2D3-OS-19a-4

ネットワークの表現学習による金融専門極性辞書の構築

Lexicon Creation for Financial Sentiment Analysis using Network Embedding

伊藤　諒 *1　　和泉　潔 *2　　須田　真太郎 *3
Ryo Ito　　Kiyoshi Izumi　　Suda Shintaro

*1*2東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo

*3株式会社 三菱 UFJ トラスト投資工学研究所
Mitsubishi UFJ Trust Investment Technology Institute Co.,Ltd.

It is necessary to build a comprehensive polarity dictionary specialized for financial policy to improve the accuracy of lexicon-based sentiment analysis in evaluating texts written on financial policy. In this research, we acquire distributed representation of words using feature learning of dependency network of words and create the polarity dictionary by bootstrap method using the distributed representation of words.

1. はじめに

金融・経済分野における新たな分析データとして，非構造化データであるテキスト情報が注目を集めており，テキストマイニングを金融・経済分野に応用した研究が盛んに行われている．テキストマイニングには，従来では指標化されていなかった，市場や企業に関する情報をテキスト情報から抽出することが期待されているが，そのような抽出対象の情報の一つとして，ある事象に対してポジティブもしくはネガティブかを評価するセンチメント指数が挙げられる．そして，Bollen and Huina(2011) の研究にみられるように，テキストから抽出したセンチメント指数と市場変動との関係性を検証する研究が多く行われている．

テキストからセンチメントを定量化する手法は，センチメント分析（Sentiment Analysis）において，これまでに数多くの研究がなされている．センチメント分析には様々な手法が存在するが，大きく分けて Pang et al.(2002) を始めとする機械学習のアプローチと，Turney(2002) を始めとする語彙ベースのアプローチに大別される．機械学習のアプローチでは，テキストの特徴量とポジティブやネガティブなどの極性ラベルとの関係性を，機械学習モデルによって学習し，未ラベルテキストに対して学習済みモデルを適用することで極性を付与する．一方，語彙ベースのアプローチでは，テキスト中に出現するポジティブな単語の出現比率とネガティブな単語の出現比率の差を以って，対象となるテキスト全体の極性を算出する方法が用いられる．センチメント分析に関するさらなる詳細は，Kumar Ravi et al.(2015) を参照されたい．

ここにおいて，語彙ベースによるアプローチを用いた場合，極性語とその極性値が組となった極性辞書が必要となるが，膨大な数の単語に対して人手で極性値を付与していくことは，コストの観点から現実的ではない．また，単語の持つ極性はその単語が出現する背景・文脈によって異なり，解析対象となるテキストに適した極性辞書が必要である．一例として，Loughran and McDonald(2011) は，語彙ベースによるアプローチにおいて広く用いられている H4N（Harvard-IV-4 TagNeg）に含まれるネガティブな単語が，ファイナンスの文脈においてネガティブな極性を有するとは限らず，ファイナンス文書のセンチメント分析において，ファイナンス専用の辞書を用いる事の重要性を指摘している．以上のように，語彙ベースによるセンチメント分析のアプローチを用いた場合，網羅的でかつ解析対象に適した辞書を構築する事が必要であり，かつ辞書構築が自動でなされる事が期待される．

このような中，Jegadeesh and Wu(2015) の研究を代表として，近年テキストマイニングを応用して金融政策の効果を分析する新たな研究が登場している．Jegadeesh and Wu(2015) は，米国の金融政策を策定する委員会である Federal Open Market Committee（FOMC）の議事録からトピックを LDA により抽出し，Loughran and McDonald(2011) により導入された辞書を用いて，トピック別にセンチメントを付与し，各トピックのセンチメントがマクロ変数や資産価格に対して与える影響を分析している．ここにおいて，Loughran and McDonald(2011) の辞書は同じファイナンスの分野であれど，企業の財務報告書を基に作成された辞書であり，金融政策の分析に特化して作成された辞書ではない．さらに，伊藤他 (2017) は Jegadeesh and Wu(2015) の手法を拡張したトピック別センチメント付与方法を用いて，得られたトピック別センチメントと市場参加者の期待形成に関するイベントスタディ分析を行っているが，ここで用いられている辞書は，金融政策の分析に特化した人手で作成された辞書であり，テキストに出現する単語の種類の多さから，全ての極性語を網羅した辞書を構築出来ているとは言えない．

