石田基広・金明哲編著『コーパスとテキストマイニング』2.金融テキストマイニングの紹介

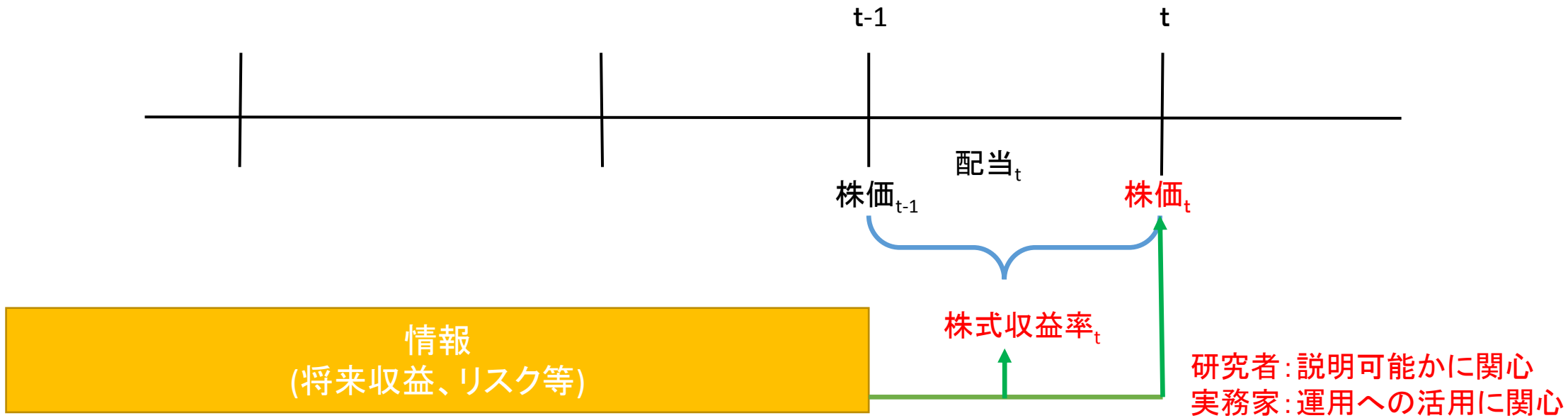
内田交謹（経済学研究院）

異分野融合テキストマイニング研究会第3回研究会

2015年12月17日(木)

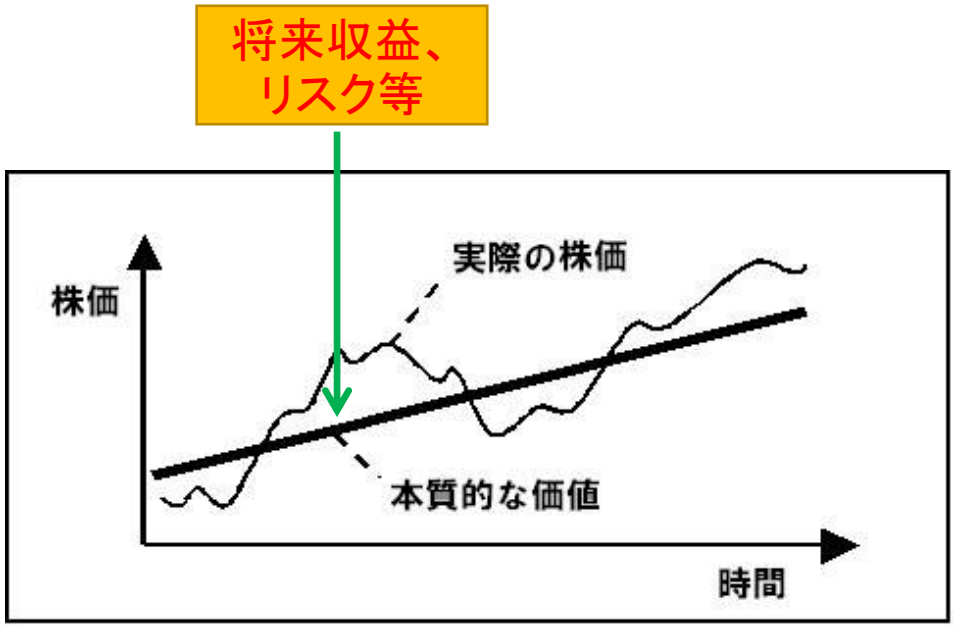
2.1 金融テキストマイニングの背景と目的

- ファイナンス (インベストメント) の主要テーマ
  - 資産価格(主に株価)、資産収益率(主に株式収益率)の説明
    - 株式収益率$_t$ = (株価$_t$ – 株価$_{t-1}$ + 配当$_t$) / (株価$_{t-1}$)
    - 研究者目線と実務目線:説明モデル構築と資産運用への活用



