがどのような情報を持っているという知識も可能な限り用いない．そこで，文字列のトークン抽出において，トークンの出現箇所に着目する．つまり，電子メールの構造において，その箇所がヘッダであるのか，また本文であるのか，またどのような名前のフィールドであるのかという情報を取得し，トークン抽出時にトークンの属性として付与する．このような手法は，Paul Graham 方式[4]や，ベイジアンフィルタの実装の 1 つである POPFile[11] においても用いられている．Paul Graham 方式においては特定のフィールドに対してのみ属性を付与しているため，より幅広い範囲のフィールドへ付与した際の効果に関しては検証されていない．また，POPFile においては，フィールド名だけではなく，HTML の構造などにも着目した属性を付与している．しかし，本論文においては，対象が電子メールであるということ以上の知識を用いないスパムメールフィルタを目指しているため，HTML に関する知識を用いることはできない．また，上記の方式では言語知識を用いたトークン抽出方式を採用しているため，言語や文字コードに対する知識を用いない場合の効果について検証する必要がある．

表 1 に，本論文において検討の対象とした属性付与方式を示す．表 1 の各項目は，方式名およびヘッダと本文別にトークンに付与される属性を表している．String 方式は，電子メールをプレーンテキストであると見なし，すべてのトークンに対して等しく ALL という属性

表 1　属性付与方式
Table 1　Attribute assignment methods.

方式名	ヘッダの属性	本文の属性
String	ALL	ALL
Raw - Raw	HEADER	BODY
Field - Raw	（フィールド名）	BODY
Raw - MIME	HEADER	（MIME タイプ名）
Field - MIME	（フィールド名）	（MIME タイプ名）

を付与する．その他 4 つの方式は，電子メールを RFC 5322 に基づく構造化文書であると見なす方式であり，「(ヘッダの属性付与方式) - (本文の属性付与方式)」という形式で表されている．これらの方式では電子メールの MIME 構造を解析し，すべての MIME パートからヘッダと本文を読み取ったのちに，それぞれに属性を付与する．Raw 方式は，ヘッダに出現するすべてのトークンに対して HEADER という属性を付与する．同様に，本文に出現するすべてのトークンに対して BODY という属性を付与する．Field 方式は，ヘッダにおいてトークンの出現したフィールド名を属性として付与する．MIME 方式は，トークンの出現した本文の MIME タイプ名を属性として付与する．属性はトークン文字列の一部のように振る舞い，トークン文字列が等しかったとしても，属性が等しくなければ同じトークンとは見なされない．そのため，短いトークン長の N-gram を用いる場合に，意味の異なるトークンの一部が一致してしまうことを防ぐ効果が期待できる．

5. 実験概要

言語知識を用いる方式に対して高い判別精度を持つ，多言語環境に対応したベイジアンフィルタを実現するために，2 段階の実験を行う．はじめに，ヘッダ，本文それぞれのトークンのみを用いた実験を行い，属性付与方式や N-gram における最適なパラメータを決定する．その後に，最適なパラメータを用いた提案手法によって 3 種類のメールコーパスに対する実験を行う．ヘッダにおいては，バイト単位の N-gram のほかに，Paul#1 方式と Paul#2 方式をトークン抽出方式として用いた．本文においては，バイト単位の N-gram のほかに，メールコーパスの言語固有の方式を用いた．主に英語の電子メールで構成される Trec06p に対してはストップワードの除去を組み合わせた Paul#1 方式と Paul#2 方式を，主に日本語の電子メールで構成される MyMail に対しては Mecab 方式を用いた．属性付与方式としては，ヘッダは Raw 方式と Field 方式を，本文は Raw 方式と MIME 方式をそれぞれ比較し，最も判別精度のすぐれている組合せを検討した．

表 2　実験に使用したメールコーパスの詳細
Table 2　Details of E-mail corpora.

