単語の類義性・対義性を考慮したドメイン特化極性辞書構築

Domain-specific dictionary construction method considering synonym and antonym

伊藤　諒[1*]　坂地　泰紀[1]　和泉　潔[1]　須田　真太郎[2†]
Ryo Ito[1]　Sakaji Hiroki[1]　Kiyoshi Izumi[1]　Shintaro Suda[2]

[1] 東京大学大学院工学系研究科
[1] Graduate School of Engineering, The University of Tokyo
[2] 株式会社 三菱 UFJ トラスト投資工学研究所
[2] Mitsubishi UFJ Trust Investment Technology Institute Co.,Ltd.

Abstract: In recent years, textual information, which is unstructured data attracts attention as new analytical data in the financial and economic fields and it is expected to structure knowledge on this domain. One such knowledge is a sentiment polarity dictionary in which each word is representing positive or negative. In building the dictionary, it is costly to add the polarity value to a vast number of words manually. Therefore, in this research, we propose a the dictionary construction model especially considering the synonymity and symmetry of words. As a result of the experiment, the proposed method is a more accurate than the model of the previous research. In addition, we extended the conventional dictionary using the proposed method, and we showed that the extended dictionary has higher accuracy than the dictionary which is not extended.

1　はじめに

自然言語処理の根幹を支える資源として，語彙資源が存在し，これまでに語彙資源の構築に関する多くの研究がなされてきた [1]．語彙資源を構成するものとして，単語と極性が組みとなった極性辞書が存在し，このような極性辞書は，語彙ベースにおけるセンチメント分析を行う際に，不可欠なものである．

ここにおいて，極性辞書の構築を考えた際，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．

そこで本研究では，解析対象となるドメインに特化した，センチメント分析のための極性辞書を自動構築することを目的とし，とりわけ対象ドメインの知識，既存の知識ベースに含まれる単語の類義・対義性に関する知識を用いた，半教師あり学習による，ドメイン特化型極性辞書自動構築手法を提案する．また，提案手法のモデルを，既存の極性辞書に対する辞書構築精度の観点から評価を行う．さらに，提案手法を有用性を評価するために，対象ドメインを金融政策ドメインとし，本ドメインに対して人手で構築された辞書を拡張した場合に，センチメント分析の精度が向上するかという観点から評価を行う．

2　関連研究

極性辞書構築手法に関する多くの研究がなされているが，コーパスベースのアプローチと，シソーラスベースのアプローチに大別される．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．シソーラスベースのアプローチとしては，シソーラス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伝搬させる事で，全ての単語に対して，極性を付与する方法が代表的である．

さらに，コーパスベースのアプローチと，シソーラスベースのアプローチを統合した研究として，Alhothai and Hoey (2017) の研究がある [2]．Alhothai and Hoey

*連絡先：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
†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．

(2017) は，まず Skip-gram モデルまたは Glove モデルによって単語分散表現を学習する事で，単語分散表現から k-近傍グラフを構築し，次に得られた k-近傍グラフとシソーラスから構築された類義語ネットワークを合わせた上で，ネットワークに対してラベル拡散法を行う事で極性語を獲得する，SNWELP モデルを提案している．

しかしながら，Alhothai and Hoey (2017) の先行研究において，類義語に関する知識は用いられているが，対義語に関する知識は用いられていない．一般に，ある単語が極性語である際に，その単語に対する対義語は，元の単語とは反対の極性を有する場合が多いが，対義語に関する知識は，極性語獲得タスクにおいて重要な情報を含むため，単語間の類義性のみならず対義性も考慮する事で，より辞書構築精度が向上すると考える．

以上を踏まえ，対象コーパスに含まれるドメインの知識，既存の知識ベースに含まれる単語の類義・対義性に関する知識を用いた，半教師あり学習による，ドメイン特化型極性辞書自動構築手法を提案する．

