1C5-2

ソーシャルメディアと時系列データを用いたイベント抽出及び自動ニュース生成に関する研究

Event Detection and Automatic News Generation Using Time Series Data and Social Media

丸井 淳己*1 榊 剛史*1 松尾 豊*1
Junki Marui Takeshi Sakaki Yutaka Matsuo

*1東京大学大学院 工学系研究科
Graduate School of Engineering, the University of Tokyo

Until today, many researches have been held to see events on social networks, while little attempts are held to observe events on specific time series data together with social media. We utilize Twitter and forex data to detect economic events. We define 5 event indices, 2 of them use both data and 3 of them use either of social media and forex data. We compare these indices with actual news, and we find event indices using both data perform better. We also propose 2 methods that generate news for detected events automatically from social media. For method 1 we use burst detection techniques together with news template, and for method 2 we extract phrase of explanation correspond to an economic event use of syntactic analysis. We conduct experiments on AUD/JPY rate and compare proposed methods to actual news. The accuracy of method 2 is higher than method 1.

1. はじめに

近年、経済的なグローバル化が進み、世界のどこかで起きた金融危機や災害がドミノ式に広範囲に波及する事例が増えた。1997 年にタイから始まり東南アジア・東アジアに波及したアジア通貨危機や、2007 年から 2008 年に起きた米国住宅バブル崩壊がきっかけとなって発生した世界的な金融危機などである。そのため資産を自国だけに留めておくことの不安が増大し、海外へ資産を逃避させる事例も多くなっている。国際金融の自由化が進み、企業だけでなく個人が様々な国への国際投資に関わることが増えていくと予想される中で、様々な地域の情報をすぐに手に入れる需要が増していると思われる。

しかし広範囲かつ詳細な金融情報を個人が手に入れるのは現状では難しい。Bloomberg などのサービスは信頼性も同時性も高いが、高価である。そこで最近リアルタイムに個人が発信するソーシャルメディアを用いてニュースを抽出することで広範囲な情報を手に入れることができると考えられる。Twitter は全世界で現在 1 億 4 千万のアクティブユーザ数を抱え [Twitter 12]、個々の発信だけでなく、リツイートという他人の発信した情報を拡散する仕組みが広く使われている。マスメディアより近しい存在経由で情報を手に入れさらに自分の周りにも共有することで、オンライン上での口コミ伝播が日々起こっている。

リアルタイムに発信されているソーシャルメディアのデータと経済指標の両方を取得し、同時に変化した時を捕捉することで、世界各地の経済指標を説明しているだろうニュースを手に入れることができる可能性が高い。本研究では、様々な経済指標の変動に関するニュースを自動的に配信する事を目的として Twitter と為替データを用いる。まずニュースとして取り上げる経済イベントを抽出し、その次にニュースソースと経済指標の変動を組み合わせてニュースを生成することにした。

様々なトピックに対してニュースがリアルタイムに自動生成されれば、現代のリアルタイムなニュースへの需要に応えることができる上に既存のメディアに対して大きな転換となる。本研究は、そのような研究の基礎となるものである。

連絡先: 丸井淳己, 工学系研究科技術経営戦略学専攻 博士課程, marui@ipr-ctr.t.u-tokyo.ac.jp

2. 関連研究

ここでは既存の関連研究について、経済イベント抽出に関わる部分とニュース生成に関わる部分の 2 つに分けて論じる。

2.1 既存のイベント抽出手法

トレンドを表すキーワードの検出をする手法として Burst detection がある。ある時間帯に多く出現したキーワードを取り出して、そのキーワードを含むイベントが起こったとするものである。Burst detection では文書の特徴語を取り出すために特徴語のスコアリングを TF-IDF[Salton 88] を用い、急激に TF-IDF 値が高くなるような後をその時点での特徴語として取り出すことが多い。Nagarajan らはこのような手法を用いている [Nagarajan 09]。Naaman らは Twitter からトレンドを抽出する上で TF-IDF を拡張し、ただ単に多くのドキュメントに含まれる単語を割り引くだけではなく、周期的な単語出現の影響を考慮している [Naaman 11]。また Lee らも移動窓による方法でバースト状態をより細かく解析できたと報告している [Lee 11]。しかし、本研究で行いたい経済イベントを抽出する事 Burst detection を行なっても、どの単語がその時点で盛り上がったかは分かるだけでそれ自身が経済イベントとして成立するかどうかは分からない。経済イベントは為替のみやソーシャルメディア上のみから抽出する必要はなく、2 つの情報を用いたほうが有効にイベントを抽出できるとかんがえられることから、本研究では一つの時系列変量だけからのイベント抽出ではなく、複数の時系列変量からイベントを抽出する手法を提案する。

