3M4-03

様々なデータ圧縮を用いた多言語に対応する tweets の話題分類法の精度比較

王 駿キ[*1] 佐藤 栄一[*1] 延原 肇[*1]

[*1] 筑波大学システム情報工学研究科知能機能システム専攻

Abstract データ圧縮手法がどのような言語で記述されたテキストにも適用できる特性を生かし、それらをソーシャルメディアの話題分類に適用する研究が行われている。本稿では、日本語と英語のtwitter情報を、deflate, gzip, bzipなどの様々なデータ圧縮手法を用いて分類を行い、それらの特性の比較検討を行う。さらに、1次経験エントロピー法もあらたに分類手法として採用する。twitterのハッシュタグを利用したテストデータセットを対象に分類実験を行い、適合率と再現率の観点から比較を行い、データ圧縮手法による分類が、言語に依存せず機能することを示す。

Keyword 多言語、データ圧縮、話題分類、1 次経験エントロピー

1. はじめに

ソーシャルメディアの代表である Twitter は、手軽に「今」を知る、「今」を知らせることができるため、驚異的なスピードで成長を遂げている。一方で、やりとりされる情報の流れが速いため、興味のある情報を見落とす可能性が高く、これを解決するために Twitter を対象とした情報推薦や情報検索などの研究が盛んに行われている[1][2]。Twitter の話題分類において、投稿(tweet)は「新語が多い」、「文法的誤りが多い」という特徴があるため、従来の形態素解析及び、bag-of-words の表現[3]による機械学習に基づく分類手法では対応が難しい。この問題を解決するために、データ圧縮に基づく情報類似度を用いた分類手法が提案されている [4]。しかし、データ圧縮の手法は数多く存在し、多様な特徴を持っているため、Twitter の話題分類に適切な圧縮手法を明らかにする必要がある。そこで、本稿では、deflate[5], gzip[6]と bzip2[7]の 3 種類の圧縮手法を用いて分類を行う。また、情報量の視点から話題分類もできると考え、1 次経験エントロピーを用いた実験も行う。さらに、データ圧縮には言語依存せず、どのようなテキストにも適応可能という特性から、多言語に対応できることを検証するために、本稿では日本語と英語の tweets 両方に分類を行い、分類の結果及び精度を比較する。

2. データ圧縮に基づく分類方法

データ圧縮とは、データが持つ冗長性と排除することで、データのサイズを小さくすることである。2 つのデータを連結して圧縮する際、2 つのデータの類似度が高いほど、冗長な部分が多くなり、圧縮時のサイズが小さくなる。本稿で提案する手法は、西田らの手法[8]に基づき、指定した文字列(キーワード、ハッシュタグなど)が含まれる tweet のテキストを時系列順に連結したものを話題モデル A、それ以外の tweet のテキストを時系列順に連結したものを比較モデル B と定義する。Benedetto らの手法[4]に基づき、本稿ではデータの圧縮量を

$$\mathrm{C_A(x)} = \mathrm{Z(A+x)} - \mathrm{Z(A)}, \quad (1)$$

$$\mathrm{C_B(x)} = \mathrm{Z(B+x)} - \mathrm{Z(B)}, \quad (2)$$

と定義する. ここで、$\mathrm{Z(A+x)}$はモデルAと入力x(分類したいtweet)の連結したデータの圧縮後のサイズを表し、$\mathrm{C_A(x)}$はxとモデルAとの非類似度を表す。そして、xに対する分類スコアは

$$\mathrm{f(x)} = \frac{\mathrm{C_A(x)} + \gamma}{\mathrm{C_B(x)} + \gamma}, \quad (3)$$

で定義する。ここで、$\mathrm{f(x)}$が分類閾値 θ より小さく、かつ、$\mathrm{C_A(x)} < \mathrm{C_B(x)}$のとき、xはモデルBよりモデルAに類似していると判断できる。ここで、できる限り情報量の多い tweet を優先的に精度よく分類することを考慮したスムージングパラメータ γ を導入する。