コーパス名	スパムメール数	ハムメール数	合計
Trec06p	3,713	2,287	6,000
Trec06c	4,079	1,921	6,000
MyMail	3,779	3,862	7,641

実験は，**表 2** に示す計 3 つのメールコーパスに対して行った．ここで，Trec06p は主に英語の電子メールで構成される TREC 2006 Public Corpus [12]，Trec06c は主に中国語の電子メールで構成される TREC 2006 Chinese Public Corpus，MyMail は主に日本語の電子メールで構成される著者の受信した電子メールを表す．Trec06p，Trec06c，MyMail はそれぞれ RFC 5322 に準拠した形式で電子メールを保持しており，ヘッダ，本文，添付ファイルなど実環境におけるすべてのデータを含む．Trec06p および Trec06c に関しては，コーパスに含まれていた電子メールのうち受信した日時の若い順に 6,000 通のみを利用して実験を行った．また，著者の受信した電子メールを用いたのは，日本語の電子メールによって構成される RFC 5322 に準拠した公開メールコーパスが存在しなかったためである．MyMail は，2007 年 4 月 25 日から 2008 年 11 月 24 日の約 1 年半の間に著者が 3 つのメールアカウントにおいて受信した電子メールであり，メールアカウントはそれぞれ，プライベート用，研究用，ハニーポット用である．これは，複数のメールアカウントから収集したメールコーパスであるため，メールアカウント固有の情報が実験結果に影響を与えてしまう可能性がある．そこで，以下の 3 点に関して前処理を行うことで，プライベート用メールアカウントにおいて受信したものと同様の形式に加工した．

- メールアドレスとドメインの置換
  メールアドレスとドメインをプライベート用メールアカウントのものへと置換した．
- ヘッダフィールドの削除
  以下のヘッダフィールドは Base64 などの MIME エンコードを変換した場合や，MTA において宛先メールアドレスを自動変換した場合に付与されるフィールドであるが，プライベート用メールアカウントにおいては該当するフィールドが存在しないため削除した．
  - X-MIME-Autoconverted
  - X-Original-To
  - Delivered-To

- Received フィールドの変換
  ヘッダには MTA が受信した順に Received フィールドが付与されるが，メールアカウントによって信頼できる MTA の名前や数が異なる．そこで，プライベート用メールアカウントにおける Received フィールドと同様の形式へと変換した．

評価指標としては，ハムメールを正しくハムメールとして識別できた割合である TAR（True Acceptance Rate）および，スパムメールを正しくスパムメールとして識別できた割合である TRR（True Rejection Rate）の調和平均である Accuracy を用いた．TAR や TRR はコーパスにおけるスパムメールとハムメールの比率の影響を受けないため，比率が偏っているコーパスにおいても適切にトークン抽出手法を評価できる．TAR や TRR そのものを評価指標として用いなかった理由としては，スパム確率の閾値が大きく影響することや，属性付与方式の比較を目的とした場合に，調和平均が良い値をとるためには，TAR, TRR ともに良い値でなければならないという特性が有利に働くことがあげられる．また一般に，Accuracy がより良い値をとる手法は，片方の判別精度をそろえた場合にもう片方の判別精度で良い値をとる．つまり，Accuracy は手法の素性の良さを表している．以下にその式を示す．

$$TAR = \frac{n_{good \to good}}{n_{good \to good} + n_{good \to bad}} \tag{8}$$

$$TRR = \frac{n_{bad \to bad}}{n_{bad \to bad} + n_{bad \to good}} \tag{9}$$

$$Accuracy = \frac{2 \cdot TAR \cdot TRR}{TAR + TRR} \tag{10}$$

$n_{good \to good}$：ハムメールをハムメールとして判定した数
$n_{good \to bad}$：ハムメールをスパムメールとして判定した数
$n_{bad \to good}$：スパムメールをハムメールとして判定した数
$n_{bad \to bad}$：スパムメールをスパムメールとして判定した数

