CS24

Computational
Intelligence Lab.

言語情報の深層生成モデルを用いた株価動向推定

2017/02/14

150X201X 秋田諒

Agenda

- 背景
  - 市場動向分析の需要
- 関連研究
  - 言語情報を用いた研究
  - 関連研究の問題点
- 提案手法
  - 生成モデル
  - 深層生成モデルを用いた株価動向予測
- 実験
  - 実験設定
  - 2値分類

Agenda

- **背景**
  - **市場動向分析の需要**
- **関連研究**
  - 言語情報を用いた研究
  - 関連研究の問題点
- **提案手法**
  - 生成モデル
  - 深層生成モデルを用いた株価動向予測
- **実験**
  - 実験設定
  - 2値分類

研究背景｜市場動向分析の需要

- 投資家はあらゆる**情報**を投資判断の材料としている

数値情報

- 消費者物価指数
- 株価

言語情報

- 新聞記事
- マイクロブログ

投資家

研究背景｜市場動向分析の需要

- 投資家はあらゆる**情報**を投資判断の材料としている

数値情報

- 消費者物価指数
- 株価

言語情報

- 新聞記事
- マイクロブログ

投資家

近年

すべての情報を分析することは困難

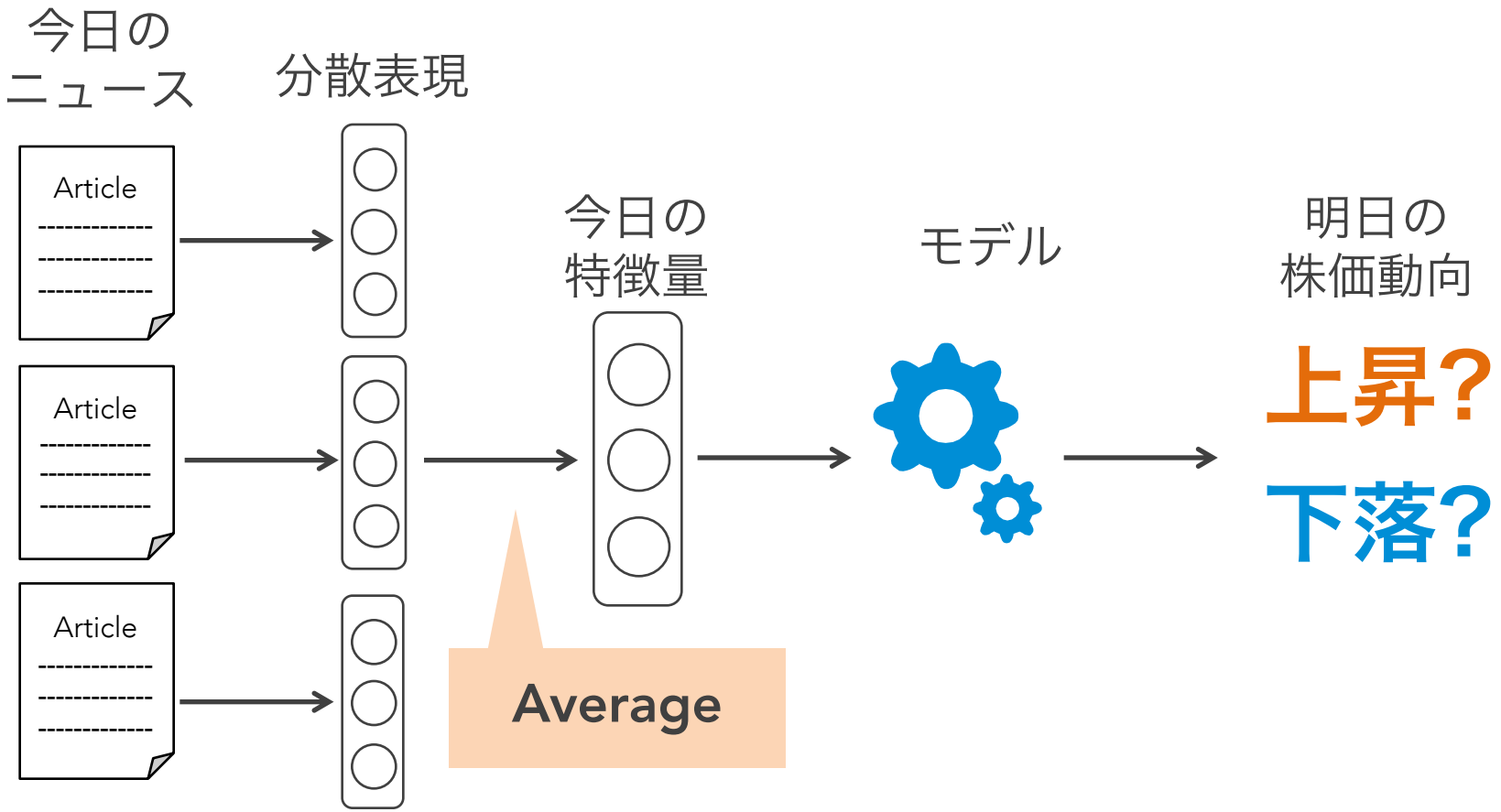
データマイニング技術を用いた
市場動向分析が注目される

Agenda

- **背景**
  - 市場動向分析の需要
- **関連研究**
  - **言語情報を用いた研究**
  - **関連研究の問題点**
- **提案手法**
  - 生成モデル
  - 深層生成モデルを用いた株価動向予測
- **実験**
  - 実験設定
  - 2値分類