2.1 金融テキストマイニングの背景と目的

- 資産運用スタイル
  - テクニカル分析：将来の株価を予測するのに、過去の株価変動、株式収益率を用いる
  - ファンダメンタルズ分析：将来の株価を予測するのに、企業収益、金利等の基礎的情報を用いて将来収益を予測する。



2.1 金融テキストマイニングの背景と目的

- 余談
  - 効率的市場仮説 (1980年代頃までの主流の考え方)
    - 株式市場は新しいファンダメンタル情報を瞬時に株価に織り込むため、ファンダメンタルズ情報を用いて株式運用を行っても超過収益をあげることはできない。
    - 株価はランダムウォークするため、テクニカル分析も意味がない。
  - アノマリー (1980年代以降、その存在が確たるものと認識される)
    - 小型株効果、バリュー株効果、モメンタム、リバーサル、流動性

2.1 金融テキストマイニングの背景

- 株価、株式収益率を説明する多くの分析は数値情報を扱う
  - テクニカル分析&アノマリー：過去の株式収益率
  - ファンダメンタルズ分析：リスク、利益、配当
  - アノマリー：企業規模、簿価時価比率

- 金融テキスト分析の背景①：ファンダメンタルズ≠数値
  - 新製品発表→将来の業績変化予測→株価変化
  - M&A →将来の業績変化予測→株価変化
  - 有能な経営者の辞任、死去→将来の業績低下予測→株価下落