2.2 自然言語生成

自然言語生成プロセスを E.Reiter らは 6 つのプロセスに分けて概説している [Reiter 97]。しかし個々のタスクがドメインに依存しているので自然言語生成は基本的にアプリケーション依存となる。言語を無から作り出すのは困難だが、文章からの抽出についてはよく研究されてきた。Smadja らは文生成のための連語を自動抽出する研究を行った [Smadja 90]。連語を 3 つにタイプ分けしている：一つ目は”New York Stock Exchange”のように切り離されずいつも同じ形で使われる連語、二つ目は補語関係で、似た文法構造で繰り返しでてくる動詞-副詞や動詞-目的語のようなペア、三つ目はフレーズテンプ

レートで、句のなかの情報は変わるが大枠が繰り返しでてくるものである。これらを Smadja らは株式市場のレポートを使って抽出し、さらにそれを文生成に用いている。テンプレートを抽出して、それを使って文を生成するアイデアは本研究でも使える手法である。本研究の対象の為替ニュースも株式市場のレポートと同じように定型文になっていることが多いので、本研究でもテンプレートを用いたニュース生成を行うこととした。Smadja らが研究を行った 90 年と違い、今はリアルタイムに経済データが簡単に手に入るので、テンプレートを用いたリアルタイムな文生成を行うことができる。

3. イベント抽出手法

本研究では為替とソーシャルメディアの動きからイベントを推定するが、比較のために為替とソーシャルメディア片方を用いたイベント度を表す数値 (以後イベント指数と呼ぶ) と、双方を用いたイベント指数を提示する。為替相場の場合には時間あたりの変化率が高い上に明確な通常状態がないので、ここでは最も単純に、ある時間あたりの為替相場の変化を用いる。為替の取引量はイベントが起きる度に大きくなると考えられるので、取引量のまま用いる。ソーシャルメディアの場合にも為替取引量と同じくあるイベントが起きるタイミングでツイート量が増えると考えられるので、ツイート量を用いる。双方を用いる場合には、イベントが起きるとどちらも同じタイミングで高い値を取ると予想される、為替相場の変化量とツイート数、為替相場の取引量とツイート数を掛けあわせたものをイベント指数として定義する。各イベント指数の絶対値が高い時がよりイベントである可能性が高いことを意味する。まとめると表 1 となる。

表 1: 5 つのイベント指数の定義

イベント指数	定義
FLUC	為替相場の終値から始値を引いた数
VOL	為替相場の取引量
TW	経済ニュースに関するツイート数
FLUC*TW	為替相場の変化量×経済ニュースに関するツイート数
VOL*TW	為替相場の取引量×経済ニュースに関するツイート数

3.1 評価

評価方法

精度の評価には一般的に、システムが提示したもののうちの正解の割合である Precision、全体の正解集合のうちシステムが提示したものの割合である Recall、2 つを総合した指標である F 値が使われる事が多いが、今回 Recall を定義することは難しい。しかし Precision だけを上げるために閾値を上げると、確実なイベントしか抽出しなくなってしまうためシステムが報告するイベントの量が減ってしまう。そこでニュースメディアが対象の期間で報じた記事数とイベント指数を用いて抽出したイベントの数が同定度になるように閾値を決めることとした。

実験データ

それぞれの値のばらつきを抑えるために 5 分間のツイート数と為替の 5 分間の変化値、取引量をそれぞれの標準偏差で割って正規化した。2012/4/3 から 2012/7/27 の 83 日分の為替データと Twitter から為替名を検索して収集したツイートを用いて 5 つのイベント指数を計算した。評価に用いるオンラインニュースはダイアモンド社が運営するザイ FX！[*1]を用いた。

結果

為替相場 USDJPY と EURJPY において、5 つのイベント指数を計算した。USDJPY における FLUC*TW の結果は図 1、VOL*TW の結果は図 2 の通りである。閾値を上げることで Precision は上がるが抽出されるイベント数は減少する。このうち 1000 以上のイベントを抽出する上で最も Precision の良い閾値を採用することとした。その結果が表 2 である。その結果、USDJPY,EURJPY ともに VOL*TW が一番良く、その次が FLUC*TW であり、1 変量よりも 2 変量を用いたイベント抽出手法が良いことを検証できた。