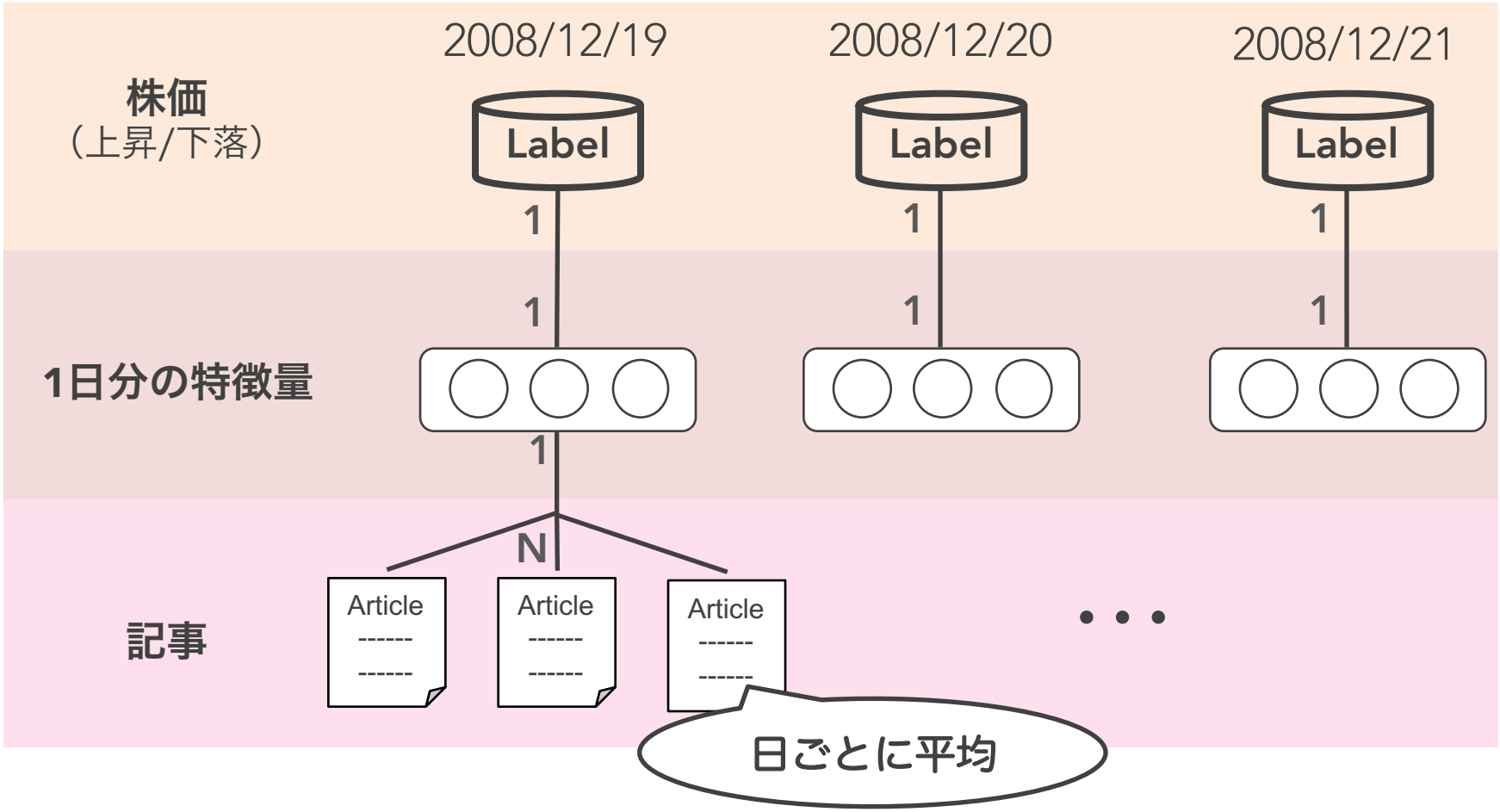
関連研究｜言語情報を用いた研究

- Deep Learning for Event-Driven Stock Prediction [Ding+, 2015]



関連研究｜問題点

- **日次予測の場合**
  - » e.g. Dingらのデータセット



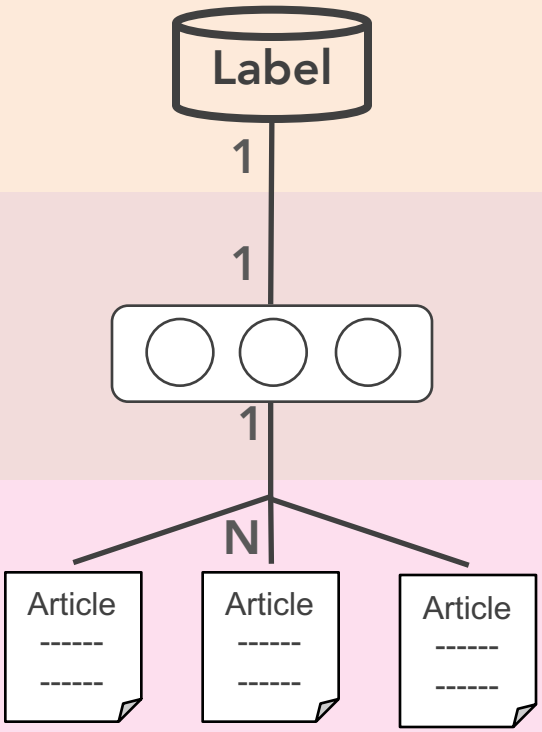
関連研究｜問題点

- **日次予測の場合**
  - » e.g. Dingらのデータセット

関連研究｜問題点

- **日次予測の場合**
  - » e.g. Dingらのデータセット

株価
(上昇/下落)



1日分の特徴量

記事

1,425 days

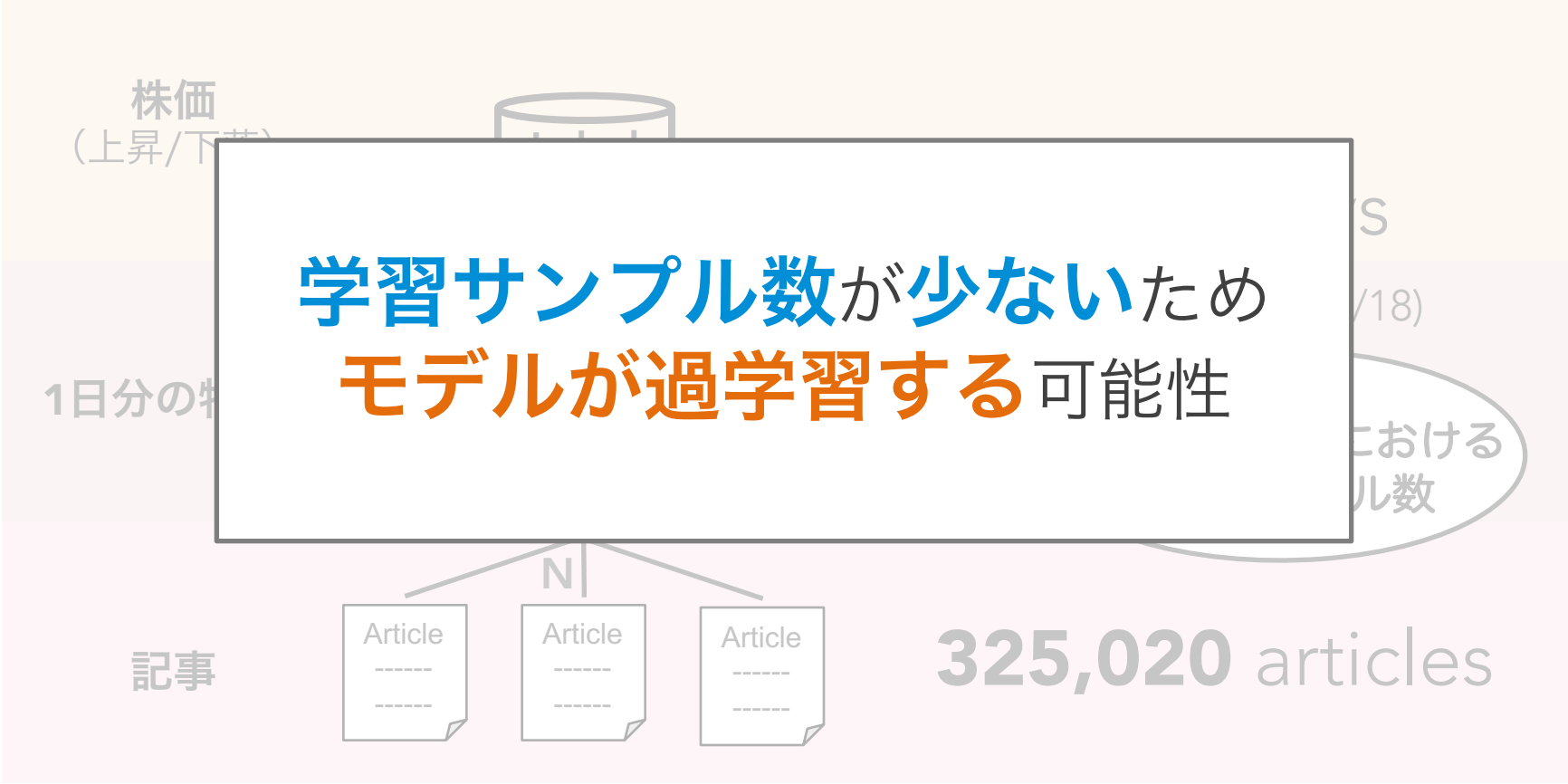
(2006/10/02 ~ 2012/06/18)

