RO-001

Twitter からの為替予測に特化したドメイン辞書構成法の提案
A Proposal of a method to construct domain dictionary that specializes in exchange prediction from Twitter

石垣 藍睦† 沼尾 雅之†
Aimu Ishigaki Masayuki Numao

1 はじめに

近年，SNS である mixi や Twitter といったサービスが普及しており，気軽にユーザが意見や感情を情報として発信することが一般的になってきている．このため，WEB 上に様々な情報が蓄積されている． しかし，人がすべての情報に対して目を通し，分析をすることは不可能に等しい．このため，テキスト情報からどのように有益な情報を抽出し分析するかの研究が活発に行われている．

その中の一つに，評価表現辞書の構築がある．評価表現辞書とは評価表現の集合であり，評価表現とは単語や語句に対して肯定的か否定的であるかどうかのラベル（以降，極性）を付加したものである．評価表現辞書は，文書に対して登録されている評価表現の極性を利用し，文書自体の極性の評価する際に用いられる．たとえば，ヘッドラインニュース等の経済に関連したテキストを評価することによって，景気動向が今後どのようになるか予測する研究も活発に行われている．

そこで本研究では，Twitter のテキスト情報を用いて評価表現辞書を構築した．Twitter には 140 文字の制限がある一方，気軽な情報発信が可能であり，多くの意見や感情が蓄積されている．このため，Twitter を用いた為替に特化したドメイン辞書構築法を提案し，この辞書を用いた為替予測について検討を加える．

2 関連研究

2.1 テキスト情報を用いた評価表現辞書構築

テキスト情報を用いた評価表現辞書構築では，人的コストを少なくするため，はじめに少数の単語のみを人間が用意し，そこから辞書中の単語を増やす手法がとられている．代表的な手法としては，語彙ネットワーク[1]や周辺文脈[2]を利用した手法が存在する．しかし，語彙ネットワークについては既存の言語資源を利用するため，未知語への対応が困難である[3]との指摘がある．周辺文脈法については，未知語には対応しやすいものの，被覆率が低いという問題がある[3].

2.2 テキスト情報を用いた景気動向予測

景気動向に関連のあるテキスト情報の評価を行うことによって，景気の予測が行われている．たとえば，Twitter のテキストを時系列的に評価することで，景気動向の一つの指標である株価との関係を見出す研究が行われている[4].

3 提案システム

3.1 概要

評価表現辞書には，単語とその極性とのペアからなる評価表現が登録されている．提案システムでは，ドメインごとのツイートに対してドメイン型評価表現辞書を構築する．

このため，ドメイン特有の表現（天気では「雨」と「晴れ」など）を登録することができ，またドメインによって異なる極性を付与することができる．

本研究では，那須川ら[2]の周辺文脈を利用した手法を用いて，種語から評価表現辞書の構築を行う．種語とははじめに用意する少数の既知である評価表現のことである．Twitter では新語が多く出現すると考えられるため，提案システムでは，那須川らの手法を Twitter に適するように拡張を行った．

提案システムの大まかなプロセスを以下に示す.

(1) ドメインごとのツイート DB からツイートを取得
(2) すべてのツイートから評価表現候補を抽出
(3) 評価表現候補が評価表現であるかの判定
(4) 更新する評価表現があれば辞書の更新し(1)に戻る．
更新する評価表現がなければ終了

(1)では，ツイート DB からテキストを取得する．

(2)の詳細については 3.2.1(2)節で説明する．

(3)では，出現回数を用いて評価表現候補が評価表現であるかの判定を行う．

(4)では，(3)より評価表現であると判定をされたものがあるのであれば，評価表現辞書に評価表現を更新した後に(1)へ戻る．評価表現であると判定をされたものがなければ，評価表現辞書が完成したとして辞書構築を終了する．提案システムの概要を図 1 に示す.

提案システムと那須川らの手法との違いを以下に示す.

