3H1-OS12a-3

英文経済レポートのテキストマイニングと市場分析

Analysis of Financial Markets' Fluctuation by English Textual Information

余野　京登[*1]　和泉 潔[*1*2]　後藤 卓[*3]　松井 藤五郎[*4]　陳 ユ[*1]
Kyoto Yono　Kiyoshi Izumi　Takashi Goto　Tohgoroh Matsui　Chen Yu

[*1]東京大学大学院　[*2]JST さきがけ　[*3]三菱東京 UFJ 銀行
The University of Tokyo　PRESTO, JST　The Bank of Tokyo-Mitsubishi UFJ, Ltd.

[*4]とうごろう機械学習研究所
Tohgoroh Machine Learning Research Institute

In this study, we applied the newly developed text-mining methods to English texts for the long-term market analyses. We analyzed monthly price data of foreign financial markets, in particular, the interest swap markets. Several extensions of the original method were suggested in order to extract English feature vectors from minutes of the monetary policy committee of The Bank of England. Trends of interest rates were estimated by using the regression analysis with the feature vectors. As a result, determination coefficients were found around 75%, and market trends were explained well. Using the predicted interest rates, we also simulated several implementation tests, which demonstrate the effectiveness of our extensions of the original method to English texts.

1. はじめに

近年，データマイニング技術を用いて，市場動向を分析する研究が多く行われている. ニューラルネットワークや遺伝的アルゴリズム等を数値データに用いて市場分析を行う研究や，ニュースなどのテキストデータの市場への影響を推測する研究がある [電気 02]. また，日本銀行の金融経済月報を題材に，テキストマイニング技術を用いて，市場動向を予測する研究もあり，一定の成果を挙げている [和泉 09]. 本研究では，今まで行われなかった英文テキストを用いたテキストマイニングを行い，英国のスワップ金利に対して市場分析を行った.

2. テキストデータによる長期市場分析手法

従来のテキストデータによる長期市場分析の研究では，日本語のテキストデータと分析の対象になる市場の時系列データを用いていた. 本研究では，テキストデータとして英国の中央銀行であるイングランド銀行（Bank of England，BOE）の金融政策委員会（Monetary Policy Committee，MPC）が発行している議事録 [イン] を，時系列データとして英国のスワップ金利を対象に選んだ. 金融政策委員会は，毎月上旬に 2 日連続で開催され，政策金利変更は，2 日目の正午に発表され，市場の注目を集める. 議事録はその 2 週間後に 10 ページ前後の分量で公表される. 分析対象として，イングランド銀行の金融政策委員会の議事録を選んだ理由は，月次のレポートであり，文章の段落構造がある程度決まっており，時系列分析を行いやすいからである. また，イングランド銀行の金融政策委員会の議事録は，金融関係者が常に注目していおり，市場への影響力が大きいと考えられる.

2.1 従来のテキストマイニング手法

和泉らの研究 [和泉 09] では，日本語のテキストデータを用いた長期的な市場分析の手法を構築した. 我々は，まず 3 つのステップからなる従来の手法をそのまま英文テキストの市場分析に用いた.

2.1.1 共起関係に基づく主要単語の抽出と可視化

具体的にはまず，英語の品詞タグを付与するソフトである GoTagger [GoTa] による単語を原形にし，名詞・動詞・形容詞等の品詞を記した. そして，出現頻度順に単語を抽出した．次に，各月のテキストデータに KeyGraph [大澤 06] を適用し，共起関係を解析した．Jaccard 係数（= p(A and B)/p(A or B); ただし A,B は抽出した単語）を段落毎に適用し，段落毎に同時に出現する単語と単語を繋ぎ，共起グラフを作成する．その後，単結合（A,B 間のみの結合部分）を切断し，結合による「島」を作成する．そして，各単語間の共起度に基づき，上位順に「橋」を作成する．これらの操作によって，各月のテキストデータから主要単語をノードとするネットワークを構築した.

2.1.2 主成分分析による単語のグループ化

KeyGraph で作成したネットワークに出現した単語のパターン (単語を月毎の出現状況に従いパターン分類したもの) に対し主成分分析を実施し，30 個の合成変数（主成分）にまとめる．各月の 30 個の主成分スコアを，分析対象期間について時系列順に並べることによって，30 次元の時系列データが作成される．これが分析対象期間のテキストデータの特徴の時間的変化を表していると考える．主成分分析の際には，単語に関して品詞を区別せずに分析を実施する．ここで注意してほしいのは，ここまで市場データは全く用いず，純粋に単語の出現パターンのみの分析を行っていることである.

