1Q-03

種々データ圧縮手法に基づく分類器の設計および性能解析とそれらの多言語 Tweets 分類への応用

王　駿き　　延原　肇

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

1. はじめに

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 の表現による機械学習・分類では対応が難しい。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[1]。この分類手法の枠組みでは、データ圧縮として様々な手法を採用することができるが、Tweets の話題分類にどの手法が適切であるかを明らかにされていない。そこで、本研究の第一の目的を、様々な圧縮手法のうち、Tweets の分類に最適な手法を明らかにすることに設定する。また、従来のテキスト分類手法では、その言語特有の文法構造を分類に利用しているため、対象言語によって自然言語解析器を変えなければならない。よって複数の言語を対象とするサービスに関して分類を行うことは難しい。これの問題に対しても、言語固有の文法などを利用しないデータ圧縮手法が有効であると考えられるため、本研究の第二目的を、複数言語への適用可能性の検証に設定する。

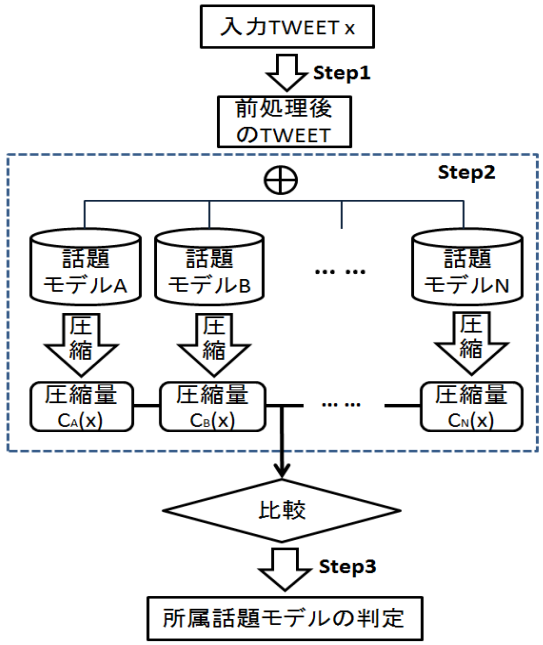

図 1　提案 N クラス分類のシステム概要図

2. 提案手法

図 1 にデータ圧縮に基づく Tweets の N クラス分類の概要を示す。Step1 では、入力 Tweet xに対して**前処理**を行う。具体的には、Tweetsに含まれる URL や画像だけではなく、句読点とハッシュタグ及びリツイートシンボルを削除する。Step2では、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 Tweetsのテキストを時系列順に連結したものを話題モデルと定義する。予め{A,B,…,N}の複数の話題モデルを設定し、これらも同様の前処理を行った複数の Tweets から構成する。その後、x とそれぞれのモデルを結合・圧縮し、圧縮量を計算する。Benedetto らの手法[2]に基づき、圧縮後のサイズが最も小さいものを、類似度が最も高いモデルとして判定し、Step3 における所属話題モデルとして決定する。

3. 評価実験

3.1 評価用実験データの収集

Tweets 自体に話題を表す属性がないため、本研究では Twitter のハッシュタグ機能を用い、話題モデルを構築する。ハッシュタグをつける時点で、ユーザがこの Tweets の内容とハッシュタグの話題に関連すると定義し、逆にハッシュタグをつけていなければ、関係がないと定義する。

本分類実験では、日本語、英語、スペイン語及びドイツ語そしてフランス語の 5 ヶ国語の Tweets に対して、Deflate、Snappy、Gzip と Bzip2 及び 1 次経験エントロピー法の 5 種類のデータ圧縮手法を適用する。

表 1　評価用 tweets データの概要

言語	話題	タグ付き(件)	その他(件)
英語	ワールドカップ	3542	281870
日本語	ワールドカップ	2794	231555
スペイン語	ワールドカップ	2389	151379
ドイツ語	アンドロイド	378	34834
フランス語	EMABiggestFans	442	37142

評価実験用に収集した Tweets の詳細を表 1 に示す。より多くの話題に関連する重複のない Tweets を収集するため、各話題に複数の関連するハッシュタグを採用し、同じ話題のハッシュタグをつける Tweets を合併して話題モデルを構築する。言語ごとに Tweets の収集も分類実験も独立して行う。

3.2 圧縮手法の分類性能比較

この評価実験では、10 分割交差検定手法を採用し、各分類手法の適合率と再現率を求め、それぞれ PR 曲線として描画する。10 分割交差検定(10-fold cross validation)では、Tweets のデータセットをランダムに 10 分割し、その中の一つをテストデータ、残りを学習データとして、分類実験を 10 回行う。この 10 回で得られた適合率と再現率の平均値を PR 曲線の描画に用いる。