(1) 更新される評価表現への極性付与
(2) Twitter 特有のノイズの除去
(3) 文書の単位変更

(1)については，那須川らの手法では好評か不評かどうかの判定のみ行う．しかし提案システムでは，評価表現の極性を利用し，未知の評価表現へ極性度合いを付与する．

(2)については，Twitter 特有の用語のうちノイズとなりえるものを除去する．

(3)については，一つの文書をドメインごとの指定した全ツイートと定義した．

(1),(2)の詳細については，3.2 システム実装において述べる．

3.2 システム実装

本節では，評価表現辞書を更新する評価表現への極性の付与と，Twitter 特有のノイズの除去について説明する．

†電気通信大学大学院 情報理工学専攻

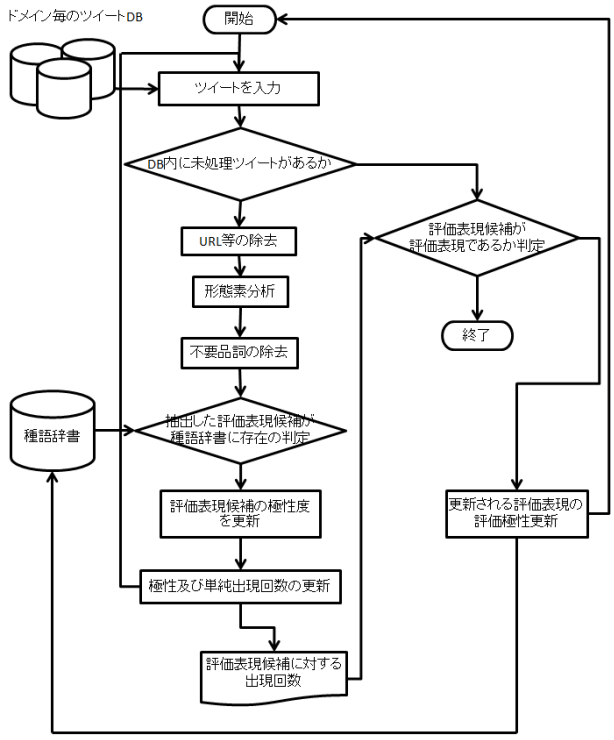

図 1.提案システムの処理フロー

3.2.1 更新時の評価表現への極性付与

評価表現辞書を構築するにあたり，少数の既知である評価表現の極性を用いて未知の評価表現に対して評価の極性を付与する．そのプロセスを以下に示す．

(1) ツイートの形態素解析
(2) 評価表現候補の抽出
(3) 極性度と種語の更新

(1)では，ツイートの本文に対して形態素解析を行う．

(2)では，(1)の結果より，品詞が｛名詞，動詞，形容詞，副詞｝である形態素を抽出し，評価表現候補とする．ただし評価表現候補とは，評価表現辞書に登録される可能性のある形態素である．

(3)では，評価表現候補の中にすでに辞書中に存在する種語が存在する場合のみ，極性度と種語数の値を更新する．極性度とは，評価表現候補と共起した種語の極性の総和である．種語数とは，評価表現候補と共起した種語の個数の総和である．

以下では，「*今日いい天気だが、明日は雨だ*」というツイートを例にとり，プロセス(1)～(3)について説明する．まず，プロセス(1)では形態素解析を行い，ツイートを形態素ごとに分ける．そして，プロセス(2)では形態素の品詞を用いて不必要な形態素を除去し，評価表現候補のみを抽出する．今回の例文に，プロセス(1),(2)を行ったものを図2に示す．

この結果，評価表現候補の中に「いい」，「雨」という種語があり，それぞれの評価の極性を，「いい」：+0.99，「雨」：-0.96 とすると，極性度はそれぞれの極性の和なので，0.99-0.96=0.03 となる．種語数は「いい」，「雨」の 2 となる．プロセス(3)では評価表現候補ごとに極性度と種語

形態素	今日	いい	天気	だ	が	明日	は	雨	だ
品詞	名詞	形容詞	名詞	助動詞	接続助詞	名詞	助詞	名詞	助動詞

不要な品詞の除去

評価表現候補	今日	いい	天気	明日	雨
種語の極性		+0.99			-0.96

図 2.プロセス(1), (2)の実行例

評価表現候補	今日	いい	天気	明日	雨
種語の極性		+0.99			-0.96

評価表現候補名	今日	評価表現候補名	天気	評価表現候補名	明日
極性度	0→+0.03	極性度	0→0.03	極性度	0→0.03
種語数	0→2	種語数	0→2	種語数	0→2