実験は，利用者からの理想的なレスポンスが得られる環境を想定して行った．すなわち，利用者は時系列順に電子メールを 1 通ずつ受信し，判別結果が間違っていた場合には即座に訂正を行う．具体的な実験手順としては以下のとおりである．

step 1. 電子メールをまったく学習していない状態にフィルタを初期化し，実験の開始時間を記録する．
step 2. 電子メールをコーパスから読み込む．
step 3. フィルタを用いて読み込んだ電子メールの種別を判別する．
step 4. 読み込んだ電子メールをコーパスの正しい判別結果を用いてフィルタに学習させる．
step 5. step 2. から step 4. をインデックスファイルの順番で所定の数の電子メールに対して行う．
step 6. 実験の終了時間を記録し，電子メール 1 通あたりの判別と学習を合計した実行時間を求める．

ここで，Trec06p，Trec06c に対しては電子メールの受信順がインデックスファイルに含まれているため，そのままの順序で実験が可能である．しかし MyMail は著者の受信した電子メールのスパムコーパスであるため，受信順を保持したインデックスファイルが存在しない．そこで，ヘッダの中で最初に出現した Received フィールドのタイムスタンプの日付が若い順に，電子メールのパスをインデックスファイルへと追加した．ヘッダの中で最初に出現した Received フィールドは，直近の MTA が記録した信頼できるフィールドである．そのため，スパム送信者が自由に操作可能な Date フィールドの日付を用いるのに対して正確な受信順でインデックスファイルを作成可能である．また，すべてのトークン抽出方式において，電子メールのヘッダの一部や本文が MIME エンコードされていた場合，デコードしたのちに処理した．形態素解析を用いる場合には必要に応じて文字コードの変換を行った．なお，実験は Xeon E5430，メモリ 16 GB を搭載した PC 上で動作する Ruby 1.9.1 を用いて行っており，実行時間を計測する際には同様の実験を 3 回試行した結果の平均によって求めている．

6. 実験結果と考察

本章では，メールコーパスごとの属性付与方式とトークン抽出方式の実験結果から発見された事実を述べ，その要因について考察する．また，フィルタの実行時間を示し，判別速度の観点から提案手法の実用性について考察する．実験結果のグラフにおいて横軸はトークン抽出方式，縦軸は TAR と TRR の調和平均 Accuracy，系列は属性付与方式である．ただし，提案方式に対する実験結果においては横軸がトークン抽出方式と属性付与方式，系列がメールコーパスとなっている．また，実験結果において，Accuracy の表示範囲は 0.9 から 1.0 とした．これは，0.9 以下の Accuracy ではスパムメールフィルタの実用上問題が発生するため，他方式と比較する意味がないと判断したためである．

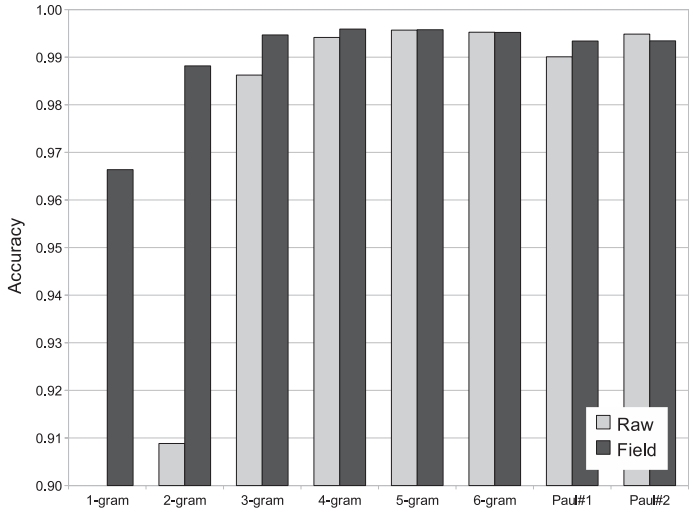


図 2　Trec06p のヘッダにおけるトークン抽出方式ごとの判別精度
Fig. 2　Accuracy of each tokenization method for Trec06p header.