極めて短い Tweets では余り有用な情報が入らなく、分類精度に大変影響があると考え、本研究では文字数制限(言わばキャラクター制限)を行うことで、性能向上を行う。実験では文字数の下限を 60 に設定し、制限以下の Tweets をすべて外す。ワールドカップを話題に指定した日本語 Tweets に対する分類結果が図 2 に示す。

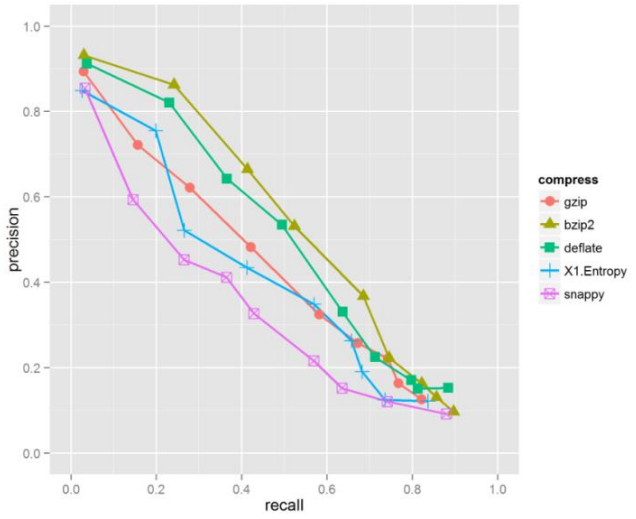

図 2　日本語 Tweets の分類結果

縦軸と横軸はそれぞれ適合率と再現率を表する。各圧縮手法の分類性能として、精度が一番高いのは Bzip2 と Deflate であり、1 次経験エントロピー法と Gzip の精度がほぼ等しく、最も精度が悪いのは Snappy であった。

3.3 多言語への適用可能性検証

日本語以外、本研究では英語、スペイン語、ドイツ語とフランス語を対象言語として Tweets の分類実験を行う。代表として、同じくワールドカップの話題を指定した対象言語が英語の Tweets の分類結果は図 3 に示す。結果として、精度の最も優秀なのは Bzip2 で、Deflate が Gzip と 1 次経験エントロピー法よりやや良く、一番精度が悪いのは同じ Snappy であることが分かった。スペイン語、ドイツ語とフランス語の分類結果は英語 Tweets の分類結果とほぼ等しい。

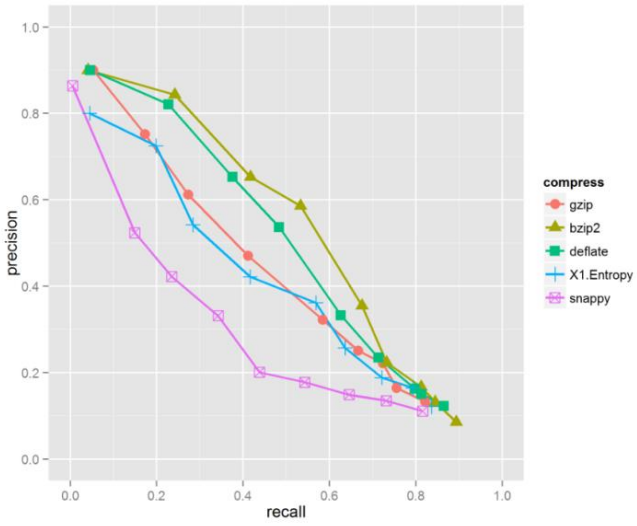

図 3　英語 Tweets の分類結果

4. まとめ

データ圧縮に基づく Tweets の分類に適正な分類手法の確定及び多言語 Tweets で応用できることの検証のため、本研究は 5 種類の圧縮手法を用い、5 ヶ国語の Tweets に対し分類実験を行う。結果として、Bzip2 が最も精度の高い手法だと判明した。また、確かに圧縮手法に基づく手法は本研究で挙げた 5 種類の言語の Tweets に対して分類手法として応用可能である。今後の展望として、1. データ圧縮手法のアルゴリズムを更に深く勉強し、より精度の高い Tweets 分類ができるデータ圧縮手法を探索する、2. 違う言語の Tweets 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] 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.
[2] D. Benedetto, E. Caglioti, and V. Loreto: Language trees and zipping, Physical Review Letters, vol.88, no.4, 2002.