図 3.プロセス(3)の実行例（極性度と種語数の更新）

数を更新する．プロセス(3)を行ったものを図 3 に示す．

最終的に，辞書に更新される評価表現候補に付与される極性は以下にように与えられる．

$$\text{更新される評価表現の極性} = \frac{\text{極性度}}{\text{種語数}}$$

3.2.2 Twitter 特有のノイズの除去

Twitter においては，新語や口語が多いことが考えられる．このような表現の中には，為替予測のために有益な兵家もあれば，単なるノイズでしかない無益な表現もある．そのため提案システムでは，以下に無益な表現である単語の選別プロセスを示す．

(1) 対象データ内のすべての単語の出現回数を調査
(2) 出現回数が高い単語の上位 100 個を抽出
(3) 抽出された単語の中から除去する単語を選別

(1)では，対象データに出現するすべての単語を抽出し，単語のデータ内での出現回数を調べる．

(2)(3)では，出現回数が多い単語の上位 100 個を抽出し，その中から評価表現辞書を構築するにあたり，ノイズであると考えられる単語を手動で選別する．

本研究では，｛やばい，っぽい，する，いる，笑，きつね，ん｝の 7 語と敬称関連の 4 語の評価表現候補を除去する単語とする．「きつね」は，あるアプリにより不特定多数が同じツイートを行っていることがわかり，極性をつけるには不適だと考えたため除去する単語とする．その他の除去される評価表現候補は，肯定的な極性を持つ評価表現と否定的な極性を持つ評価表現どちらのツイートに含まれることが多い．そのため，出現頻度が高いこれらの単語にどちらかの極性がついてしまうとノイズとなると考え除去した．

4 実験

4.1 予備実験

4.1.1 概要と目的

評価表現辞書を構築するにあたり，はじめに種語をいくつか用意しなければならない．この予備実験では，種語の適切な個数を評価し選択するために行う．

4.1.2 実験方法

日本語のみの約 220 万ツイートに対して，種語を高村ら[1]の単語感情極性対応表から肯定・否定極性が明確な単語

上位 5, 50, 500 個ずつ選択し，評価表現辞書を構築する．以下の表 1 に，種語の数が肯定・否定の上位 5 個ずつ計 10 個を示す．

表 1.初期の種語 10 個の場合

種語	読み	極性
優れる	すぐれる	1
良い	よい	0.999995
喜ぶ	よろこぶ	0.999979
褒める	ほめる	0.999979
めでたい	めでたい	0.999645
ない	ない	-0.999997
酷い	ひどい	-0.999997
病気	びょうき	-0.999998
死ぬ	しぬ	-0.999999
悪い	わるい	-1

種語の適正な数の評価尺度は，初期の種語の数ごとに構築された評価表現辞書に登録された評価表現の数の比較，また評価表現の一致率とする．

辞書に登録された評価表現の数の比較は，初期の種語の数に辞書に登録される影響があるか調べるためである．一致率は，登録される評価表現に変化がない場合，初期の種語の数に関係なく同様の評価表現が登録することができるか調べるためである．

4.1.3 実験結果

辞書を更新する際の辞書に登録されている評価表現の数を図 4 に示す．また，構築された評価表現辞書ごとに登録されている評価表現の一致率を表 2 に示す．

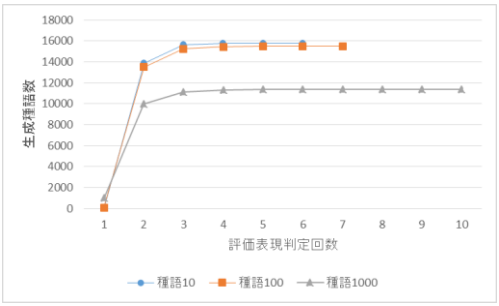

図 4 更新された評価表現数

表 2.各組み合わせの評価表現一致率

組み合わせ	一致率
種語 10 個と種語 100 個	53.55%
種語 100 個と種語 1000 個	24.13%
種語 10 個と種語 1000 個	33.55%

4.1.4 考察

図 4 より，辞書へ評価表現を更新するごとに徐々にある一定の数で収束していくことが見て取れる．

辞書の評価表現数の増加度を比較すると，種語数が 10 個と 100 個の場合では，ほぼ同数の評価表現が登録されたことが分かる．しかし，種語数が 1000 個で構築した場合では，著しく収束する数が異なる．