そこで本研究では，金融政策の分析に特化した極性辞書の自動構築を行うことを目的とし，単語の特徴を定量的に扱うために，とりわけ近年研究の進んでいるネットワークの表現学習を用いた辞書自動構築方法を提案する．

2. 関連研究

本章では，極性語を対象にした辞書構築（Lexicon Creation）に関する先行研究について述べる．

極性辞書構築に関して数多くの研究が行われているが，大きく分類して，辞書ベースのアプローチと，コーパスベースのアプローチに分けられる．

辞書ベースのアプローチとしては，辞書から語彙ネットワークを構築し，その語彙ネットワーク上に種表現を元にして極性を

*1 連絡先: 伊藤諒，東京大学大学院工学系研究科システム創成学専攻和泉研究室，〒 113-8654 東京都文京区本郷 7-3-1，E-mail: m2016rito@socsim.org

*3 留意事項: 本稿の内容は筆者が所属する組織を代表するものではなく，すべて個人的な見解である．また，当然のことながら，本稿における誤りは全て筆者の責に帰するものである．

伝搬させる事で，各単語に対して，極性を付与する方法が代表的である．Kamps et al.(2004) は，英語概念辞書の WordNet[*1] を用いて種表現から極性値を伝搬させる方法を提案している．

また，コーパスベースのアプローチでは，単語の共起情報や文脈情報を用いて極性語を取得する方法が代表的である．Turney(2002) は，2 単語間の共起度合いを定量的に表す PMI（Pointwise Mutual Information）を用い，ある単語が種表現として与えられたポジティブな単語とネガティブな単語のどちらと共起しやすいかを以って，単語の極性を付与する方法を提案している．

さらに，近年単語の分散表現や教師あり学習を用いた，極性辞書構築方法が提案されている．分散表現を用いた研究として，片倉・高橋 (2015) は，極性語同士は類似した分散表現を持つという考えの下，CBOW モデル（Continuous bag-of-words model）によって単語の分散表現を得た後に，シードとなる極性語と類似した分散表現を持つ単語を取得する事で，Loughran and McDonald(2011) の辞書の拡張を行なっている．また，教師あり学習を用いた手法として，坪内・山下 (2014) は，リッジ回帰を用いて，文書に付与されたポジティブ・ネガティブのフラグを学習し，学習モデルの各フレーズに対する係数をフレーズの極性値として付与している．

コーパスベースのアプローチでは人手で構造化されていない情報を入力とするため，一般に辞書ベースのアプローチよりも辞書構築の精度は劣るが，入力として専門領域について書かれたテキストを入力にする事で，専門領域に特化した極性辞書を構築できるという利点がある．

以上の背景の下，本研究ではコーパスベースのアプローチを用いて，金融政策のセンチメント分析に特化した極性辞書を作成する事を目的とする．また，本研究ではネットワークの表現学習によって得られた分散表現を用いた，新たな極性辞書構築の為の手法を提案する．

3. ネットワークの表現学習を用いた極性辞書の構築

本章では，ネットワークの表現学習を用いた極性辞書の構築方法について提案するが，はじめに本研究の着想について述べ，次に具体的なフレームワークについて述べる．

3.1 本研究の着想

極性を持つセンテンス中において出現する極性語が，他の単語とどのような関係を有しているかを考える．図 1 は極性を持つセンテンスについて，極性語の係り先を図示したものである．