Dingのモデルにおける
学習サンプル数

325,020 articles

関連研究｜問題点

- **日次予測の場合**
  - » e.g. Dingらのデータセット



学習サンプル数が**少ない**ため
モデルが過学習する可能性

関連研究｜問題点

① 1日の特徴量の作成方法

» 各記事の平均による記事の影響の重みの無視

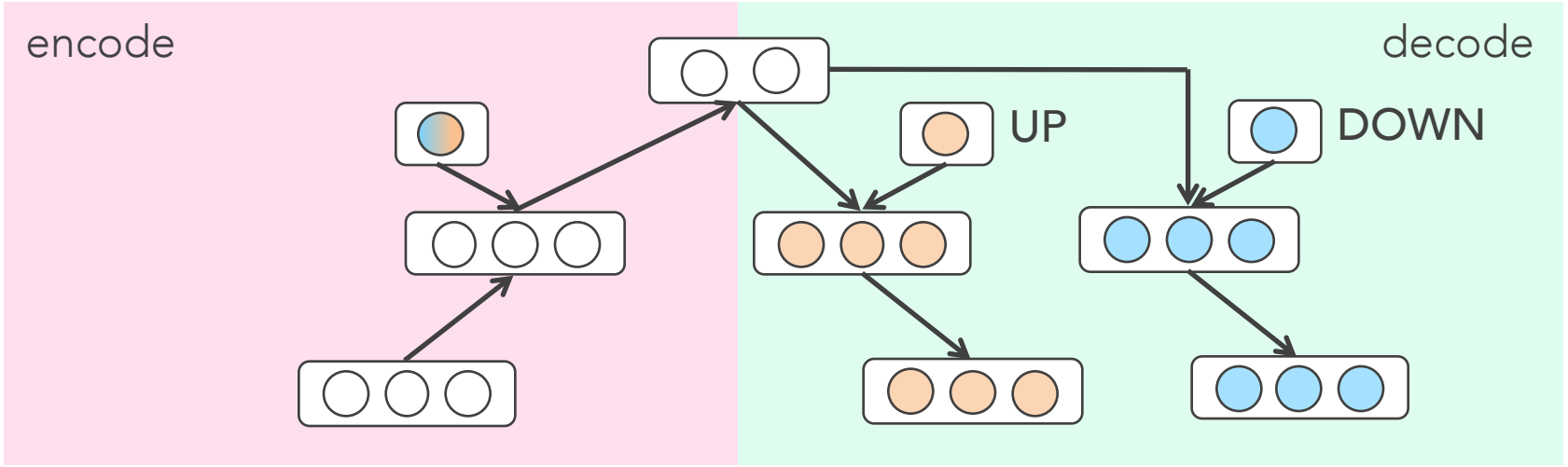
② サンプル数

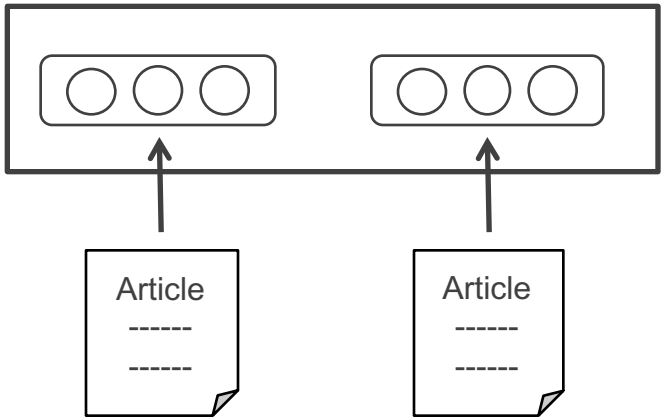
» 学習サンプル数が少ないため**過学習**する可能性

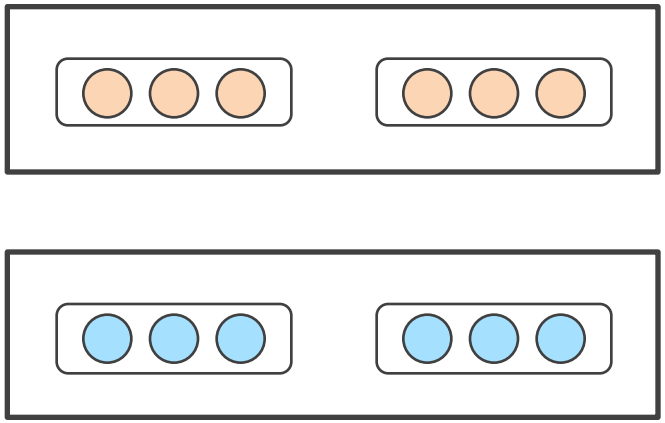
Agenda

背景	市場動向分析の需要	
関連研究	言語情報を用いた研究	関連研究の問題点
提案手法	**生成モデル**	
	深層生成モデルを用いた株価動向予測	
実験	実験設定	
	2値分類	