テキストの 1 次経験エントロピー(平均情報量)の計算も行う。これにより、情報の圧縮限界が分かるため、各圧縮手法の性能を比較するには重要な指標として利用できる。0 次経験エントロピーと 1 次経験エントロピーは

$$H_0(\mathrm{x}) = -\sum_{c\in\sum} \frac{n_c}{|x|}\log\frac{|x|}{n_c}, \quad (4)$$

と

$$H_1(\mathrm{x}) = -\sum_{s\in\sum k} \frac{|x^s|}{|x|} H_0(x^s), \quad (5)$$

で示す。ここで、n_cはxの中の文字cの出現回数で、x^sは$s \in \sum k$の直前に出現した文字の連結である。1 次経験エントロピー法では、

$$\mathrm{f(x)} = \frac{H_1(\mathrm{A+x}) + \gamma}{H_1(\mathrm{B+x}) + \gamma}, \quad (6)$$

により分類スコアを計算する。$\mathrm{f(x)}$が分類閾値 θ より小さい場合、xはモデルBよりモデルAに類似していると判断する。

3. 評価実験

各圧縮手法の分類性能を評価するために、本研究では交差検定を用いる。分類閾値 θ を変化させ、各手法の分類の再現率と適合率を示すことにより、分類精度を分析する。本稿では、JAVA 言語を用いて分類システムを構築し、また、Twitter Streaming API を利用し tweets データを収集した。

3.1 実験データ

ハッシュタグとは、Twitter において、#記号と半角英数字で構成された特別な文字列である。ユーザが自分の投稿内容を他人により分かりやすくために、文末にハッシュタグを付ける。実験として、人気のハッシュタグを話題に設定、日本語と英語の二つのグループに分けて行う。話題の分かりやすさを考慮し、日

連絡先:〒305-8573 茨城県つくば市天王台 1-1-1
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, 王駿キ, wjq@cmu.iit.tsukuba.ac.jp

本語 tweets には為替話題の「#fxch」とラジオ番組の「#aniaca」、また英語 tweets には海外流行のゲーム話題の「#gameinsight」とアプリ話題の「#android」のハッシュタグを付けている tweets をそれぞれ話題モデル、ハッシュタグを付けていない tweets を比較モデルとする。実験はハッシュタグ毎に独立に行う。

データ圧縮の精度をより高く確保するため、本研究は tweets からリングや retweet の記号 RT などのマークをすべて除き、テキストだけに注目した。また、ユーザが検索を行う時極端に短い tweets はユーザにとって有用ではない可能性が高いため、15 文字以下の tweets をすべて取り除いている。実験に用いた tweets の具体なデータは表1および表2で示す。