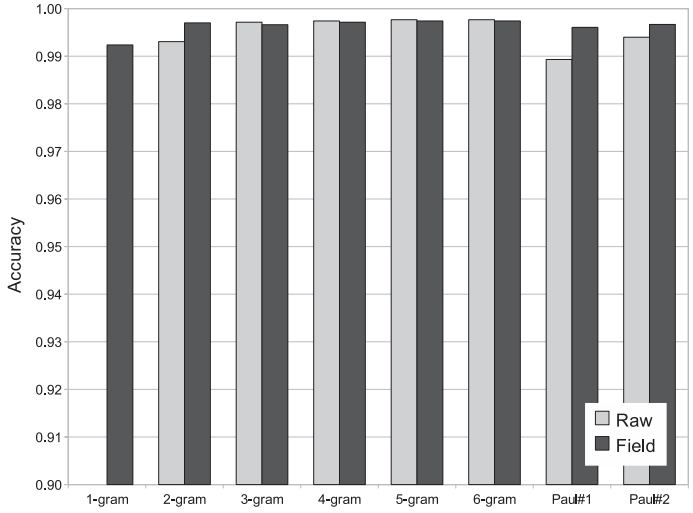


図 3　Trec06c のヘッダにおけるトークン抽出方式ごとの判別精度
Fig. 3　Accuracy of each tokenization method for Trec06c header.

6.1　ヘッダに対する実験結果

図 2，**図 3** および **図 4** にヘッダを対象とした 3 種類のメールコーパスにおける属性付与方式とトークン抽出方式の比較実験の結果を示す．

まず，すべてのコーパスにおいて，1-gram や 2-gram など短いトークン長の場合には，Raw 方式に対して Field 方式が高い判別精度を示している．これは短いトークン長において，異なる意味を持つ単語の一部であるトークンどうしが同じものとして扱われてしまうことを属性の付与によって防げているためである．Field 方式を採用する効果は，おおむねトークン長を 1 から 2 大きくする効果に等しい．トークン長が 2 倍になると，判別対象の電子メールから抽出されるトークン数も約 2 倍に増え，判別に時間がかかる．そのため，処理速度の向上に関して Field 方式の有用性が期待できる．

次に，すべてのコーパス，すべての属性付与手法において 3 程度のトークン長で判別精度は飽和している．特に Field 方式では Trec06p を除いて，1-gram の段階から非常に高い判別精度を達成しており，ヘッダにおいて長いトークンは必要ないことが分かる．

最後に，言語知識を用いるトークン抽出方式である Paul#1 方式と Paul#2 方式をバイト単位の N-gram と比較すると，おおむね 2-gram から 4-gram と同等の判別精度となって

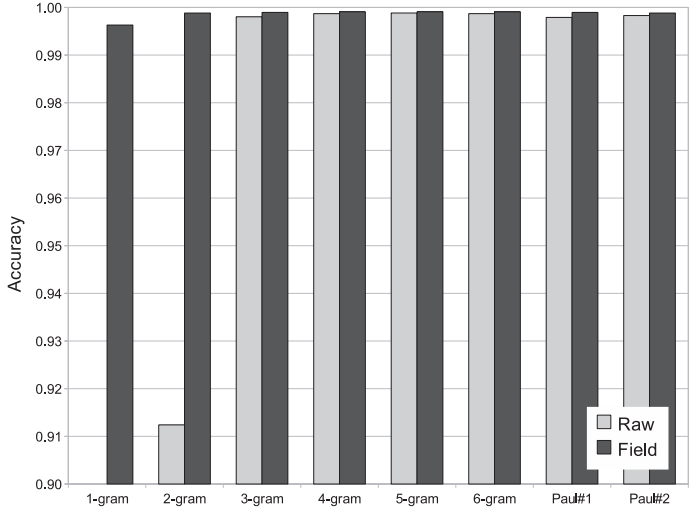


図 4　MyMail のヘッダにおけるトークン抽出方式ごとの判別精度
Fig. 4　Accuracy of each tokenization method for MyMail header.