提案手法｜概要図

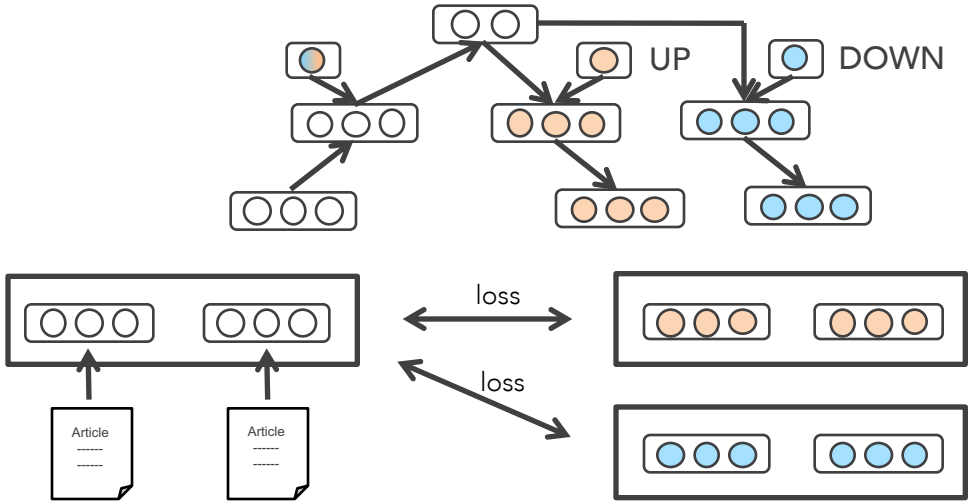






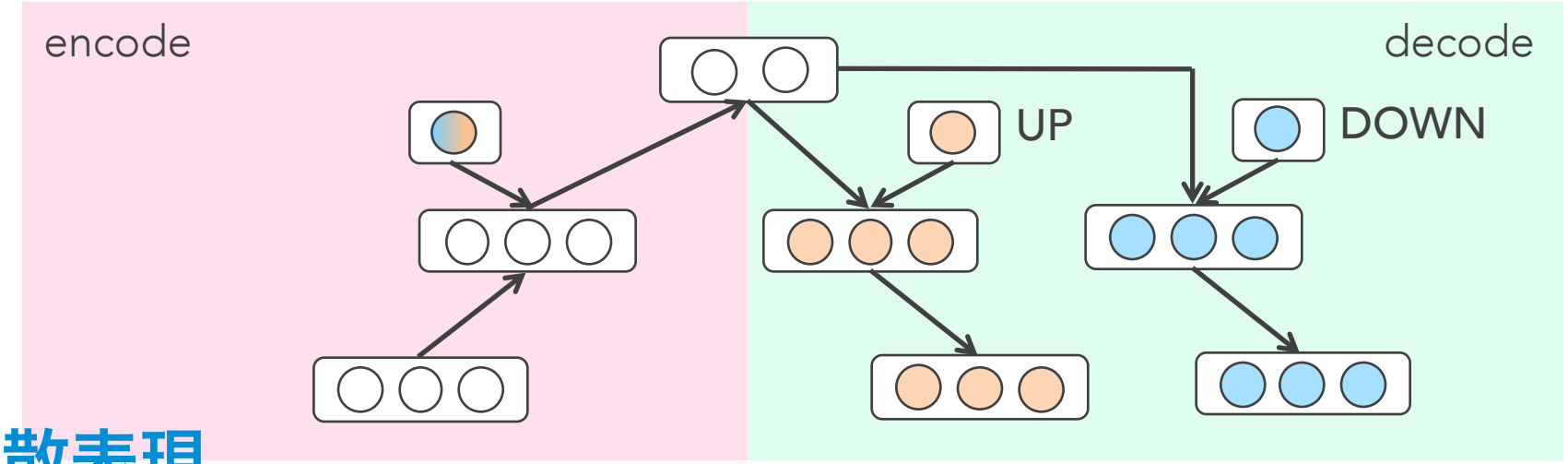
提案手法｜特徴

- **深層生成モデルによる株価動向推定**
  1. **正しいラベルと共に**再構築するとき
     元の記事$\boldsymbol{x}$を正確に再構築するように学習
  2. 1日に発行された**各記事に対して**繰り返し
     再構築によって生じた誤差を計算
  3. **誤差が小さいラベル**をその日の予測とする

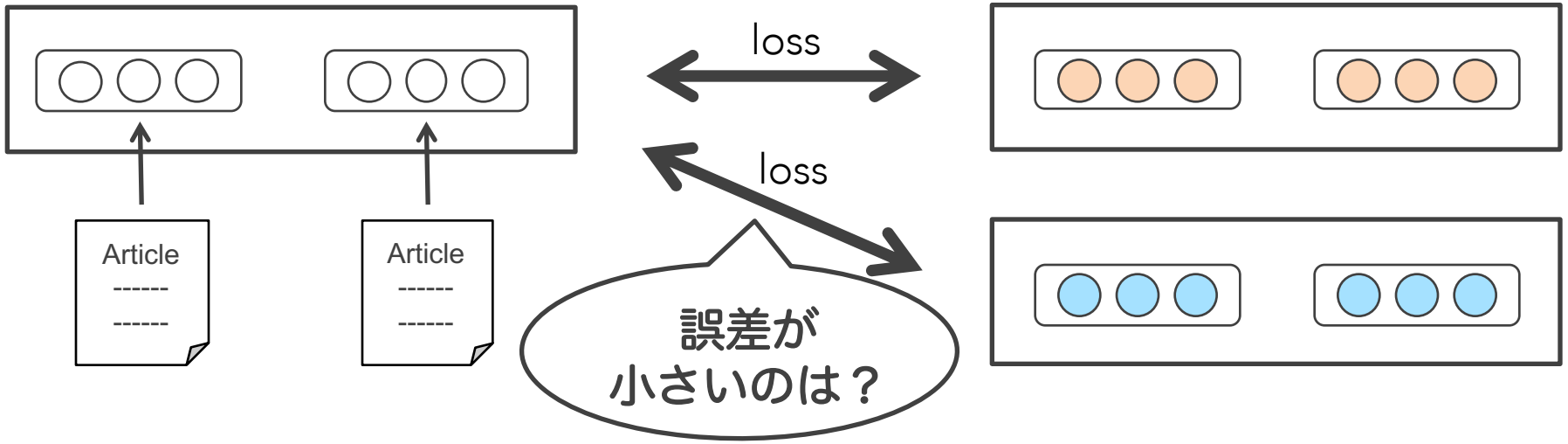


提案手法

深層生成モデル



分散表現



Agenda

- 背景
  - 市場動向分析の需要
- 関連研究
  - 言語情報を用いた研究
  - 関連研究の問題点
- 提案手法
  - 生成モデル
  - 深層生成モデルを用いた株価動向予測
- 実験
  - **実験設定**
  - **2値分類**

評価実験｜データセット

日経データセット

- **言語情報：新聞記事**
  - 日本経済新聞 本紙朝刊 8年分
    （見出し, 日経225に関する記事のみ）
    - 訓練データ　：2001年 ~ 2006年, 1,541日, 69,160件
    - 検証データ　：2007年　　,　236日, 11,666件
    - テストデータ：2008年　　,　236日, 11,699件
  - **Paragraph Vector** [Le, 2014] を用いて**分散表現**を獲得
- **数値情報：日経平均株価**
  - 言語情報と同じ期間を用意

評価実験｜実験設定

- **上がる**か**下がる**かの**2値分類**

- 評価指標
  - **Accuracy（Acc.）**
    - データの偏りに敏感
  - **Matthews Correlation Coefficient（MCC）**
    - データの偏りに左右されない

評価実験｜実験設定

- 比較手法と目的

	入力	比較目的
SVM	avg	既存研究の基準値
MLP	all/avg	**生成モデル**利用の有効性
提案手法	all/avg	---

all / avg：全ての記事/日ごとに平均
→ **サンプル数増加・各記事の考慮**の有効性

SVM：Support Vector Machine
MLP：Multilayer Perceptron

評価実験｜結果

- テストデータに対する予測結果

	Acc. (%)	MCC
majority	51.27	0.0
SVM (avg)	49.15	-0.0241
MLP (all/avg)	52.12 / 51.69	0.0486 / 0.0179
提案手法 (all/avg)	**56.35** / 47.88	**0.1248** / -0.0345

評価実験｜結果

- テストデータに対する予測結果

	Acc. (%)	MCC
majority	51.27	0.0
SVM (avg)	49.15	-0.0241
MLP (all/avg)	**52.12** / 51.69	**0.0486** / 0.0179
提案手法 (all/avg)	**56.35** / 47.88	**0.1248** / -0.0345

MLP・提案手法においてallがavgの結果を上回る

評価実験｜結果

- テストデータに対する予測結果

各記事の影響の大きさの考慮と
学習サンプル数の増加が
株価動向推定に有効

評価実験｜結果

- テストデータに対する予測結果

	Acc. (%)	MCC
majority	51.27	0.0
SVM (avg)	49.15	-0.0241
MLP (all/avg)	52.12 / 51.69	0.0486 / 0.0179
提案手法 (all/avg)	**56.35** / 47.88	**0.1248** / -0.0345

提案手法が**MLP**を上回る

評価実験｜結果

- テストデータに対する予測結果

	Acc. (%)	MCC
majority	51.69	0.0
SVM (all/avg)	51.69	0.0
MLP (all/avg)	52.12 / 51.69	0.0486 / 0.0179
提案手法 (all/avg)	**56.35** / 47.88	**0.1248** / -0.0345

生成モデルの利用が
株価動向推定に対して有効

結論と今後の課題

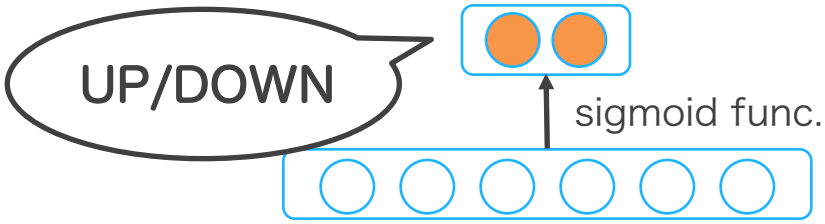
- **結論**
  - 言語情報を用いた日次の株価動向予測
    - 各記事の影響の大きさを考慮
  - 2値分類・投資シミュレーションの評価実験
    - 各記事を考慮する有効性を示した
    - 生成モデル利用の有効性を示した