この例で特徴的な点は，これらの極性語が同じ単語である rate に掛かっている点である．increased は decreased と対義語の関係にあるが，両者とも数量の大きさを評価するという概念を持ち，評価対象の単語は類似した単語になることが考えられる．また，increased や decreased に対する類義語も，同様に数量の大きさを評価するという概念を持つことから，評価対象の単語は類似した単語になることが考えられる．以上の観察から，極性語同士は係り先となる単語が類似しており，単語間の係り受け関係には，極性語としての特徴が含まれると考えられる．

さて，単語間の係り受け関係を表現するものとして，ネットワークによる表現がある．ここで，先の観察をネットワークの観点で表現すると，ネットワーク上において極性語同士はある別の単語を介して繋がっており，二次近接の関係にあると言える．よって，単語間の係り受けネットワークにおける二次近接の関係性を以って単語を定量化する事で，単語の持つ極性語らしさを定量的に扱えると考える．そこで，本研究では初めに金融政策テキストから単語間の係り受けネットワークを構築する．次に，ネットワークの各ノード（単語）に対して，二次近接の関係性を，J Tang et al.(2015) によってネットワークの表現学習手法として提案された LINE（Large-scale Information Network Embedding）を用いて定量化する．そして，定量化された単語の分散表現と種表現の極性語を元に，ブートストラップ法を用いて極性語辞書構築を試みる．

分散表現を用いて極性辞書構築を行う研究として，前章で述べた片倉・高橋 (2015) の研究が存在するが，この研究において単語の分散表現の獲得には CBOW モデル（Mikolov et al. 2013）が用いられている．CBOW モデルではある単語 x が現れている周辺の単語を用いて x を推定する中で，単語の分散表現を獲得するが，周辺単語の組み合わせの多さ故，CBOW モデルで獲得される分散表現は LINE よりも意味の広い分散表現となることが考えられる．一方，LINE では CBOW モデルよりも焦点を絞った近接関係を元にして分散表現を得る為，極性語の獲得において，より良い分散表現であると期待される．

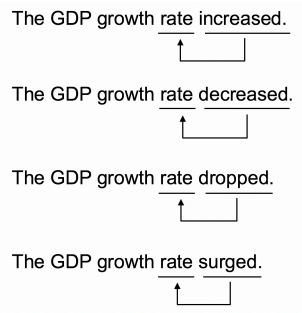

図 1: 極性語の係り先の類似性

3.2 係り受けネットワークの構築

はじめに，解析対象となるテキストから単語間の係り受けネットワークを構築する．本研究では金融政策の評価に特化した辞書を作成するという目的から，解析対象となるテキストとして，米国の金融政策を策定する委員会の議事録である FOMC 議事録を用いる．

まず FOMC 議事録の HTML を，FRB のホームページ [*2] よりウェブクローラーによって収集する．次に収集した議事録の HTML をパースし，テキスト部分のみを抽出する．そして，取得したテキストに対して係り受け解析を行うことで，単語の係り受けペアを抽出する．この際，各単語に対してレンマ化を行う．さらに，単語の係り受けペアから重み付き有向グラフを作成する．ここで，ネットワークの各ノードは各単語に相当し，ネットワーク上のあるノード v_i から v_j に貼られるエッジの重み w_{ij} は，その単語の係り受けペアの出現回数に相当する．また，エッジの方向性は係り元から係り先としてネットワークを構成する．

3.3 LINE による単語の分散表現獲得

前節で得られた単語間の係り受けネットワークを用いて，単語の分散表現をネットワークの表現学習手法である LINE によって得る．LINE には一次の LINE と二次の LINE が存在