2.1.3 重回帰分析による市場データの動向分析

最後に，各主成分スコアの毎月の動きから月次での市場金利の動きを解析する．具体的には， 前節で述べた 30 個の主成分スコアの時系列データを説明変数として，月次の市場データを被説明変数とする重回帰分析を行う．分析対象期間内の金利の動きを推定するだけでなく，分析対象外のテキストデータを与えれば外挿予測を行うこともできる．この外挿予測は，月中に発表される金融政策委員会の議事録から，約 2 週間後の月末のスワップ金利を推定することになる.

: 余野 京登，東京大学大学院　工学系研究科 システム創成学専攻 大橋研究室 修士 1 年，〒 113–8656 東京都文京区本郷 7–3–1， yono@crimson.q.t.u-tokyo.ac.jp

2.1.4 外挿予測力の運用テスト

英文テキストに対する従来の手法の有効性を確かめるために，外挿予測力の運用テストを行った. 実際に提案テキストマイニング手法が使われる場面と同様に，直近のデータまでを訓練データとして毎月新しいデータを追加して新たに分析を更新した場合の外挿予測力の比較を行った．最初に 1998 年 1 月から 2007 年 9 月までのテキストデータと市場データを訓練データとして回帰式を推定し，その式に 2007 年 10 月のテキストデータを入力して，2007 年 10 月末の金利を外挿予測によって推定した．次に 2007 年 10 月のテキストデータと市場データを訓練データに追加して，1998 年 1 月から 2007 年 10 月までのテキストデータと市場データを訓練データとして回帰式を推定し，その式に 2007 年 11 月のテキストデータを入力して，2007 年 11 月末の金利を外挿予測によって推定した．そして，以下のルールで，予測金利を用いて，運用テストを行った．

1. 取引ルール 1 (金利水準の比較): 予測金利と発表時の金利を比較して，予測金利が高ければ 1 単位の資本を買う．低ければ，1 単位の金利を売る．

2. 取引ルール 2 (金利変動の比較): 予測金利の前月からの変動幅と，発表時の金利の前月末金利からの変動幅を比較して，予測金利の変動幅が高ければ，1 単位の資本を買う．低ければ，1 単位の資本を売る．

運用テストの結果は以下の図 1 のようになった. ルール 1，2 共に平均年率リターンがマイナスであった. つまり, 日本語テキストに用いられた従来の手法は, 英文テキストの分析には使えなったのである.

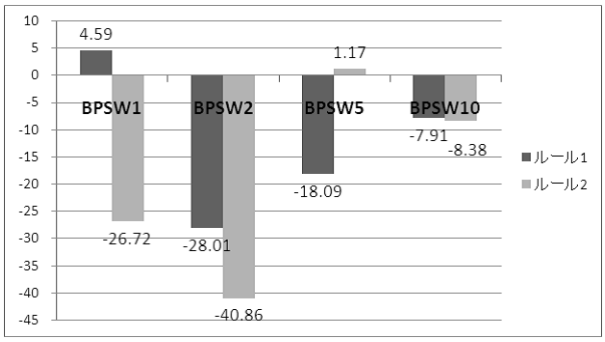

図 1: 従来の手法による運用テスト結果．2008 年 1 月〜2008 年 12 月までの各月末金利に関して外挿予測を行った．灰色が取引ルール 1(金利水準の比較) での平均年率リターン (%). 黒色が取引ルール 2(金利変動の比較) での平均年率リターン (%).

3. 英文テキストマイニングのための拡張

上記の従来の手法で英文テキストマイニングがうまくいかない理由として, 日本語にはない英語の特徴が考えられる. 例えば, 日本語で一語の単語でも英語では 2 語以上の連語となることや，同じ意味でも英語は多様性の富んだ表現を使うことが挙げられる. そこで, 我々は英文テキストマイニングのために, 従来のテキストマイニングの長期的な市場分析手法に対して, 新たに二つの拡張を施した.

3.1 手法の拡張

本研究では従来の手法に「連語抽出」と「1 回出現単語の削除」という二つの工程を加えた.

連語抽出

2.1.1 節で述べた「共起関係に基づく重要単語の抽出と可視化」において, 連語の抽出機能を加えた. 名詞＋名詞や, 形容詞＋名詞のなど順番で出てきた単語を 1 語の連語として変換し, 抽出した．この機能を加えることで, 2 語に分けられていた単語が 1 語として捉えることができ，より正確な単語抽出が可能となる.