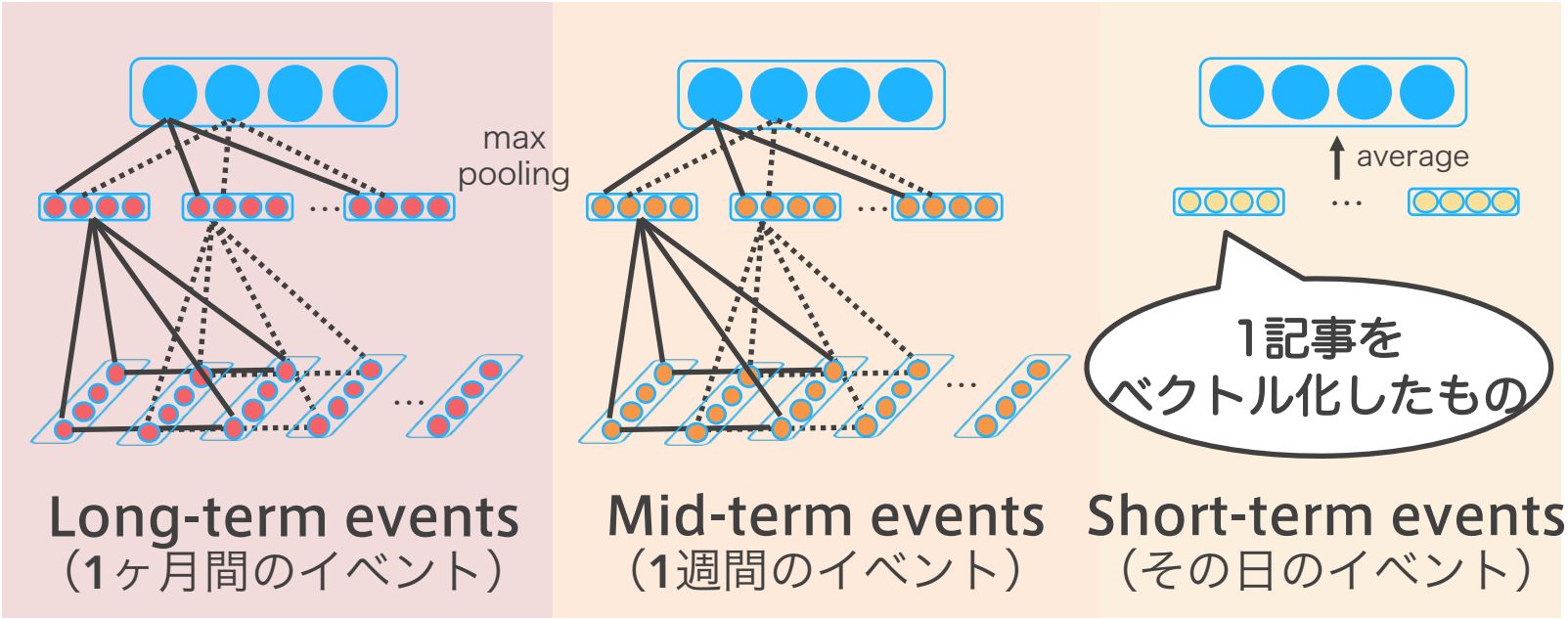
- **今後の課題**
  - 時系列情報の考慮
    - 直前のニュースしか考慮していない
    - 隠れマルコフモデルなどの時系列情報を考慮できるモデルの利用の検討
  - 数値情報の考慮
    - テクニカル指標などの使用を検討

APPENDIX

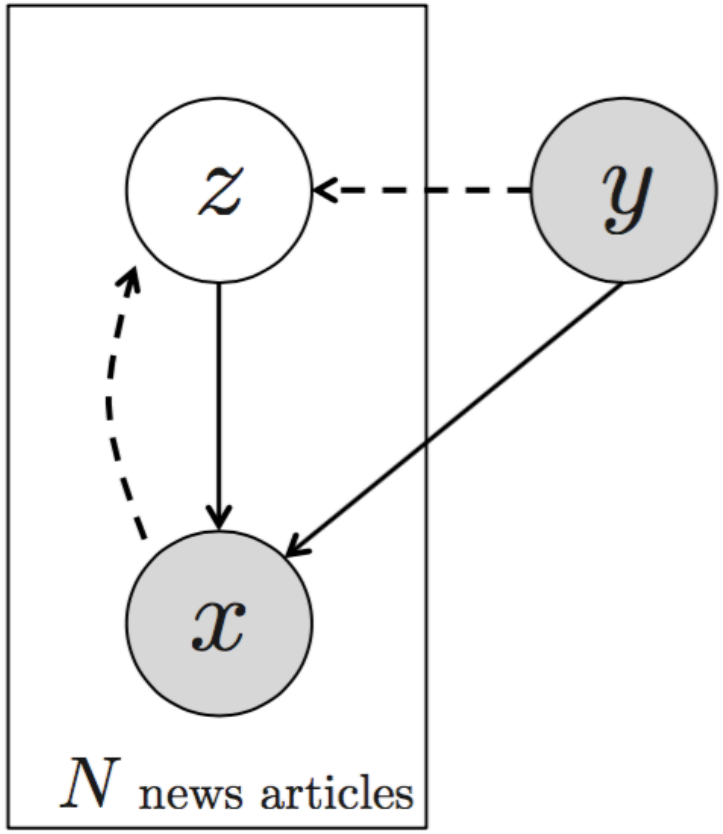
関連研究｜言語情報を用いた研究

- Deep Learning for Event-Driven Stock Prediction [Ding+, 2015]