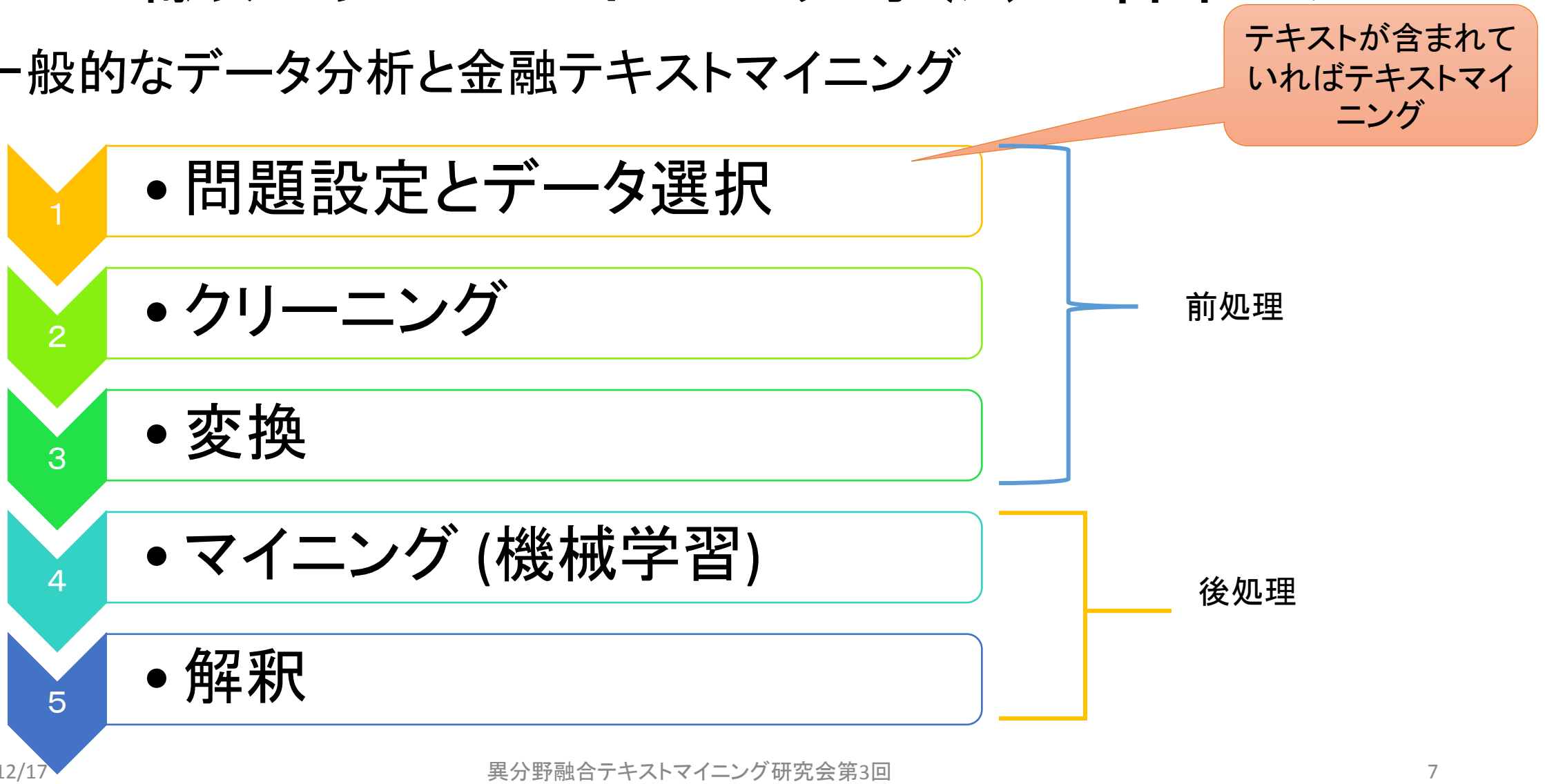
2.1 金融テキストマイニングの背景

- 金融テキスト分析の背景②：運用現場でテキスト利用
  - 新聞記事などテキスト情報を用いた運用
  - 経営計画を考慮した運用

- 背景①、②のいずれの意味でも、テキストを用いた株価の分析は古くからある。
- しかし最近では、コンピューターでテキストから内容を抽出して分析に用いることや高速処理が多くなってきている。
  - 東京証券取引所アローヘッドの採用
  - テキストの自動分析による自動注文

2.2 金融テキストマイニング手法の枠組み

- 一般的なデータ分析と金融テキストマイニング



2.2 金融テキストマイニング手法の枠組み

- 実際のデータマイニングの処理の流れ

1. 問題設定
2. 機械学習アルゴリズムの選択
3. データ選択
4. 前処理方法の選択
5. モデルの評価

- 以下、各処理について説明

2.2 金融テキストマイニング手法の枠組み

- データマイニング
  - 記述的データマイニング
    - データの特徴を明らかにする。
    - (例) 2008年のリーマンショック前の新聞報道内容
  - 予測的データマイニング
    - 特徴が分かっているデータから未知のデータの特徴を予測するモデルの構築
    - (例)朝刊の新聞記事を分析してその日の株価終値を予測
- 以下では、予測データマイニングに焦点を当て、テキストマイニング手法の枠組みを説明

2.2 金融テキストマイニング手法の枠組み
2.2.1 問題設定

- 何を予測するのか決める
  - 株価、ボラティリティ、出来高、金利、為替レートなど
- どれだけ後の将来を予測するのか決める

2.2 金融テキストマイニング手法の枠組み
2.2.2 機械学習アルゴリズムの選択

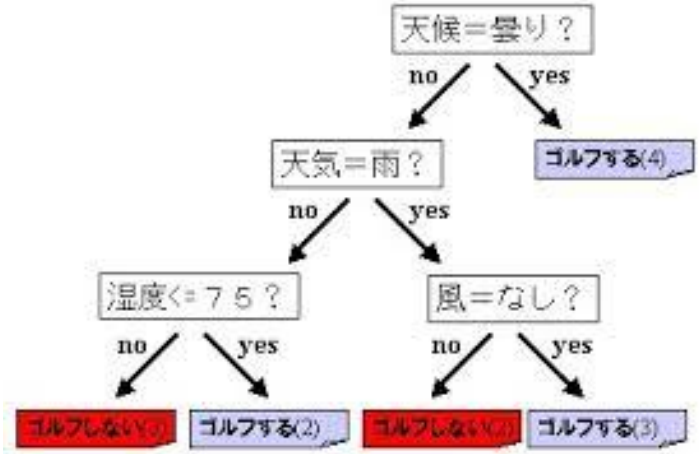
- Bag-of-words
  - どの語が何回出現したかでテキストを表現
- 教師付き学習と教師なし学習
  - 説明、予測したいものの答えの値を教師データと呼び、答えの値がある時は教師付き学習、ない時は教師なし学習を用いる。
    - (例) テキストから株価を予測するときは教師付き学習を用い、テキストの内容によって東証一部の銘柄を三つのグループに分ける時は教師なし学習を用いる