1 回出現単語の削除

2.1.2 節で述べた「主成分分析による単語のグループ化」において, 主成分分析を施す前に, 重要単語のうち, 一ヶ月のみに出現した単語を削除した．これは, 主成分分析時に一ヶ月のみに出現した重要単語がその月の特徴を表しすぎてしまい, 2.1.3 節で述べた回帰分析の外挿予測に利用しにくいと考えたためである.

3.2 拡張した手法による月次市場分析の結果

最初に，KeyGraph アルゴリズムと主成分分析を用いて，30 次元の特徴量を金融政策委員会の議事録のテキストデータから抽出した．抽出された主成分には大きく分けて 2 つのタイプがあった．一つは市場の動きに関する特徴量である．例えば，1 番目の主成分は,「bank rate」「unchanged」「hold」といった動きを表す単語から構成されていた．他にも，3 番目の主成分は,「weaken」「increase」「high」「rise」といった単語の寄与が高かった．もう一つのタイプは，経済のファンダメンタルズに関する特徴量である．例えば，2 番目の主成分は「inflation」「housing market」といった実態に関する単語から構成されていた．他にも，12 番目の主成分は,「inflation target」「policy」「official data」といった金融政策に関する単語の寄与が高かった．次に，これらの 30 次元の特徴量の時系列データを用いて，各市場データの回帰分析を行った．回帰分析の際に，AIC 基準を用いたステップワイズ選択により，説明変数の絞り込みを行った．英国のスワップ金利の 1 年物，2 年物，5 年物，10 年物について，23-25 個の説明変数による回帰式を得ることができた．決定係数 R2 をみると，サンプルデータについて十分な説明力を持つことがわかった．R2= 73.45%(SWAP1 年物), 75.15% (SWAP 2 年物)，75.67% (SWAP 5 年物)， 74.66% (SWAP 10 年物)．1998 年 1 月から 2007 年 12 月までの過去 10 年間の訓練データを用いた回帰式に，2008 年 1 月から 12 月までのテキストデータを入力して，各金融市場における外挿予測テストを行った．図 1a-d に，推定されたパスと実際のパスを示す．外挿期間における推定パスと実際のパスを比較すると，英国スワップ金利 10 年物がトレンドの方向性 (上昇と下降) および金利の全体的な水準が一致しており，最も精度の高い外挿予測を行っていた．他スワップ金利に関しても方向性の一致が見られた.

3.3 拡張した手法による外挿予測力の運用テスト

2.1.4 節で解説した運用テストを行った (図 3). 従来の手法と運用結果（図 1）を比較すると, すべての金利 SWAP で運用成績が向上した. ルール 1 は, 平均で 29%近くリターンが上昇し, ルール 2 においても 21%のリターンが上昇した. このことから「連語を含む単語の抽出」と「ある 1 つ月のテキストのみで抽出される重要単語の削除」という拡張した手法の方が, 外挿入予測力が強いことが示させた．また, 外挿期間において,