また，表 1 より種語数が 10 個と 100 個で登録された評価表現を比較すると，種語 10 個と 100 個の一致率は 50% 以上であるが，種語 1000 との比較では一致率が低くなった．これは，10 個, 100 個と 1000 個の間で異なる種語がノイズになると考えられる．

以上をまとめると，初期の種語を 1000 個とした場合は，種語の数が多いのにもかかわらず，登録された評価表現の数が少ない．一方初期の種語数が 10 個と 100 個では，登録された評価表現の数が多く，また評価表現の一致率が高いことから，ノイズになりにくい種語が含まれていると考えられる．そのため，初期の種語数を 10 個にすることは適切であると考えられる．

4.2 ドメイン型辞書の評価

4.2.1 概要

本実験では，ドメインごとのツイートにより辞書を構築する．また，それぞれ辞書のカスタマイズの度合いを調べることで，ドメインに特化した辞書が構築されたかを調べる．

4.2.2 実験方法

本実験では，ドメインごとの辞書を構築するために表 3 取得条件でツイートの収集を行う．

表 3.ツイートの取得条件

対象データ名	対象データの取得条件
一般	日本を囲むように緯度経度を指定，日本語の含まれるもの
為替	「#usdjpy」を含むもの，ツイートを行ったデバイスが``FX 実況ちゃんねる”

各対象データの 2 万ツイートに対して，評価表現辞書を構築することで，ドメインごとの辞書の構築を行う．辞書のカスタマイズ度合いの評価には，辞書に登録された評価表現の重複割合に比較により行う．比較した辞書の重複割合が低いほど，それぞれの対象データに依存した評価表現が存在すると考えられる．

4.2.3 実験結果

一般のツイートで構築した辞書に登録されている単語の数は，約 4000 語であり，為替のツイートで構築した辞書に登録されている単数は，約 500 語であった．これら二つの辞書間で重複した単語は，24 単語であった．

4.2.4 考察

実験結果より，為替と一般の辞書の重複割合は高くないことが分かる．そのため，ドメイン型辞書として構築されたと考えられる．

4.3 ドメイン辞書と為替レートの関係評価

4.3.1 概要

為替と一般のドメイン辞書を用いて，ツイートの極性判定を時系列のデータと為替の値動きの比較を行う．

4.3.2 実験方法

2013/11/26～2013/12/7 までのツイートの極性判定を行う．次に，否定極性と肯定極性のツイート数の割合を一日単位で算出する．そして，2013/12/1～2013/12/7 までの為替の値との比較をグラフにより評価する．

4.3.3 実験結果

1 週間の為替のツイートに対して，二つのドメイン辞書により極性判定を行い，日ごとの否定極性のツイート数をグラフにしたものを図 5 に示す．

図 6 は為替レートを表し，図 7 は為替レートの傾きを表し，図 8 は 11/26～12/2 の期間のツイートに対して，それぞれの極性の値を基に，日ごとの肯定ツイート数/(否定ツイート数+肯定ツイート数)をグラフで表したものである．

4.3.4 考察

一般のツイートを基に構築した辞書を一般辞書，為替のツイートを基に構築した辞書を為替辞書とする．

図 5 より，日にちごとの否定極性であるツイートの数を比較したところ，為替辞書を用いて否定極性と判定されたツイートの数の方が多かった．これより，ドメイン辞書は専用のドメインのツイートに対してはセンシティブに反応すると考えられる．

図 6,7,8 より，五日前の肯定極性判定されたツイートの割合と為替レートを比較すると 12/5～6 の上昇が一致していることが見て取れる．これより，五日目のデータと為替レートとの間に関係があると考えられる．

5 まとめ

本研究では，周辺文脈法を利用した辞書構築によってドメイン型辞書の構築する見通しを得ることができた．そして為替辞書を用いて，為替のツイートに対して極性判定を行いそれぞれの極性のツイート数を算出した．その結果，為替辞書の方がより感度がよいことがわかった．また，実際の日にちの五日前の肯定極性の割合と為替レートには相関があることが見受けられた．そのため，為替予測に特化したドメイン辞書構築法の見通しを得ることができた．