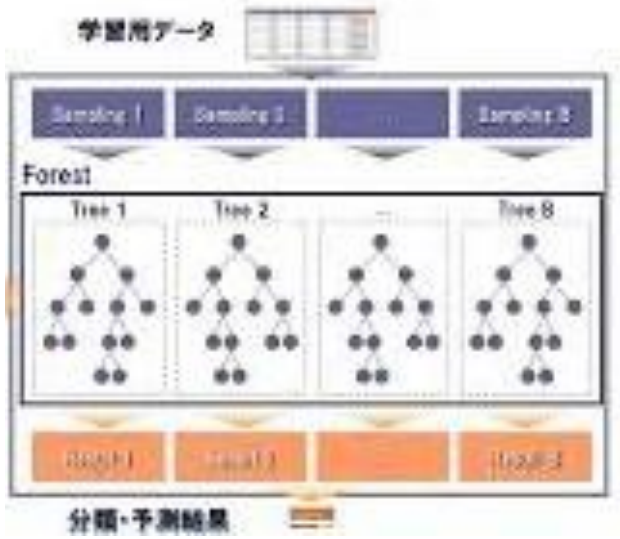
2.2 金融テキストマイニング手法の枠組み
2.2.2 機械学習アルゴリズムの選択

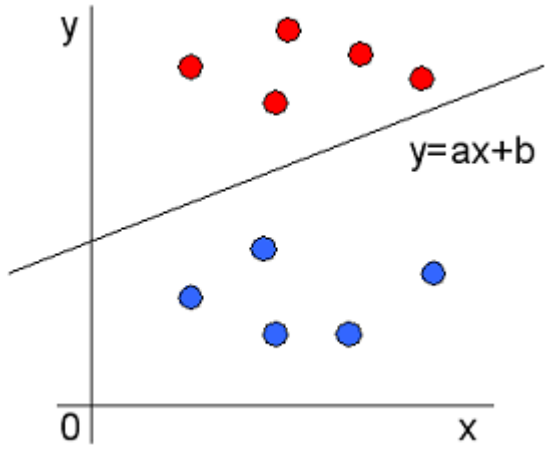
- 教師付き学習
  - 目的変数 (従属変数、被説明変数) : 予測する値
  - 説明変数(独立変数) : 予測に用いる値
  - 分類 (目的変数がカテゴリー; 上昇／下降，安定／不安定等)
    - 決定木学習、ナイーブベイズ、サポートベクター分類（SVC）、ランダムフォレスト等
  - 回帰 (目的変数が数値)
    - 重回帰、Lasso、Ridge、サポートベクター回帰（SVR）等
- 教師なし学習
  - クラスタリング : 階層的クラスタリング、K平均法、スペクトラルクラスタリング、混合正規分布法

2.2 金融テキストマイニング手法の枠組み
2.2.2 機械学習アルゴリズムの選択

- 分類 (教師付き学習 ; カテゴリー目的変数)
  - 決定木 :
    - シンプルな木構造の If-then ルールの学習
  - ナイーブベイズ
    - テキストに含まれる語が独立に出現すると仮定し、それぞれの語の出現確率から同時出現確率を求めてテキストを分類
  - SVC :
    - 高次元空間上で正事例と負事例を分割する超平面を学習する方法
  - ランダムフォレスト :
    - 複数の決定木をランダムに作成し、統合して予測







2.2 金融テキストマイニング手法の枠組み
2.2.2 機械学習アルゴリズムの選択

- 回帰(教師付き学習 ; 目的変数が数値)
  - 重回帰 :
    - 目的変数を説明変数の線形関数で表現する (パラメーター推定)
  - Lasso, Ridge :
    - 重回帰分析に正則化(L1正則化、L2正則化)を行い、モデルの複雑さにペナルティを課すことで、過学習を防ぐ
  - SVR :
    - 高次元空間上で回帰式を学習する方法

2.2 金融テキストマイニング手法の枠組み
2.2.2 機械学習アルゴリズムの選択

- クラスタリング(教師なし学習)
  - 階層的クラスタリング:
    - 最も似ている事例同士、事例とクラスタ、クラスタ同士を一つのクラスタにまとめることを繰り返すことで事例とクラスタの階層関係を構築。
  - K平均法:
    - ①K個のクラスタを生成し、②それぞれの事例を最も近いクラスタに振り分け、③それぞれのクラスタに属する事例の平均値を各クラスタの中心とし、②、③を繰り返すことで、事例をK個のグループに分ける。
  - スペクトラルクラスタリング:
    - 次元圧縮を行ってからK平均法を行う
  - 混合正規分布法:
    - 事例が複数の正規分布から確率的に発生していると仮定し、正規分布のパラメータを推定