*1 https://wordnet.princeton.edu/wordnet/

*2 https://www.federalreserve.gov/

するが，ここでは二次近接の関係にある単語に対して，類似した分散表現を獲得させたいため，二次の LINE を用いる．

二次の LINE ではノード間で共通して隣接しているノードが多ければ，それら 2 つのノードが類似した分散表現を持つように学習をする．具体的には，あるノード v_i から他のノードに対してエッジが貼られる確率分布が，分散表現から予測される分布と観測される分布とで類似した分布になるように学習を行う．ここであるノード v_i に対してノード自身のベクトル $\overrightarrow{u_i}$ とノードの文脈ベクトル $\overrightarrow{u_i}'$ を導入する．分散表現から予測されるあるノード v_i から他のノード v_j に対してエッジが貼られる確率は以下の式で表される．

$$p_2(v_j|v_i) = \frac{\exp(\overrightarrow{u_j}'^{\mathrm{T}} \cdot \overrightarrow{u_i})}{\sum_{k=1}^{|V|} \exp(\overrightarrow{u_k}'^{\mathrm{T}} \cdot \overrightarrow{u_i})}$$

また，ノードの集合として V，エッジの集合として E を持つ，ネットワーク G から観測可能な，あるノード v_i から他のノード v_j に対してエッジが貼られる確率は以下となる．

$$\hat{p_2}(v_j|v_i) = \frac{w_{ij}}{\sum_{k \in N(i)} w_{ik}}$$

$N(i)$ はノード i からエッジを貼られているノードの集合を表す．

ここで，上記の 2 つの分布間の距離の重み付き総和を目的関数とし，これが最小になるように学習を行う．目的関数は以下の式で表される．

$$O_2 = \sum_{i \in V} \lambda_i d(\hat{p_2}(\cdot|v_i), p_2(\cdot|v_i))$$

λ_i はノード i のネットワークにおける重要度を表し，$\sum_{k \in N(i)} w_{ik}$ の値が用いられる．二つの分布間の距離 $d(\cdot|\cdot)$ を KL 情報量として，定数項を除くと以下の形を得る．

$$O_2 = -\sum_{(i,j) \in E} w_{ij} \log p_2(v_j|v_i)$$

そして，この目的関数を負例サンプリングを用いて最適化する事で，各ノードに対する分散表現を得る．

3.4 ブートストラップ法による極性語の獲得

次に LINE によって得られた単語の分散表現を用いて，ブートストラップ法により，極性語候補の獲得を行う．

まず，ブートスラップを行う際にシードとなる極性語を人手により与え，これらのシードとなる単語を極性語候補リスト C へ登録する．また，ある単語 d がどれほど極性語らしいかを表す値である信頼度 P_d を導入するが，ここでシードとなる全ての極性語の信頼度を 1 として与える．

次に極性語候補リストとなる単語以外の各単語に対し，信頼度 P_d を，単語 d と極性語候補リスト中の各単語 c との類似度と，単語 c が持つ信頼度 P_c との積の平均により算出する．

$$P_d = \frac{1}{|C|} \sum_{c \in C} sim(\overrightarrow{u_c}, \overrightarrow{u_d}) P_c$$

$sim(\overrightarrow{u_c}, \overrightarrow{u_d})$ は 2 単語間の分散表現の類似度を表す関数であり，ここではコサイン類似度を用いる．また $\overrightarrow{u_d}$ は単語 d の持つ，LINE によって得られたノード自身のベクトルに対応する．

$$sim(\overrightarrow{u_c}, \overrightarrow{u_d}) = \frac{\overrightarrow{u_c} \cdot \overrightarrow{u_d}}{|\overrightarrow{u_c}||\overrightarrow{u_d}|}$$

さらに，極性語候補リストとなる単語以外の各単語に対し信頼度 P_d を計算した後，信頼度 P_d のスコアが上位 L 件以内に含まれる単語を，次の反復において使用するために極性語候補リスト C に追加する．

以上のステップを繰り返し行うことによって，極性語候補リスト C を拡張し，停止条件の回数 M に達した場合，信頼度のスコアが上位 N 件以内に含まれる単語を，システムの最終的なアウトプットとして出力する．