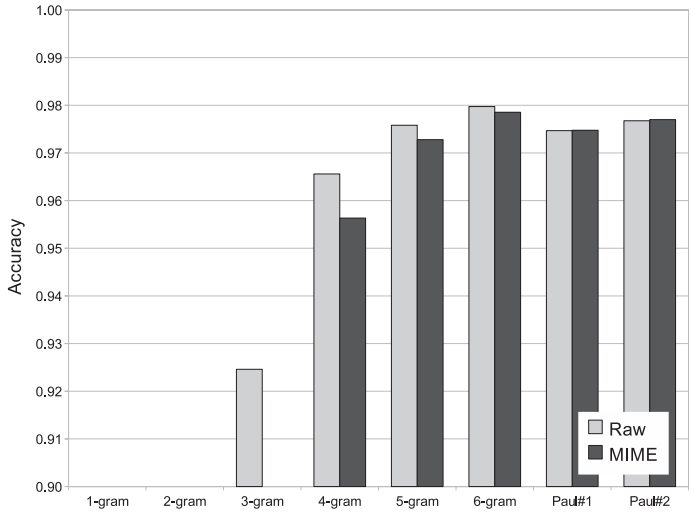


図 5　Trec06p の本文におけるトークン抽出方式ごとの判別精度
Fig. 5　Accuracy of each tokenization method for Trec06p body.

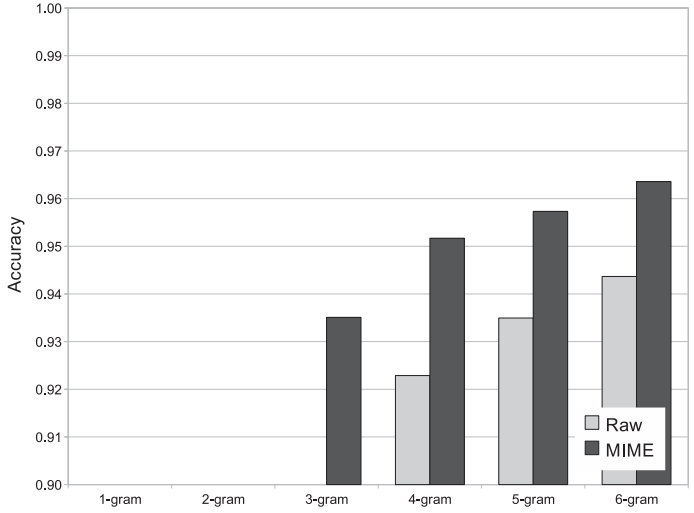


図 6　Trec06c の本文におけるトークン抽出方式ごとの判別精度
Fig. 6　Accuracy of each tokenization method for Trec06c body.

いることが分かる．そのため，言語知識を用いる方式に対して同等以上の判別精度を得ようと考えた場合，Field 方式における 2-gram 以上のトークン長を用いればよいと考えられる．

6.2　本文に対する実験結果

図 5，**図 6** および **図 7** に本文を対象とした 3 種類のメールコーパスにおける属性付与方式とトークン抽出方式の比較実験の結果を示す．

初めに，ヘッダにおいては 4-gram 程度のトークン長で飽和していた判別精度が，本文ではトークン長を長くすれば長くするほど向上する傾向がある．これは，本文がヘッダに対してより多くの語彙を含んでいるため，短いトークン長では処理できない単語が存在したためだと思われる．また，トークン長を長くした場合，トークンは文脈をある程度含むことができる．そのため，判別精度が向上していっているのだと考えられる．

次に，Raw 方式と MIME 方式の判別精度における差であるが，Trec06p 以外においてはおおむね MIME 方式の方が良い結果を残している．また，ヘッダにおける Raw 方式と Field 方式の関係と違い，トークン長が長くなっていった場合にも判別精度の差はあまり狭まっていない．これは，ヘッダではフィールド間であまり文字列が重ならないのに対し，本文では text/plain や text/html などのように非常に重なりやすい場合が存在するためだと