2.2 金融テキストマイニング手法の枠組み
2.2.3 データの選択

- 分析対象となりえるテキスト例
  - ① 各国中央銀行の報告書
  - ② ニュース
  - ③ アナリスト・レポート
  - ④ ツイッター
  - ⑤ ブログ
  - ⑥ 電子掲示板
- 留意点
  - ①：内容が確かで数が少なく、テキストのサイズが大きい
  - ②・③：内容が確かで数が比較的多く、テキストのサイズは中程度
  - ④～⑥：内容が不確かで数が非常に多く、テキストのサイズが小さい

2.2 金融テキストマイニング手法の枠組み
2.2.4 前処理の選択

- データクリーニング
  - ゴミやムダな情報の削除。
    - スパム記事；同じ記事の重複；文字化け等
- ベクトルデータへの変換
  - 機械学習への入力に、ベクトルデータへの変換が必要
  - Bag-of-words の要素となっている語(特徴語)のそれぞれに次元を割り当て、テキストデータを特徴語の特徴量を要素とするベクトルデータとして表す
  - 文章を形態素に分割
    - MeCab:「東京の金が反発」→「東京-名詞／の-助詞／金−名詞／が−助詞／反発−名詞」
    - 連続した名詞を連結するには、TermExtract などの専門用語抽出システムを利用

2.2 金融テキストマイニング手法の枠組み
2.2.4 前処理の選択

- ベクトルデータへの変換:
  - ベクトルデータの次元数が大きいと機械学習がうまくできない
  - 不要な特徴語は取り除く:(例) 助詞
  - 名詞、動詞、形容詞など必要な品詞や「よい」「悪い」などの評価表現を取り出す
  - 名寄せ:言い換え表現の統一
  - それでも特徴語の数が多ければ、重要な特徴語だけ選択して利用
    - tf-idf (その語のテキスト出現頻度×全テキスト数／その語を含むテキスト数)
    - AIC; BIC
  - 各特徴語について特徴量を求め、ベクトルデータに変換
    - 含まれる=1、含まれない=0
    - 出現頻度;tf-idf

2.2 金融テキストマイニング手法の枠組み
2.2.5 モデルの評価(予測精度の推定)

- 予測精度(予測誤差)の推定：
  - モデルに説明変数を入れ、予測値を計算。
  - 予測値と実際の目的変数の値を比較して、予測精度 (予測誤差) を推定
- 金融テキストマイニングにおける予測精度推定
  - データが時系列であるため、評価データよりも過去の訓練データから学習したモデルで予測精度を推定。

2.3 金融テキストマイニングの研究事例

2.2.5 分析目的と利用テキストから見た研究事例の分類

ロイター／Bloomberg／日経Quick等

		ツイッター／ブログ	オンラインのニュース・新聞記事	金融機関等発行の経済リポート
テキストの性質	量	1Gバイト以上／日	数百Kバイト／日	数十Kバイト／月
	書き手	1億人以上，一般	数百人，記者	数十人，専門家
	内容	多様	ある程度限定	経済専門
	様式	不定	ある程度固定	固定
問題設定	予測対象	市場平均株価 (個別銘柄??)	(市場平均株価??) 個別銘柄	市場平均株価，国債市場、(個別銘柄??)
	予測機関長	1日から数日間	数十分間から数時間	2週間から1か月間

膨大な量／多様な内容と書き手

専門的な内容／統一的テキスト

2.3金融テキストマイニングの研究事例

2.3.2ツイッターに現れる心理状態と株価平均

- Bollen et al. (2011)

2008年2月28日～11月28日
270万人のユーザーによるツイッター9,853,498個のテキスト

“I feel”, “I am feeling”, “I’m feeling”, “I don’t feel”, “I‘m”, “I am”, “makes me” を含むテキストを抽出し、心理状態を表す表現を分析

平穏・警戒・確信・活気・善意・幸福の6次元の尺度のスコアを計算

米国ダウ平均株価指数

①Granger 因果性検定
→「平穏」が2～5日後の平均株価に影響

②1日前から3日前までの「平穏」と平均株価で翌日の平均株価を予測：
訓練期間2008年2月28日～11月28日
テスト期間2008年12月1日～19日
→翌日の株価の騰落を86.7%の精度で予測。ただし、テキスト情報なしでも73.3%の予測精度