4. 実験

LINE によって得られた分散表現を用いて，ブートストラップ法を行う事で極性語を取得する実験を行った．また比較実験として，CBOW モデルによって得られた分散表現を入力とした場合と比べて，どのような語が取得されるかを検証した．ブートストラップにおける各パラメーター L，M，N はそれぞれ 1，30，30 とした．さらにシードとなる極性語として，increase・high・improvement の 3 単語を用い，LINE と CBOW モデルにおける分散表現は，両手法とも 50 次元とした．また，CBOW モデルにおけるウィンドウサイズは 4 として学習を行った．

5. 結果と考察

本章では，実験結果と考察について述べる．とりわけ，LINE によって得た単語の分散表現と CBOW モデルによって得た単語の分散表現の相違と極性語の取得可能性の観点から論じる．

表 1 は，各モデルによる分散表現を入力として，ブートストラップ法を用いる事で獲得された単語の内，信頼度上位の 20 単語を示したものである．

表 1: ブートストラップ法で取得された単語の内，信頼度上位 20 件以内の単語

LINE	CBOW モデル
rise, decline, decrease, drop, advance, climb, showing, fall, jump, there, faster-than-anticipated, quickened, constraining, weaker-than-anticipated, slower-than-expected, differently, better-than-anticipated, differing, higher-than-expected, approximating	decline, rise, drop, decrease, advance, fall, fell, surge, jump, decelerate, climb, rebound, contract, accelerate, deceleration, run-up, gain, step-up, weaken, recover

LINE による分散表現を用いた場合において，decrease, drop, advance, fall などをはじめとする極性語が取得できている事が分かる．とりわけ，climb などの単語は日常的に用いられる文脈では極性を持たないが，物価等が上昇するという意味合いで，金融政策テキスト中には多く出現する表現であり，このような単語を極性語として取得する事が出来ている．しかしながら，there などの極性語ではない単語も含まれる結果となった．

一方，CBOW モデルによる分散表現を用いた場合においても，decline, rise, drop などの極性語を取得する事が出来てお

り，さらに climb などの，金融政策の文脈において極性を持つ単語も取得する事が出来ている．

LINE による分散表現と CBOW モデルによる分散表現を，それぞれ用いた場合の差異について述べる．ブートストラップ法において，反復処理を行う中で目的となるインスタンス以外のインスタンスが抽出されてしまう問題は，意味ドリフト（Curran et al. 2007）と呼ばれているが，LINE による分散表現を用いた場合，この意味ドリフトが比較的早く起きている事が観察された．具体的には極性語ではない differently が獲得された後に，differing が獲得されるというように，意味がドリフトしていく様子が観察された．意味ドリフトが起こる原因の解明や抑制は今後の課題である．

6. まとめ

専門領域について書かれたテキストに対して語彙ベースのセンチメント分析を行う上で，専門領域に特化した網羅的な辞書がセンチメント分析の精度向上のために必要である．本研究では，金融政策の分析に特化した極性辞書の自動構築を行うことを目的とし，極性語同士は係り受けネットワークにおいて二次近接に位置する度合いが高いという考えの下，二次近接の関係を定量化するネットワークの表現学習手法である，二次の LINE を用いた辞書自動構築方法を提案した．

今後の課題として多くの課題が挙げられるが，主に次の四点が重要な課題として挙げられる．

まず第一点目として，提案手法による辞書構築の精度評価を定量的に行うことである．またその中において，CBOW モデルを用いた場合と提案手法を用いた場合の精度を比較する必要がある．さらに，シードとして与える極性語に対する辞書構築精度のロバストさを両手法とも検証する必要がある．