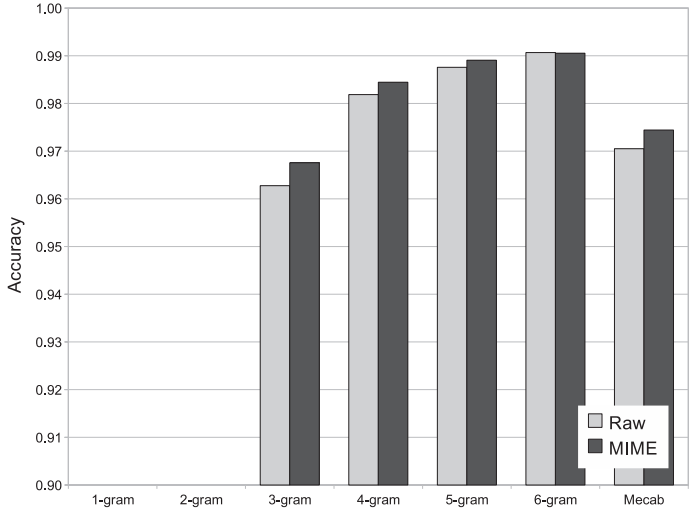


図 7　MyMail の本文におけるトークン抽出方式ごとの判別精度
Fig. 7　Accuracy of each tokenization method for MyMail body.

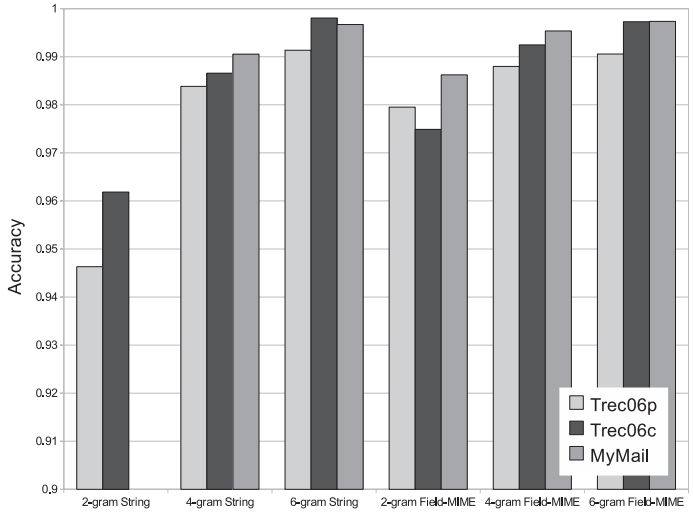


図 8　すべてのメールコーパスにおけるトークン抽出方式ごとの判別精度
Fig. 8　Accuracy of each tokenization method for all E-mail corpora.

考えられる．

最後に，言語知識を用いるトークン抽出方式とバイト単位の N-gram を比較すると，Trec06p では 5-gram において Paul 方式と同程度の判別精度を達成できることが分かる．また，MyMail では 3-gram から 4-gram において Mecab 方式と同程度の判別精度を達成できている．そのため，言語知識を用いる方式に対して同等以上の判別精度を得ようと考えた場合，4-gram 以上のトークン長を用いればよいと考えられる．

6.3　提案方式に対する実験結果

ヘッダ，本文それぞれの比較実験の結果から，おおむねヘッダにおいては 2-gram 以上，本文においては 4-gram 以上のトークン長とすることで，知識を用いる方式に対して高い判別精度を得られることが判明した．また，属性付与方式としては Trec06p の本文において Raw 方式が優れることを除いて，Field - MIME 方式が最も優れるであろうことが明らかになった．そこで，2-gram，4-gram，6-gram の Field - MIME 方式において，電子メール全体のトークンを用いた場合に，どの程度の判別精度が達成できるのかについて実験を行った．また，比較対象として電子メールの構造を用いない属性付与方式である String 方式についても実験を行った．**図 8** にその実験結果を示す．

表 3　Trec06p における電子メール 1 通あたりの実行時間（ms）
Table 3　Processing time per E-mail for Trec06p (ms).