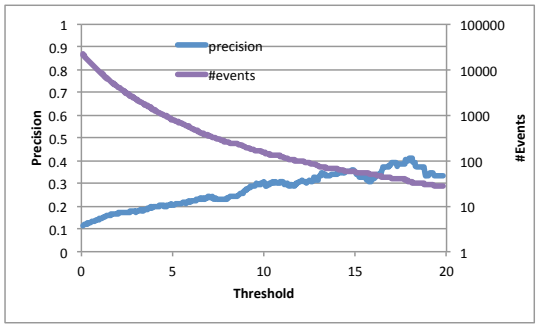

図 1: FLUC*TW:(5 分為替変動幅) × (ツイート数) を用いたイベント抽出 (USDJPY)

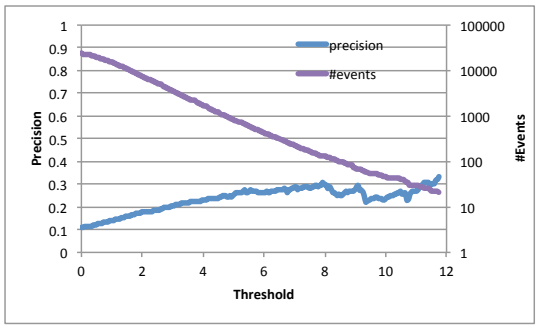

図 2: VOL*TW:(5 分為替取引高) × (ツイート数) を用いたイベント抽出 (USDJPY)

表 2: 5 つのイベント指数の結果

イベント指数	USDJPY Precision	USDJPY 閾値	EURJPY Precision	EURJPY 閾値
FLUC	16.2%	1.95	9.2%	2.1
VOL	20.3%	1.65	12.4%	1.45
TW	18.7%	4	12.9%	4
FLUC*TW	20.5%	4.4	13.2%	4.9
VOL*TW	**24.9%**	4.7	**16.3%**	4.5

3.2 議論

今回はドルとユーロを検索語句に経済イベントに関するツイートを収集したので、USDJPY・EURJPY の為替変動とツ

∗1 http://zai.diamond.jp/fx

イートが高い相関をもっているとしても不思議ではない。しかし本手法のように２変量を用いることで、ドルとユーロ以外の為替とのペアでもイベント指数を定義し、イベントを抽出することが可能であると考えられる。それを確かめるためにドル、ユーロ以外のイベントをドルで定めた閾値から抽出し、ニュースを自動生成する手法を次章で議論する。抽出されたイベントのタイミングでツイートされた情報を使って以下の２種類の手法で自動ニュース生成を行った。

4. ニュース生成手法

4.1 テンプレートによる手法

経済イベントの結果何が起きたかは為替の上げ下げを確認すればよいので、以下の様な簡単なテンプレートが作れる。

- (時間) 頃、(原因) の影響で (通貨名) 高/安 (通貨名) 安/高になりました。

原因の部分にイベントを表す特徴語を入れることとし、特徴語算出の最も一般的な方法である TF-IDF を用いる。直前のトレンドを打ち消すために DF に用いる文書の母集合をイベントが起きる前全て (DF_{all})、2 時間前 (DF_{120})、1 時間前 (DF_{60})、30 分前 (DF_{30}) の 4 つを用いた。一番高い TF-IDF 値を持つ単語をそれぞれの DF の定義毎に算出し、テンプレートを用いてニュース生成を行った。

4.2 係り受けによる方法

経済イベントを説明した表現の後に続く言葉を見つけ出せれば、フレーズとしてその部分を抜き出すことができる。為替のニュースはあるパターンで書かれることが多く、「含む」という動詞は頻出する言葉だがそれが出現する文節を抽出する。日本語は係り受けという文節の依存関係があるが、最初に抽出した文節に係る全ての文節を抽出することでフレーズも抽出できる。例えば図 3 のように「米雇用指数の強含みで」という一節が抽出できることになる。このようなニュース記事特有のパターンを抽出するために人とボットのツイートでの各単語の DF 値を比較し、ボットのツイートでの DF 値が高いものをパターンとして採用する。その結果、表 3 にある説明表現を抽出した。

なお、５分間に複数該当した場合には最長のフレーズが一番経済イベントを良く説明するフレーズであると考えられるので、最長フレーズを選択した。また、一つ前のニュースと今のニュースが同じヘッドラインから同じ表現を抜き出してしまうことがある。TF-IDF での特徴語抽出では DF の範囲を短くすることでそのような問題を防いでいたが、この場合は次に長いフレーズを選択することで回避する。