提案手法｜生成モデル



：生成

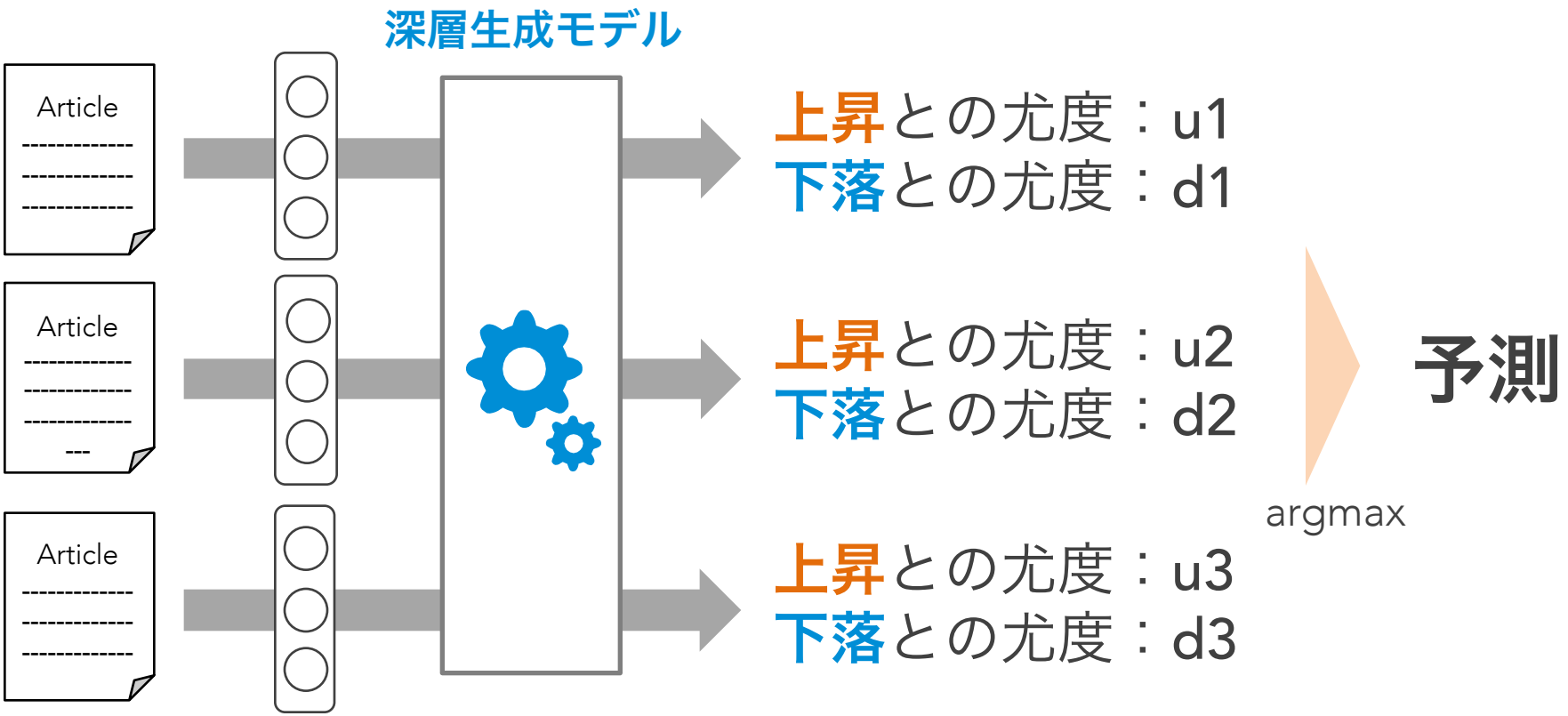
：推定

x：ニュース

y：株価

z：ノイズ

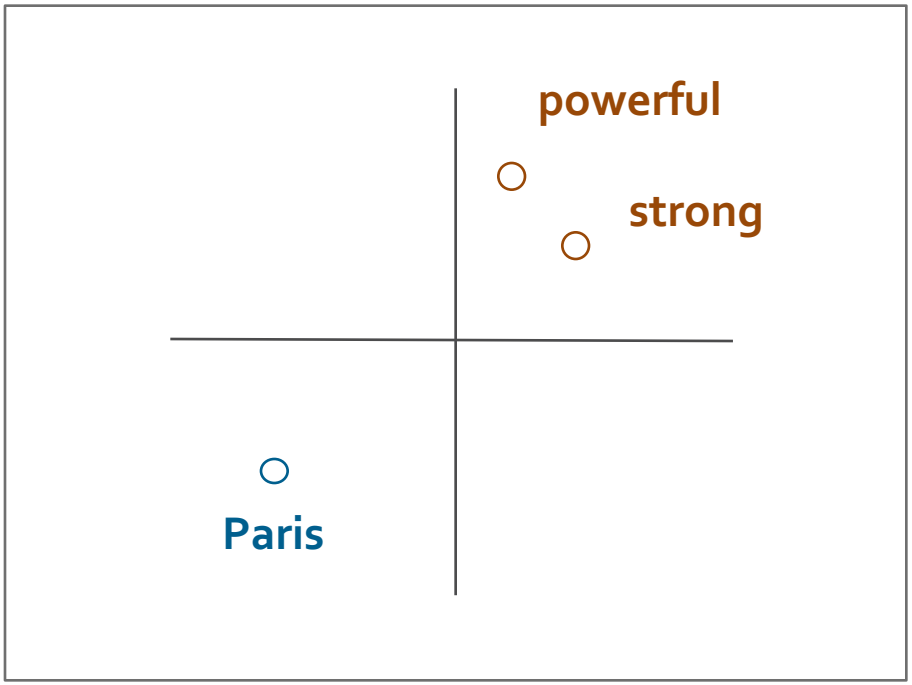
提案手法


深層生成モデル
Article
Article

Article
上昇との尤度：u1
下落との尤度：d1
上昇との尤度：u2
下落との尤度：d2
上昇との尤度：u3
下落との尤度：d3
argmax
予測

提案手法｜テキストの表現

- **分散表現の利用**
  - » 分散表現 = 単語・文の意味が近いほど類似するベクトル



分散表現の例（単語）

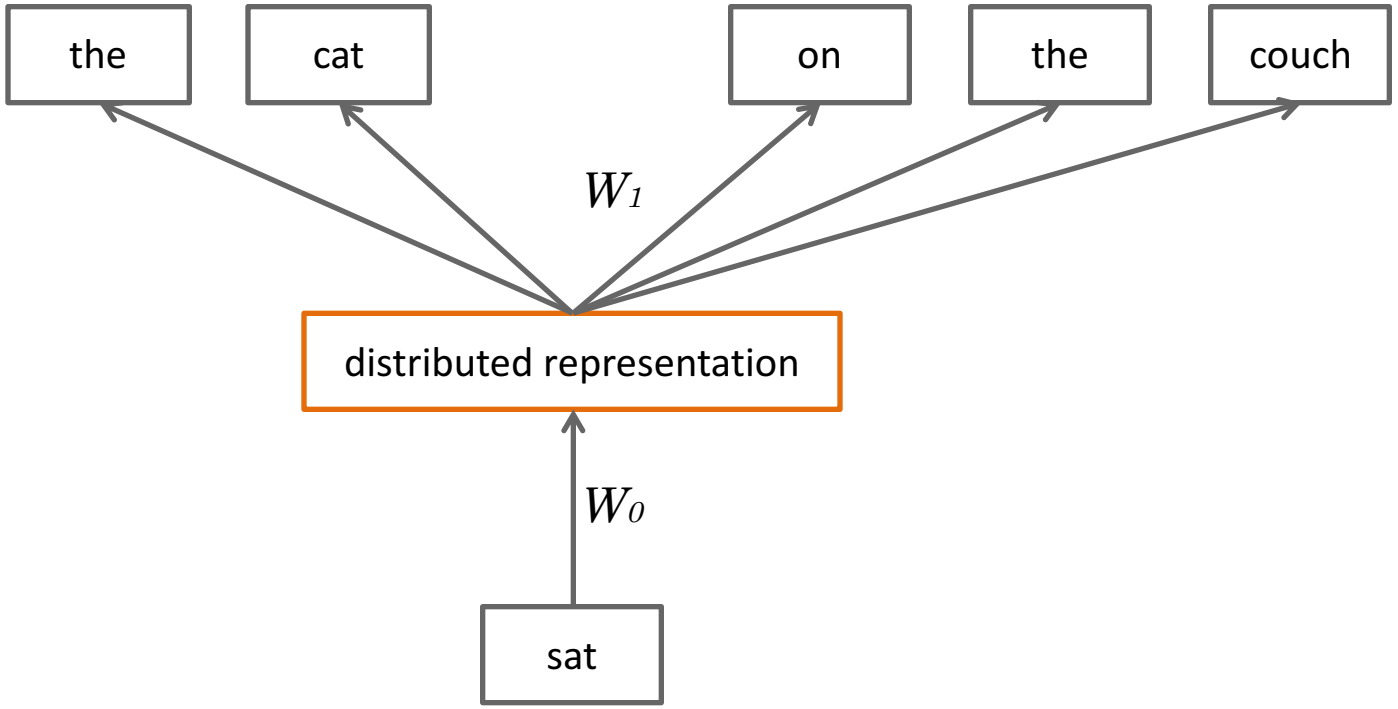
Paragraph Vector

- **Word2Vec（Skip gram）の拡張**
  - » 周辺単語から単語のベクトル表現を学習

The cat __?__ on the couch last night .