トークン抽出方式	属性付与方式			
	ヘッダ		本文	
	Raw	Field	Raw	MIME
1-gram	13	24	22	22
2-gram	34	38	43	46
3-gram	43	43	78	92
4-gram	47	46	140	153
5-gram	53	47	239	209
6-gram	56	48	277	257

2-gram，4-gram においては Field - MIME 方式が優れているが，6-gram においては両者の差はなくなっている．これは，トークンの出現箇所という属性を付与しなくてもトークンの文字列そのものが異なっていっているためで，十分に長いトークン長をとるのであれば電子メールの構造を解釈せずに単純な文字列として扱っても判別精度的に差がないことを示している．また，判別精度としては，4-gram において 0.99 近い値をとっており，実用的な判別精度を持っていると結論づけられる．

6.4　実 行 時 間

本節では，実行時間の観点から属性付与方式とトークン抽出方式の実用性について考察する．まず，属性付与方式による実行時間の変化を示すために，代表して Trec06p における電子メール 1 通あたりの実行時間を**表 3** に示す．属性付与方式による影響を明確にするために，ヘッダもしくは本文のトークンのみを用いている．

初めに，Field 方式，MIME 方式ともに短いトークン長において，同じトークン長では実行時間が Raw 方式に対して長くなる．しかし，Raw 方式のトークン長を 1 伸ばす場合に比べると実行時間の増加量は少ない．そのため，6.1 節や 6.2 節における実験結果と考察から，同じ判別精度で実行時間を短縮するという観点において属性付与方式が有効であることが判明した．

次に，提案手法における電子メール 1 通あたりの実行時間を**表 4** に示す．

初めに，Field - MIME 方式と String 方式を同じトークン長において比較すると，すべて String 方式の方が実行時間が短い．これは，電子メール構造の解釈による処理の負荷が影響しているためだと考えられる．そのため，同程度の判別精度が得られるのであれば，Field - MIME 方式よりも String 方式がより好ましい．図 8 から，2-gram や 4-gram では判別精度において Field - MIME 方式の方が優れるが，6-gram においては両方式において差がな

表 4　提案手法における電子メール 1 通あたりの実行時間（ms）
Table 4　Processing time per E-mail for proposed methods (ms).

トークン抽出方式	属性付与方式	コーパス名		
		Trec06p	Trec06c	MyMail
2-gram	String	53	54	79
4-gram	String	131	104	252
6-gram	String	256	190	512
2-gram	Field - MIME	81	66	114
4-gram	Field - MIME	266	139	352
6-gram	Field - MIME	447	227	724

い．よって，長いトークン長においては String 方式を用いるべきであると結論づけられる．

また，それぞれの属性付与方式でのトークン長の変化に対する実行時間の変化に着目すると，表 3 の本文における実験結果と同様に，抽出されるトークン数の差以上に実行時間に差が出ている．そのため，トークン長はより短い方が好ましい．ここで，図 8 から，String 方式に比べ Field - MIME 方式では 2-gram においても比較的高い精度を達成できている．よって，判別速度を重視するのであれば，2-gram の Field - MIME 方式を用いるべきであると結論づけられる．

7. まとめと今後の課題

本論文では，多言語環境においても有効なベイジアンフィルタを構築するために，言語や文字コードの知識を用いないトークン抽出手法について比較検討した．また，実験を通して，それらの手法間における判別精度の差の要因について考察した．

結論としては，バイト単位の N-gram と属性付与方式を適切なパラメータで用いることによって，言語知識を用いる方式に対して高い判別精度を達成できることが明らかになった．具体的な方式としては，判別速度を重視するのであれば，2-gram の Field - MIME 方式が最適であり，知識を用いる方式に対してつねに高い判別精度を得たいのであれば，4-gram の Field - MIME 方式が最適である．また，6-gram など十分に長いトークン長を用いるのであれば，電子メールの構造を解釈しない String 方式においても，他の電子メールの構造を用いる方式と差のない判別精度を得られることが判明した．