2.3金融テキストマイニングの研究事例

2.3.3　リアルタイム配信ニュースと短期市場変動

- Survey by Mittermayer and Knolmayer (2006)

ニュース記事テキスト(ロイター，日経Quickなど)

重要そうな単語やカテゴリーまたは単語の組み合わせの頻度、tf-idf値を計算し、テキストの特徴ベクトルとする

数時間後の価格トレンド(上昇／下降／横這い)やボラティリティ

[課題] (1)市場分析に重要な単語の抽出；カテゴライズのための単語リスト (辞書)
(2) ニュース配信時の経済状況等を考慮したテキストの特徴量化

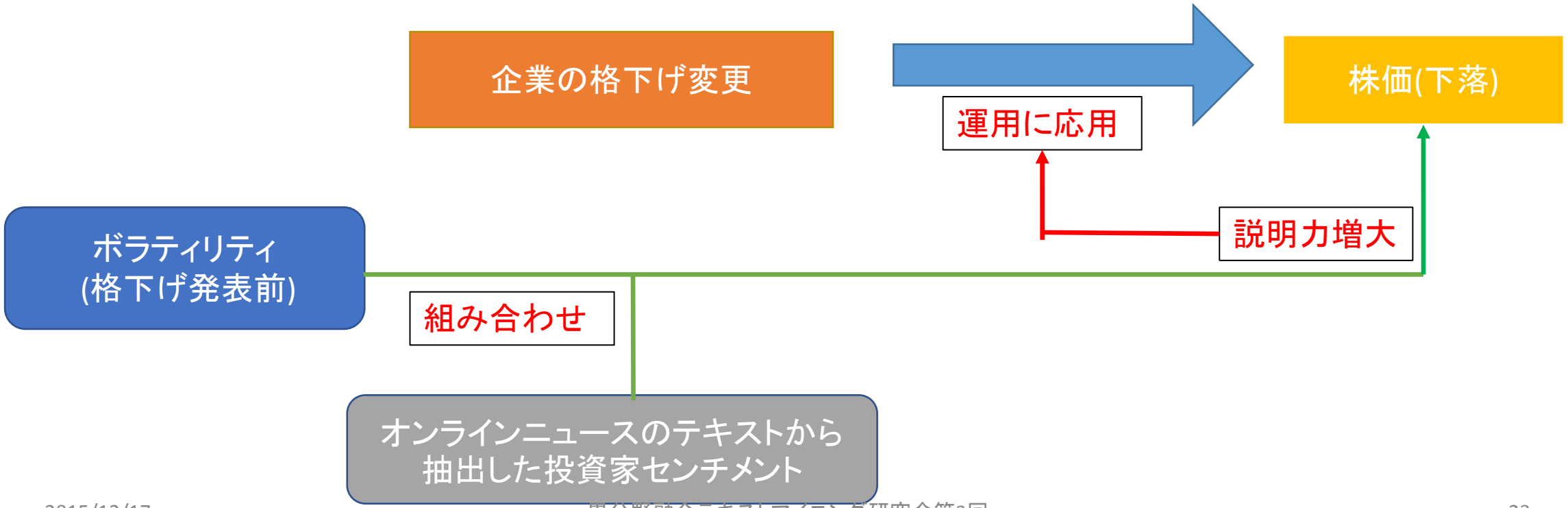
①機械学習で関係を分析：ナイーブベイズ、SVC等

②過去データ（訓練期間）の機械学習で得たルールに、新しいニューステキストを入力して予測
→40～50%の予測精度：精度向上の必要性

2.3金融テキストマイニングの研究事例

2.3.3　リアルタイム配信ニュースと短期市場変動

- 実際の運用への応用：Schmaker and Chen (2010); 岡田・羽室 (2011)
  - 岡田・羽室 (2011): 数値データの時系列解析とテキストマイニングの組み合わせ



2.3金融テキストマイニングの研究事例

2.3.4　経済リポートと長期市場変動

- 和泉他 (2010, 2011)
  - 同様の分析を、1998年～2008年の国債・株式・外国為替市場で実施：前月末から価格が上昇するか下落するかを予測
  - 日経平均株価と国債1年、2年、5年物
    - 76.92%から84.62%の高い正答率
  - 国債10年物、円ドルレートについては正答率約40～50%しかなかった
- 金融関係者の見解：妥当な結果
- 運用への適用
  - 国債2年、5年、10年物での運用で、既存のSVRや計量経済モデルに比べてほぼ最高水準の運用益をあげた