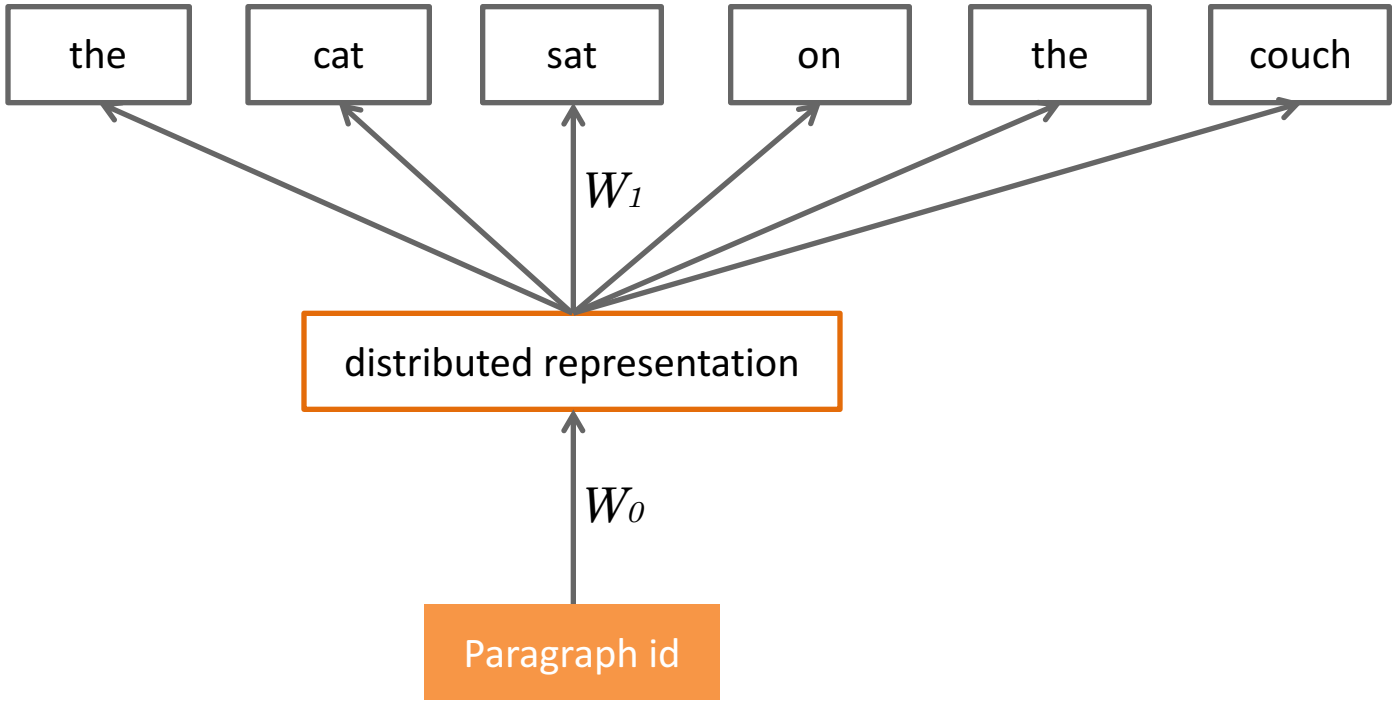
__?__ = sat

Word2Vec（Skip gram）



- 各単語の前後数単語の出現確率が高くなるようNNを学習させる
  - » 目的関数： $\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$
- 中間層の出力をその単語の特徴とする

Paragraph Vector（Distribute Bag-of-Words）



- Word2Vecで単語を学習させる際にparagraph idも入力に含めてSkip gramを学習させる
- パラグラフに共起する単語特徴に関する分散表現を獲得することができる