今後の課題としては，トークン出現位置によって異なるトークンスパム確率計算方式を用いる方式の検討，トークン選別方式との組合せによる評価があげられる．

謝辞　本研究の諸段階において適切なご助言をくださいました青山学院大学社会情報学部伊藤一成助教に深く感謝いたします．

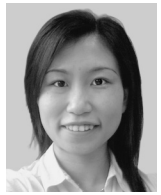
参 考 文 献

1) Pantel, P. and Lin, D.: Spamcop: A spam classification & organization program, *Proc. AAAI-98 Workshop on Learning for Text*, pp.55–62 (1998).
2) Graham, P.: A Plan for Spam. http://www.paulgraham.com/spam.html
3) Robinson, G.: Spam Detection (2002). http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html
4) Graham, P.: Better bayesian filtering, *Spam Conference* (2003).
5) Lai, C.-C. and Tsai, M.-C.: An Empirical Performance Comparison of Machine Learning Methods for Spam E-mail Categorization, *Hybrid Intelligent Systems* (2004).
6) 岩永　学，田端利宏，櫻井幸一：ベイジアンフィルタによる迷惑メール対策の効果的な利用に関する考察，火の国情報シンポジウム 2004 (2004).
7) Kudo, T.: MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/
8) Kanaris, I., Kanaris, K., Houvardas, I. and Stamatatos, E.: WORDS VERSUS CHARACTER N-GRAMS FOR ANTI-SPAM FILTERING, *International Journal on Artificial Intelligence Tools*, Vol.16, No.6, pp.1047–1067 (2007).
9) Zendejas, S.V.J. and Kanta, M.: Using time to classify spam, *IPSJ SIG Notes*, Vol.2007, No.126, pp.19–24 (2007).
10) 伊藤朋哉，寺田真敏，土居範久：発信元情報を適用したベイジアンスパムフィルタ方式の提案，情報処理学会研究報告，Vol.2008, No.21, pp.285–290 (2008).
11) Graham-Cumming, J.: Pseudo-words for spam filtering in an unmodified Naive Bayesian Text Classifier (2005). http://www.jgc.org/pdf/vb2005.pdf
12) Cormack, G.: TREC 2006 Spam Track Overview (2006). http://trec.nist.gov/pubs/trec15/papers/SPAM06.OVERVIEW.pdf

(平成 20 年 12 月 2 日受付)
(平成 21 年 6 月 4 日採録)

藤田　拓也（正会員）

2007 年青山学院大学理工学部情報テクノロジー学科卒業．2009 年同大学大学院理工学研究科理工学専攻知能情報コース博士前期課程修了．同年ソニー株式会社入社．在学中はネットワークセキュリティ，特にスパムメールフィルタに関する研究に従事．



松本　章代（正会員）

2008 年静岡大学大学院理工学研究科博士後期課程修了．2005 年東京工業高等専門学校情報工学科助手．2008 年青山学院大学理工学部情報テクノロジー学科助手．現在，同大学同学部同学科助教．博士（情報学）．自然言語処理，情報検索に興味を持つ．電子情報通信学会，日本データベース学会，教育システム情報学会各会員．

テュールスト マーティン ヤコブ（正会員）

1986 年チューリッヒ大学経済学部修士．1990 年東京大学大学院理学研究科情報科学専攻博士課程修了．理学博士．チューリッヒ大学情報科学科主任助手，慶応義塾大学特別研究助教授を経て，現在青山学院大学理工学部情報テクノロジー学科教授．専門はソフトウェア科学，インターネットやワールドワイドウェブの技術，ウェブやソフトウェアの国際化．IEEE Computer Society，ACM，Unicode 各会員．