また第二点目として，提案手法を用いた場合，単語のポジティブ・ネガティブに関係なく極性語候補に追加されるが，取得された単語に対してポジティブ・ネガティブに分類した辞書を構築できるよう，提案手法を拡張することが課題である．これを実現する上で，単語の類義語・対義語関係を考慮した分散表現を得る必要があるが，K. A. Nguyen et al.(2016) は，Skip-gram モデルの目的関数に類義語・対義語関係を考慮した項を追加した上で目的関数を最適化することで，単語の類義語・対義語関係を考慮した分散表現を得ている．本研究でも，この手法をベースに LINE を拡張することで，単語の類義語・対義語関係を考慮した分散表現を得ることが出来，取得された極性語をポジティブ・ネガティブに振り分けた辞書を構築することが可能であると考える．

さらに第三点目として，ブートストラップ法を用いた場合に，反復を進めるにつれて極性語以外の単語が取得される意味ドリフトに対処する必要がある．意味ドリフトに対処する方法としては，各単語に対する信頼度を算出する関数の再設計や，ブートストラップ法と学習器を組み合わせる方法が考えられる．また，Curran et al. (2007) の研究に見られるように，意味ドリフトを起こす事が分かっている単語をストップクラスとして用意し，単語の抽出に制限をかける方法を用いる事が考えられる．

最後に第四点目として，本提案手法によって拡張された辞書を用いた場合と，拡張前の辞書を用いた場合とに分けた上で，金融政策テキストに対して伊藤他 (2017) の手法によってセンチメント分析を行い，拡張された辞書を用いた場合に，センチメント分析の精度が向上するかを検証する必要がある．拡張された辞書が正しい極性値のついた網羅的な辞書であるならば，その辞書を用いる事でセンチメント分析の精度が向上する事が期待される．さらに，センチメント分析の精度が向上しているのであれば，得られたセンチメントの値と市場期待やエコノミスト予想との関連性について検証を行いたい．

参考文献

[1] Bollen, J., and Huina, M. (2011) Twitter mood as a stock market predictor, *Computer* **44**: 91-94.

[2] Curran, J. R., Murphy, T., and Scholz, B. (2007) Minimising semantic drift with mutual exclusion bootstrapping, *In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*: 172-180.

[3] Jegadeesh, N., and Wu, D. (2015) Deciphering Fedspeak: The Information Content of FOMC Meetings, *2016 AFA Anual Meeting Working Paper* (https://www.aeaweb.org/conference/2016 / retrieve.php?pdfid=1136).

[4] Kamps, J., Marx, M., Mokken, R. J., and de Rijke, M. (2004) Using WordNet to Measure Semantic Orientations of Adjectives, *In Proceedings of the 4th International Conference on Language Resources and Evaluation*.

[5] Loughran, T. and McDonald, B. (2011) When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *Journal of Finance* **66(1)**: 35-65.

[6] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013) Efficient estimation of word representations in vector space, *arXiv preprint arXiv*: 1301.3781.

[7] Nguyen, K. A., Walde, S. S. I., and Vu, N. T. (2016) Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction, *Proceedings of 54th Annual Meeting of the Association for Computational Linguistics*.

[8] Pang, B., Lee, L., and Vaithyanathan, S. (2002) Thumbs up?: sentiment classification using machine learning techniques, *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing*: 79-86.

[9] Turney, P. (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics ACL'02, Association for Computational Linguistics, Stroudsburg*: 417-424.

[10] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015) Line: Large-scale information network embedding, *In Proceedings of the 24th International Conference on World Wide Web*: 1067-1077.

[11] 伊藤諒, 須田真太郎, 和泉潔 (2017) フォワードガイダンスの市場期待への影響分析 - テキストマイニング・アプローチ - 第 *46* 回 *2016* 年度冬季 *JAFEE* 大会: 60-71.

[12] 片倉賢治, 高橋大志 (2015) 金融市場ニュースの分散表現学習による辞書作成と金融市場分析 *2015* 年度人工知能学会全国大会（第 *29* 回） .

[13] 坪内孝太, 山下達雄 (2014) 株価掲示板データを用いたファイナンス用ポジネガ辞書の生成 *2014* 年度人工知能学会全国大会（第 *28* 回） .