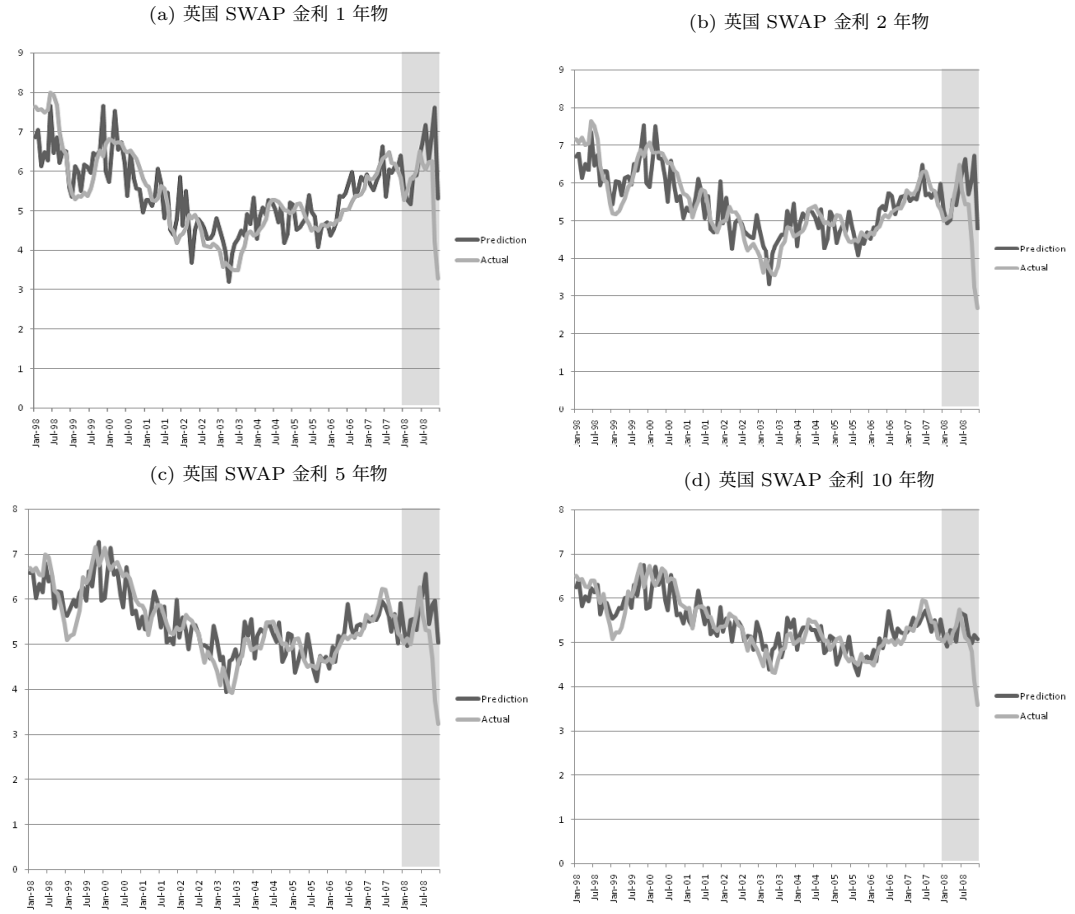

図 2: 各市場トレンドの推定．訓練期間: 1998 年 1 月〜2007 年 12 月，外挿期間: 2008 年 1 月〜12 月．

主成分 1		主成分 2		主成分 3		主成分 4		主成分 5	
bank rate	0.712	euro area	-0.462	assume	-0.436	volatility	0.449	inflation report	-0.484
credit	0.676	little	-0.416	generate	-0.380	business investment	0.439	large	-0.464
exclude	-0.639	assume	0.406	policy	-0.349	output	0.436	expect	0.381
unchanged	-0.621	minute	-0.398	weaken	0.341	warrant	0.436	governor	0.349
hold	0.562	reduce	0.359	increase	-0.340	less	0.410	rise	0.344
rate	-0.525	central projection	0.345	bank	0.336	market interest rate	0.406	development	-0.343
financial market	0.523	united kingdom	-0.344	view	-0.329	global	-0.403	prove	-0.341
euro	-0.494	inflation	0.341	export	-0.328	temporary	-0.403	remain	0.340
minute	0.487	housing market	-0.336	high	-0.327	mark	0.396	peak	-0.323
revise	-0.475	view	0.334	rise	-0.317	picture	0.385	global	-0.315

主成分 6		主成分 7		主成分 8		主成分 9		主成分 10	
global	-0.417	set	-0.346	increase	0.426	uncertainty	0.477	household spend	0.350
temporary	-0.417	growth rate	0.344	deteriorate	0.423	need	-0.365	near term	0.346
inflation report	-0.405	depreciation	-0.339	margin	0.408	iraq	0.341	shift	-0.342
historical average	-0.383	contrast	-0.327	condition	-0.406	economy	-0.325	first	0.341
light	-0.381	government	-0.323	sub	-0.375	war	0.324	clear	-0.321
positive	-0.348	indicator	-0.318	subdue	0.374	financial	0.323	determinant	0.318
oil price	0.326	flat	-0.316	inflation expectation	0.359	contribution	-0.322	maintain	0.314
example	-0.324	rise	-0.316	supply	0.354	underlie	-0.301	go	-0.308
expect	-0.323	previous quarter	-0.311	lead	0.334	risk	0.299	consumption	0.306
recover	-0.312	expect	-0.305	cut	-0.326	previous quarter	-0.296	bank 's	-0.302