表 1　実験用の日本語の tweets データ

ハッシュタグ	タグ付き	タグ付きではない
#fxch(jp)	765	65443
#aniaca(jp)	295	37501

表 2　実験用の英語の tweets データ

ハッシュタグ	タグ付き	タグ付きではない
#gameinsight(en)	468	94651
#android(en)	287	83264

3.2 実験結果——日本語 tweets の分類

図 1 と図 2 は、「#fxch」と「#aniaca」の二つの話題に対して分類精度の結果である。

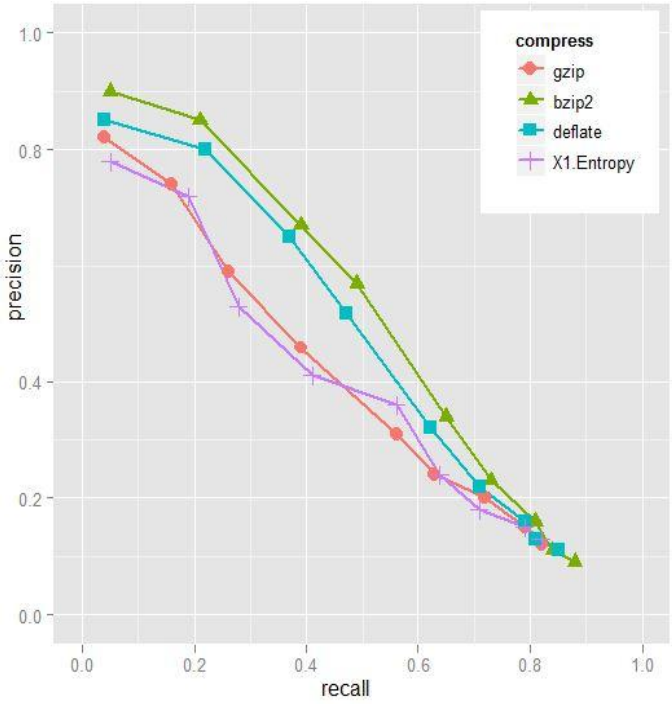

図 1　#fxch の分類結果

図 1 及び図 2 で示されるように、話題が違っても、分類精度の状況がほぼ変わらない。圧縮法では gzip と比べ、bzip2 と deflate が分類精度の観点で少々優れていることを示している。また、1 次経験エントロピーの分類精度は bzip2 と deflate より低く、gzip の精度とほぼ同じであることが見られる。

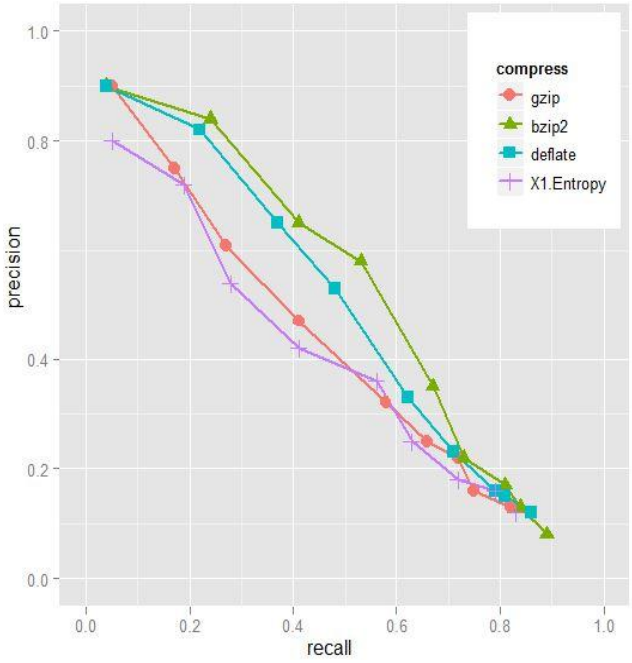

図 2　#aniaca の分類結果

3.3 実験結果——英語 tweets の分類

図 3 と図 4 は、「#gameinsight」と「#android」の二つの話題に対して、英語 tweets を分類した結果を示す。日本語の分類結果とほぼ同じように、bzip2 と deflate のほうが分類精度が上だと分かった。また、1 次経験エントロピーの曲線と gzip のと随分近いことも分かった。