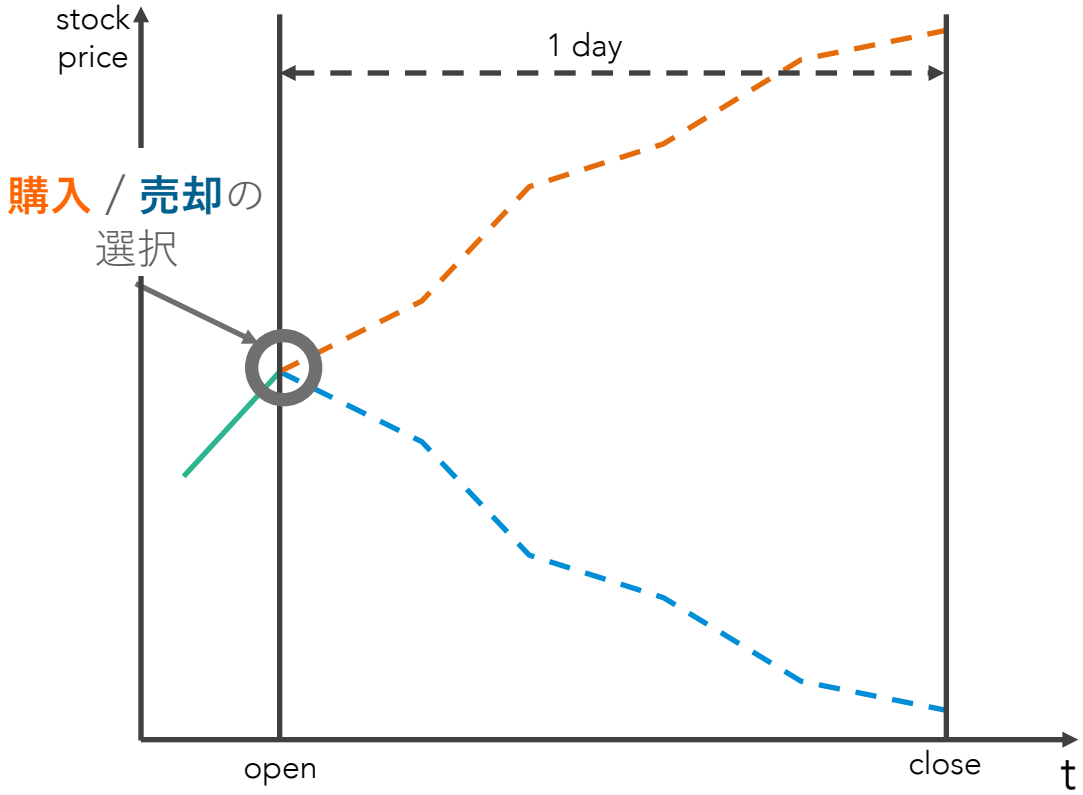
Sentiment Analysis

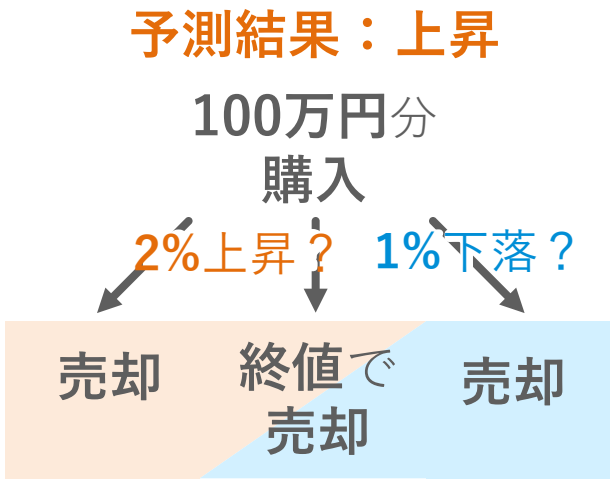
- Results

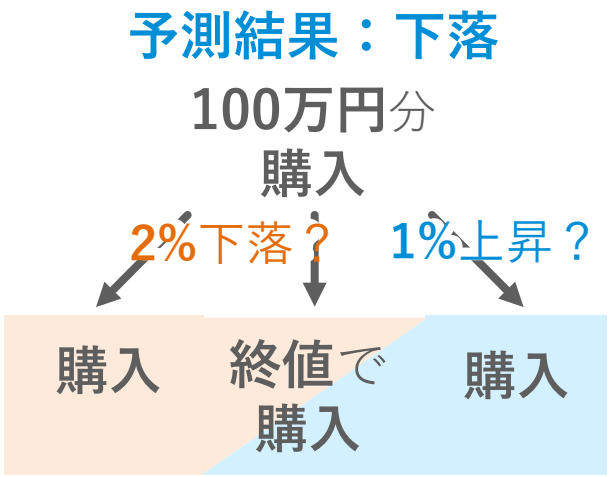
Model	Error rate
BoW (bnc) (Maas et al., 2011)	12.20 %
BoW (bΔt'c) (Maas et al., 2011)	11.77%
LDA (Maas et al., 2011)	32.58%
Full+BoW (Maas et al., 2011)	11.67%
Full+Unlabeled+BoW (Maas et al., 2011)	11.11%
WRRBM (Dahl et al., 2012)	12.58%
WRRBM + BoW (bnc) (Dahl et al., 2012)	10.77%
MNB-uni (Wang & Manning, 2012)	16.45%
MNB-bi (Wang & Manning, 2012)	13.41%
SVM-uni (Wang & Manning, 2012)	13.05%
SVM-bi (Wang & Manning, 2012)	10.84%
NBSVM-uni (Wang & Manning, 2012)	11.71%
NBSVM-bi (Wang & Manning, 2012)	8.78%
Paragraph Vector	**7.42%**

評価実験｜シミュレーション設定

- 取引方法[Lavrenko+, 2002]







：利益＋　：利益－

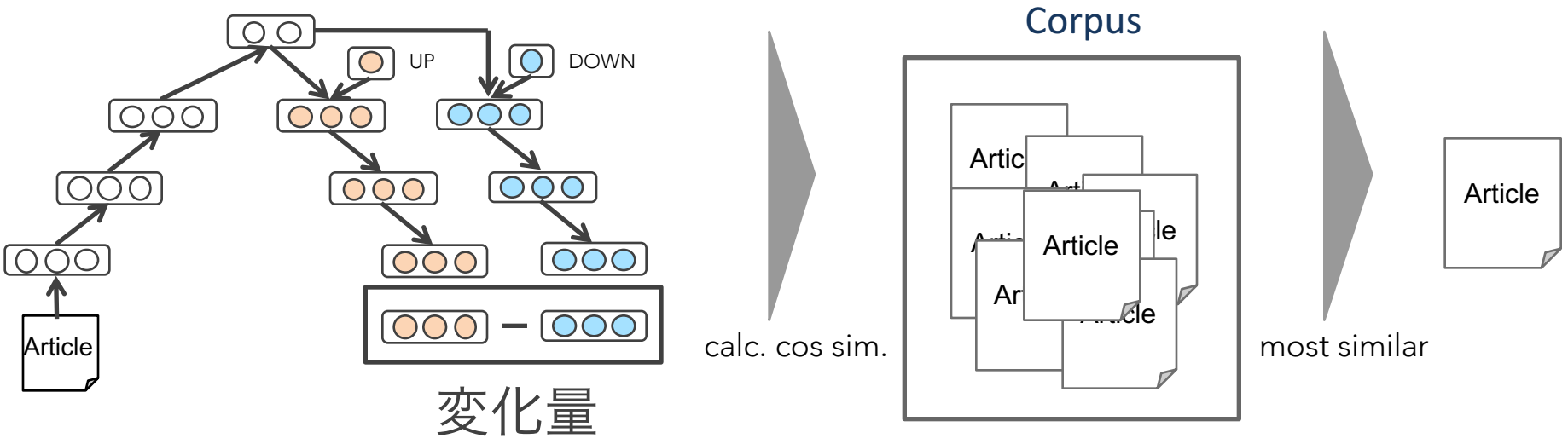
評価実験｜株売買シミュレーション

- テストデータに対する利益

	利益 (円)
majority	228,273
SVM (avg)	102,277
MLP (all/avg)	290,816 / 148,766
提案手法 (all/avg)	**336,681** / 142,687

定性的な評価｜実験設定

- **実験方法**
  - » 2種類のラベルでモデルが生む極性の変化量を獲得
  - » 元のベクトルと足し合わせ、記事の極性を変化
  - » 最もコサイン類似度が高い記事を獲得することで確認



定性的な評価｜実験結果

UP → DOWN	3月のビール出荷，2年2ヵ月ぶり増――新製品が需要喚起 → 第2部企業の興亡 (1) 値崩れの波をくぐれ (デフレが蝕む) キリン HD 純利益 12%増，前期 600 億円，年間配当 19―20 円に → TDK の前期，純利益を訂正――333 億円に減額 Nestle, Sara Lee profits lifted by price increases → Toshiba, Fujitsu hit by price falls, outlook rough Hyundai targets 2007 revenue growth spurt → Qualcomm profit falls, cuts '09 revenue target
DOWN → UP	9月中間経常，キッコーマン，13%減益――冷夏・円高が打撃 → 円高一服好感し反転 ――電機・自動車株が上げ主導 (株式往来) 高成長中国，潜むリスク――日本企業にも影響，鉄鋼は減産，輸出も減少 → 非鉄金属，軒並み高，銅は7ヵ月ぶり水準，実需買いに影響も McClatchy sees ad revenue down in first half of 2007 → SES sees pay TV revenue increasing 34% worldwide in 2016 January housing starts down 14.3 percent → Instant view: CPI rises; housing starts up 15 percent

評価実験｜データセット

S&Pデータセット

- **言語情報：Webニュース**
  - Reuter・Bloomberg（見出しのみ）
    - 訓練データ：2006.2.10 ~ 2012.06.18,　1,425日, 442,933件
    - 検証データ：2012.06.19 ~ 2013.02.21,　169日,　34,868件
    - テストデータ：2013.02.22 ~ 2013.11.21, 191日,　35,603件
- **数値情報：S&P500**
  - 言語情報と同じ期間を用意

評価実験｜結果

- 2値分類結果（S&Pデータセット）

	Acc. (%)	MCC
majority	59.47	0.0
SVM (avg)	58.95	0.0366
MLP (all/avg)	54.74 / 39.47	0.0252 / -0.0961
提案手法 (all/avg)	**61.05** / 50.00	**0.1379** / -0.1297

評価実験｜株売買シミュレーション

- シミュレーションの利益（S&Pデータセット）

	利益 ($)
majority	1,914
SVM (avg)	1,558
MLP (all)	670
提案手法 (all)	**2,319**

評価実験｜株売買シミュレーション

- テストデータに対する利益

	利益 (円)	利益 ($)
majority	228,273	1,914
SVM (avg)	102,277	1,558
MLP (all)	290,816	670
提案手法 (all)	**336,681**	**2,319**