表 3: 為替ニュースにおける理由表現

単語	DF_{bot}	DF_{human}	単語	DF_{bot}	DF_{human}
受ける	0.0226	0.0059	好感	0.0034	0.0007
懸念	0.0189	0.0037	嫌気	0.0026	0.0006
優勢	0.0152	0.0032	焦点	0.0022	0.0005
強まる	0.0119	0.0024	内容	0.0020	0.0015
継続	0.0111	0.0029	失速	0.0017	0.0005
期待	0.0108	0.0060	急騰	0.0017	0.0008
急落	0.0060	0.0025	含み	0.0017	0.0004
加速	0.0043	0.0014	下支え	0.0016	0.0003
背景	0.0038	0.0012	検討	0.0013	0.0008

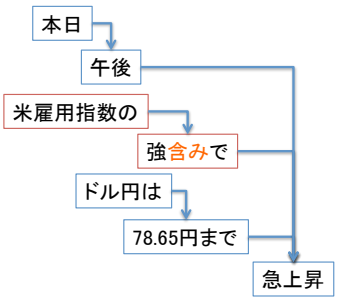

図 3: 経済ニュースの係り受けの例

4.3 評価

評価方法

２つの手法で自動的に生成したニュースと近い時間帯に配信された実際のニュースと比較することで、自動生成されたニュースの評価をすることができる。そこで、その時間帯に配信されたニュースのうちのイベントを説明するようなものから表現が抽出されているかで評価する。

評価実験

2012/7/23 朝から 7/24 夜まで、VOL*TW で判定された 56 の経済イベントについて、テンプレートによる方法 (DF のとり方で 4 種類) と係り受けによる方法でニュース生成を行った。テンプレートによる方法で各 56 ずつニュースが生成され、係り受けによる方法では 40 のニュースが生成された。係り受けによる方法では経済イベントを説明するようなフレーズがないと判定されるとニュースが生成されない。その結果、実際のニュースでその時間帯にどのように報じられているかを確認した上で生成されたニュースと比較し、正しいニュースから語やフレーズが抽出できているか判定した。その結果が図 4 である。テンプレートに依る方法では、２時間の DF を用いて TF-IDF 値をとったものが一番良い精度であった。係り受けによる方法が全体の中で一番精度が良く、60%のニュースは正しく経済ニュースから表現を抽出できていた。

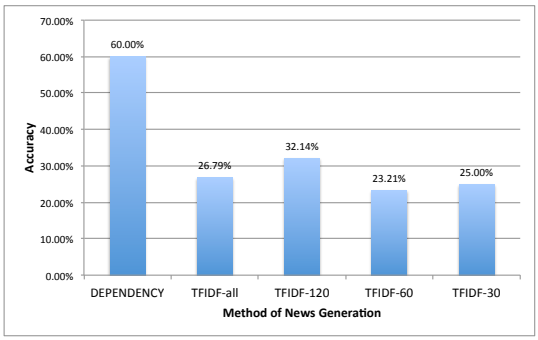

図 4: 各手法での評価結果 (7/23-24)

では本手法によりどのようにニュースが生成されたのか見てみよう。テンプレートによる方法では TF-IDF 値ランキングをそれぞれの DF で算出し、テンプレートによる手法で生成したニュースが表 4 である。上位３語に絞っているため、ニュースのキーワードが消えてしまったものもいくつかある。また Twitter で遅れて流れていた情報を掴んだり、ニュース元の別の通貨との関係では正しいが豪ドルとの関係が逆になり、実際には円安になっているにもかかわらず円高というキーワードを出してしまったケースもある。

表 4: テンプレートによる手法でのニュース生成の結果

DF の種類	ニュース文
DF_{all}	2012-07-24 20:00 頃、予想,PMI, 円の影響で豪ドル高円安になりました。
DF_{120}	2012-07-24 20:00 頃、予想,PMI, 円の影響で豪ドル高円安になりました。
DF_{60}	2012-07-24 20:00 頃、円, 予想,PMI の影響で豪ドル高円安になりました。
DF_{30}	2012-07-24 20:00 頃、円, 予想,PMI の影響で豪ドル高円安になりました。