主成分 11		主成分 12		主成分 13		主成分 14		主成分 15	
consider	-0.417	reduce	0.398	turn	0.338	february inflation	-0.351	asian crisis	0.398
mark	0.361	confidence	0.341	household	0.315	go	0.330	asia	0.398
help	-0.354	judge	0.319	cost	-0.312	central projection	-0.313	balance	-0.335
flat	-0.344	contrast	0.311	household spend	-0.305	line	-0.310	demand growth	0.314
set	-0.324	maintain	0.305	movement	0.303	account	0.304	export	0.303
boost	-0.313	oil	0.301	inflationary pressure	-0.287	recovery	-0.302	country	0.301
underlie	-0.306	particular	0.289	bank 's repo rate	-0.274	business survey	0.296	help	0.298
consumer spend	0.303	firm	0.272	domestic	0.269	agent	0.293	influence	0.293
weakness	-0.291	bank 's repo rate	0.264	risk	-0.269	regional	0.289	sterling 's	-0.278
business	-0.286	november inflation	-0.263	little news	-0.268	ons	-0.284	prospect	-0.276

表 1: 1997 年 1 月から 2007 年 12 月までのテキストから抽出された主成分 (上位 15 個) と， 各主成分で負荷量の絶対値が上位 10 個のキーワード.

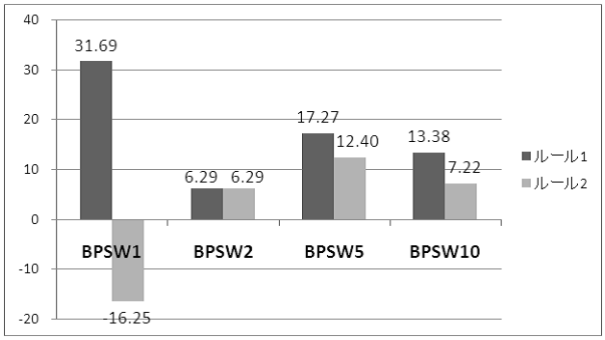

図 3: 拡張した手法による運用テスト結果. 2008 年 1 月〜2008 年 12 月までの各月末金利に関して外挿予測を行った. 灰色が取引ルール 1(金利水準の比較) での平均年率リターン (%). 黒色が取引ルール 2(金利変動の比較) での平均年率リターン (%).

価格変動の予測値と実際の金利変動と一致していたかどうかの割合（正答率）を拡張した手法と従来の手法で比較した.(図 4) その結果, ルール 1 とルール 2 共にすべての英国 SWAP 金利で正答率が上昇した. このことからも, 拡張した手法の方が外挿において強い予測力を有していることがわかる.

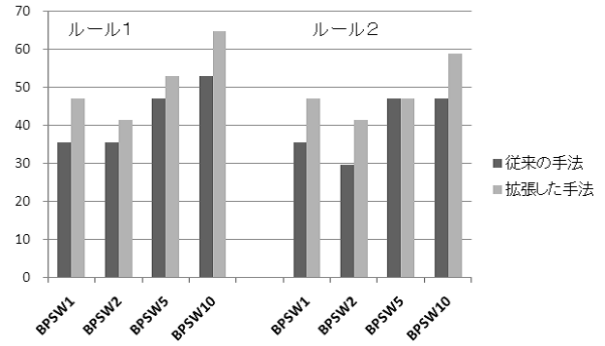

図 4: 外層期間の変動予測の正答率の変化

4. まとめ

本研究では, 英文テキストデータを用いた分析にも適応できるように従来のテキストデータを用いた長期的な市場分析に二つの拡張を加えた. 本手法により, 英国 SWAP 金利の分析を行い, 運用テストを行った結果, 従来の手法では, 英文テキストに対して得られなかった高いリターンを得ることができた. また, 訓練データの決定係数に関しても, 75 %前後と高く, 訓練データについても十分な説明力を持つことがわかった. 本研究では, 英国中央銀行の金融政策委員会の議事録を用いたが, 今後は, 連邦準備制度や欧州中央銀行など他の国の中央銀行英文テキストを試みる予定である.

参考文献

[GoTa] GoTagger ホームページ：http://uluru.lang.osaka-u.ac.jp/~k-goto/use_gotagg% er.html

[イン] イングランド銀行：金融政策委員会が発行している議事録：http://www.bankofe% ngland.co.uk/publications/minutes/mpc/index.htm

[大澤 06] 大澤 幸生：チャンス発見のデータ分析―モデル化+可視化+コミュニケーション→シナリオ創発, 東京電機大学出版局 (2006)

[電気 02] 電気学会（編）：学習とそのアルゴリズム, 第 6 章, 森北出版 (2002)

[和泉 09] 和泉 潔, 後藤 卓, 松井 藤五郎：テキスト情報による金融市場変動の要因分析, 2009 年度人工知能学会全国大会 (2009)