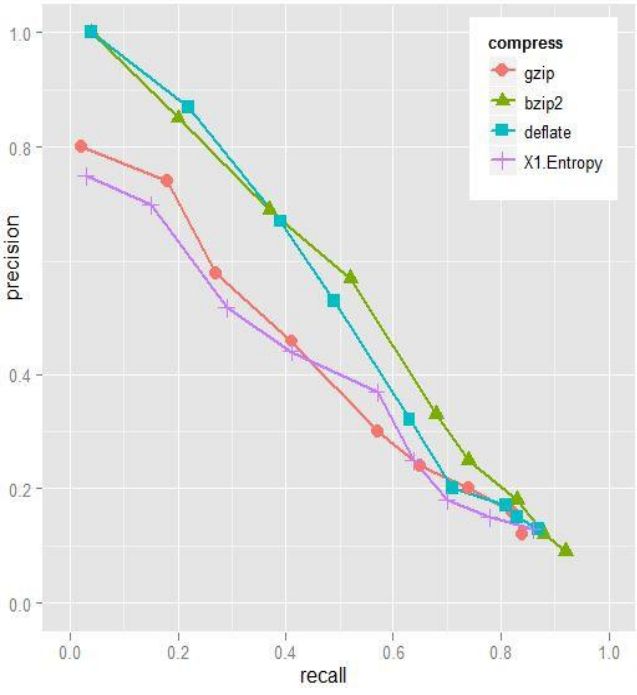

図 3　#gameinsight の分類結果

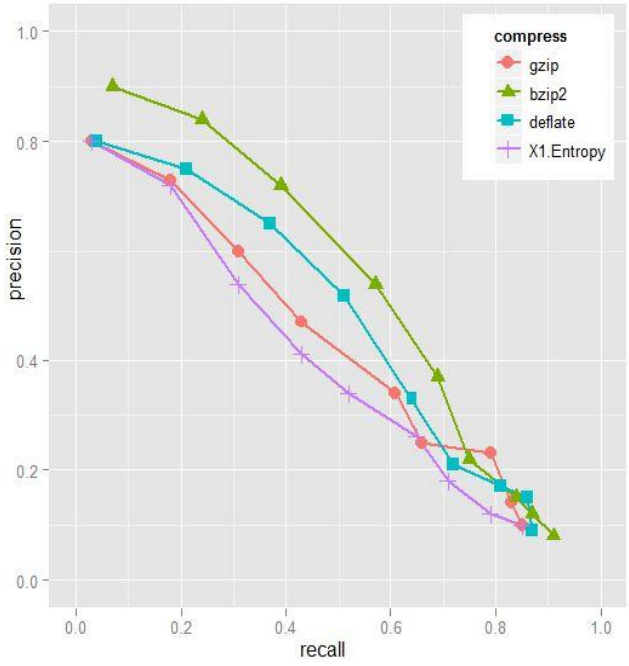

図 4　#android の分類結果

3.4 まとめ

結果として、bzip2 と deflate の分類精度が gzip 及び 1 次経験エントロピーより少し高いを分かった。1 次経験エントロピー法の場合、gzip と同じ程度の分類精度を持つことが見られる。また、日本語と英語の tweets の分類結果を総合的に見ると、分類法それぞれの精度は大した変りがない。すなわち、英語と日本語の場合、データ圧縮分類法は言語に影響されていないと判断できる。さらに、日本語より英語 tweets の場合では分類精度が少し高く、原因として、日本語 tweets では片仮名、平仮名と漢字が混ざっているため、分類するには影響になると思われる。

4. おわりに

Twitter におけて、tweets の新語や言葉的な誤りが多数存在することで、構文解析や自然言語処理の難易度が上がり、データ圧縮手法が検討される。本稿では、データ圧縮手法の多言語に対応することを検証するために、bzip2,gzip と deflate の三種類の圧縮手法を用いて tweet の話題分類を行い、また 1 次経験エントロピーを用いた分類も比較した。評価実験の結果、日本語と英語において、分類精度はあまり大きな違いはなく、つまり、データ圧縮分類法は日本語も英語も対応できると判断する。また、英語の場合分類精度が若干高いと分かり、推測として、英語 tweets の文字列は日本語より単純で、日本語でのカタカナ、平仮名と漢字変換などの影響により、分類精度が悪くなる可能性が考えられる。

今後の展望として、1. 実験の言語を増やし、より正確に多言語に対応するデータ圧縮法を把握する、2. 違う言語の tweet 検索をより精確できるため、データ圧縮に基づく tweets のカテゴリ分類システムを構築すること、が挙げられる。

参考文献

[1] 三浦 大樹 , 諏訪 博彦 , 鳥海 不二夫 , 鬼塚 真:ソーシャルサーチのための効率的な検索アルゴリズムの提案。情報処理学会論文誌. データベース 6(3), 29-39, 2013-06-28

[2] 宮西 大樹 , 関 和広 , 上原 邦昭:マイクロブログ文書の選択による擬似適合フィードバックデータベース・システム研究会報告 2013-DBS-157(15), 1-6, 2013-07-15

[3] Sivic, Josef (April, 2009): Efficient visual search of videos cast as text retrieval. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 4. IEEE. pp. 591-605.

[4] D. Benedetto, E. Caglioti, and V. Loreto: Language trees and zipping, Physical Review Letters,vol.88, no.4, 2002.

[5]Wikipedia, http://ja.wikipedia.org/wiki/Deflate.

[6] Wikipedia, http://ja.wikipedia.org/wiki/Gzip.

[7] Wikipedia, http://ja.wikipedia.org/wiki/Bzip2.

[8] 西田京介, 坂野遼平, 藤村考ほか:データ圧縮による Twitter のツイート話題分類, 日本データベース学会論文誌 10(1), 1-6,2011-06-00.