周辺文脈を利用した手法では，重要であるが出現回数が低い単語が評価表現辞書に登録されにくいことが挙げられる．そのため，周辺文脈法に大まかな評価表現辞書を構築し，重要であるが出現回数が低い単語を TF-IDF を用いることで改善出来るのではないかと考えられる．例として，「ジブリ」を挙げる．理由としては， 第一金曜日にアメリカの雇用統計とジブリが重なると，為替市場が暴落するようなジンクスがある．このような単語は一時的にしか現れない単語を登録することが今後の課題である．

6 謝辞

評価表現辞書を構築するにあたり，「単語感情極性対応表」の使用を許可していただきました東京工業大学の高村准教授に感謝いたします．

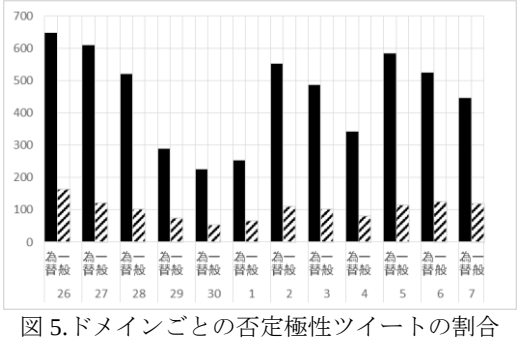

図 5.ドメインごとの否定極性ツイートの割合

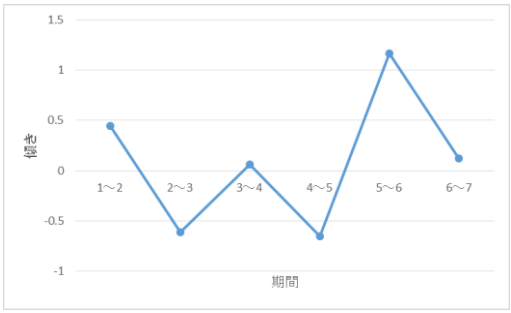

図 6.為替レート

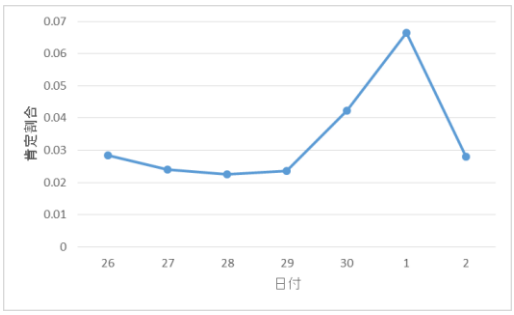

図 7.為替レートの傾き

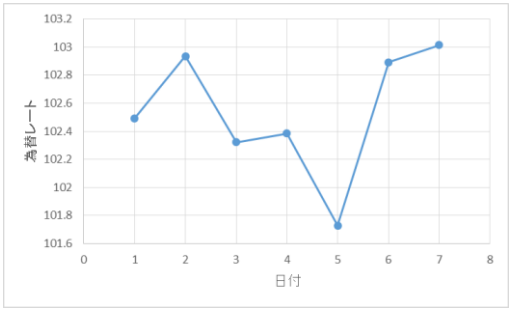

図 8.肯定極性の割合

参考文献

1) 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，Vol.46, No.2, pp.627-637 (2006).

2) 那須川哲哉，金山博：文脈一貫性を利用した極性付評価表現の語彙獲得，情報処理学会研究報告会．自然言語処理研究会報告，Vol.2004, No.73, pp.109-116 (2004).

3) 乾孝司，奥村学：テキストを対象とした評価情報の分析に関する研究動向，自然言語処理，Vol.13, No.3, pp.201-241 (2006).

4) Bollen, J., Mao, F. and Zeng, X.: Twitter mood predicts the stock market, Joumal of Coumputional Science, Vol.2, pp.1-8 (2011).

5) Mochizuki,T. and Inagaki, K.: 日本の株・外為投資家が身構える「ジブリの呪い」，THE WALL STREET JOURNAL.（オンライン），入手先（http://jp.wsj.com/news/articles/SB10001424127887323451804578642561445230042）（参照2014-06-30）．