係り受けによる手法でのニュース生成の結果の一部が表 5 である。テンプレートによる手法に比べ、自然言語文として成立していることが多いのでニュースとして読みやすい。そしてニュースから意味のひとかたまりで抽出できているためか、経済イベントを説明するニュースをうまく抽出できている。しかしテンプレートによる手法と同じく、円高と円安が混ざっているケースもある。また、実際のニュースでは直接豪ドルについては言及されてない場合でも豪ドルについてそれらしい説明をするケースもある。例えば 7/23 21:20 の「介入警戒感が下支えし、」というフレーズは米ドルのニュースから抽出したものだが、おそらく豪ドルでも成り立つニュースと思われる。

表 5: 係り受けによる手法でのニュース生成の結果

日付	ニュース文
2012-07-23 21:20	介入警戒感が下支えし、2012-07-23 21:20 頃豪ドル高円安になりました。
2012-07-24 17:15	豪ドル反発、中国ＰＭＩを好感して、2012-07-24 17:15 頃豪ドル高円安になりました。
2012-07-24 17:45	引き続き豪ドル反発、中国ＰＭＩを好感して、2012-07-24 17:45 頃豪ドル高円安になりました。
2012-07-24 18:20	円高続く展開で、2012-07-24 18:20 頃豪ドル安円高になりました。

以上、実際のニュースと比較し精度を計算した後に実際に生成されたニュースを見ながら２つの手法を評価した。テンプレートによるニュース生成では DF のとり方を幾つか試行してニュースと比較した制度を取ることで経済イベントが一番良く取れる DF の範囲が２時間であることが分かった。一番精度の良かったものは係り受けによる手法で、ニュースのヘッドラインからフレーズを切り出して為替の文に足しているため、自然な文が生成できた。

5. まとめ

本研究では為替変動のイベントを為替データに加えてソーシャルメディアを用いることで検出する方法について議論し、さらにその出来事についてのニュースを自動生成する手法を提案した。

為替変動のイベント抽出では、２変量のイベントを用いる手法で経済イベントである度合いの指標を提案した。抽出するイベントの数が Precision よりも重要であるため、既存のオンラインメディアの記事数と同等な一定数以上のイベントを抽出させ、その上で一番良い Precision をとるような閾値を用いた。その結果、１変量でのイベント抽出よりも精度よくイベントを抽出できることが分かった。

自動ニュース生成タスクでは、米ドルのツイートデータと豪ドルの為替データを用いてイベントを抽出し、豪ドル・日本円市場でのイベントを抽出した。イベントが起きている時の特徴語とテンプレートを用いてニュース文生成する手法と、表現パターンと係り受けに着目して経済ニュースからイベントが起きた原因の記述を抽出する手法の２つを提案した。テンプレートと特徴語によるニュース生成は DF 値の計算期間を変えながら実験し、係り受けによる手法と比べたところ、係り受けによるニュース生成手法が一番精度良く、また自然な文が生成できることが分かった。

これから一層国同士の垣根が低くなっていき、個人においても国際経済の連関の影響を受けると考えられ一層の情報速度の需要が高まる中で、新聞やテレビといった紙面・電波・時間などで多くの制約を受けるマスメディアは試練に立たされている。本研究において行ったイベント抽出や自動ニュース文生成は、そのような問題を扱う上で基礎となる技術である。

参考文献

[Lee 11] Lee, C.-H., Wu, C.-H., and Chien, T.-F.: BursT: a dynamic term weighting scheme for mining microblogging messages, in *Proceedings of the 8th international conference on Advances in neural networks - Volume Part III*, pp. 548–557, Springer-Verlag (2011)

[Naaman 11] Naaman, M., Becker, H., and Gravano, L.: Hip and trendy: Characterizing emerging trends on Twitter, *Journal of the American Society for Information Science and Technology*, Vol. 62, No. 5, pp. 902–918 (2011)

[Nagarajan 09] Nagarajan, M., Gomadam, K., Sheth, A. P., Ranabahu, A., Mutharaju, R., and Jadhav, A.: Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences, in *Proceedings of the 10th International Conference on Web Information Systems Engineering*, pp. 539–553 (2009)

[Reiter 97] Reiter, E. and Dale, R.: Building applied natural language generation systems, *Nat. Lang. Eng.*, Vol. 3, No. 1, pp. 57–87 (1997)

[Salton 88] Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, in *INFORMATION PROCESSING AND MANAGEMENT*, pp. 513–523 (1988)

[Smadja 90] Smadja, F. A. and McKeown, K. R.: Automatically extracting and representing collocations for language generation, in *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pp. 252–259, Association for Computational Linguistics (1990)

[Twitter 12] Twitter, : Twitter turns six, *Twitter Blog* (2012)