3 ドメイン特化型極性辞書自動構築手法の提案

本章では，SMLS モデルと DLS モデルという，二つのドメイン特化型極性辞書自動構築手法の提案モデルについて述べる．

3.1 SMLS モデル

はじめに，SMLS モデルを提案する．SMLS モデルでは，はじめに対象コーパスのテキストに対して文分割をし，文分割されたセンテンスに対して形態素解析を行う．次に，単語分割された各センテンスを元に，Mikolov et al. (2013) による Skip-gram モデルを用いて単語分散表現を学習する [3]．そして，各単語に対して，類似度上位の単語 k 個をエッジで結んだ k-近傍グラフを構築する．ここで，単語 c と単語 d の分散表現をそれぞれ $\overrightarrow{u_c}$，$\overrightarrow{u_d}$ とした時，類似度をコサイン類似度とし，エッジの重みとしてコサイン類似度の値を付与する．

次に，得られた k-近傍グラフを元に，単語間がエッジで結ばれていれば，要素として，そのエッジ重みを，結ばれていなければ 0 を与えた，隣接行列 $\mathbf{M}$ を作成する．また，シソーラスから単語間の類義・対義関係を抽出し，単語間に類義関係もしくは対義関係が存在していればエッジで結んだ，類義語グラフ・対義語グラフをそれぞれ作成する．そして，得られた類義語グラフ・対義語グラフを元に，単語間がエッジで結ばれていれば 1 を，結ばれていなければ 0 を要素として格納した，隣接行列 $\mathbf{S}$，$\mathbf{A}$ をそれぞれ作成する．

ここで，分散表現から構築された隣接行列 $\mathbf{M}$ と，類義語・対義語グラフから構築された隣接行列 $\mathbf{S}$，$\mathbf{A}$ を結合した行列 $\mathbf{E}$ を作成する．なお，隣接行列 $\mathbf{E}$ の各要素 $E_{i,j}$ は，隣接行列 $\mathbf{M}$，$\mathbf{S}$，$\mathbf{A}$ の各要素を平均化した値とする．

$$E_{i,j} = \frac{M_{i,j} + S_{i,j} - A_{i,j}}{3}$$

次に，得られた行列 E に対して，シードを付与した上で，Zhou et al., (2004) によって提案されたラベル拡散法の手続きに基づき，ノードのラベル推定を行う [4]．

さて，V を行列 $\mathbf{E}$ の行数とした時，$\mathbf{p}$ を $\mathbf{p} \in \mathbb{R}^{|\mathbf{V}|}$ を満たす，単語の極性値ベクトルとする．ここで，極性値ベクトル $\mathbf{p}$ の要素は，$\frac{1}{|V|}$ で初期化されている．次に $\mathbf{D}$ を行列 $\mathbf{E}$ の次数行列とし，次数行列の各成分に絶対値をとった行列を $\mathbf{D}'$ とした際，以下の式に基づいて行列 $\mathbf{T}$ を計算する．

$$\mathbf{T} = \mathbf{D}'^{\frac{1}{2}}\mathbf{E}\mathbf{D}'^{\frac{1}{2}}$$

そして，得られた行列 $\mathbf{T}$ を用いて，以下の式を反復的に計算することで，ラベル拡散法を行う．

$$\mathbf{p}^{(t+1)} = \beta\mathbf{T}\mathbf{p}^{(t)} + (1-\beta)\mathbf{s}$$

ここで $\mathbf{s}$ は，$\mathbf{s} \in \mathbb{R}^{|\mathbf{V}|}$ を満たすベクトルであり，シードとして付与された単語に対応するベクトルの要素は $\frac{1}{|S|}$ を，シードとして付与されていない単語に対応するベクトルの要素は 0 を，要素として与えられたベクトルである．また，β は推定ラベルの，局所整合性・大域整合性を調整するパラメーターである．

そして，単語 w_i の推定極性値を得るために，ポジティブ単語．ネガティブ単語のシードセットを用いて，各々ラベル拡散法によって単語の極性値推定を行い，推定極性値 $\mathbf{P}^P(w_i)$ と $\mathbf{P}^N(w_i)$ を，それぞれ得る．さらに，得られた推定極性値を用い，以下の式によって単語 w_i の調整極性値 $\overline{\mathbf{P}}^P(w_i)$ を求める．

$$\overline{\mathbf{P}}^P(w_i) = \frac{\mathbf{P}^P(w_i)}{\mathbf{P}^P(w_i) + \mathbf{P}^N(w_i)}$$

最後に，得られた調整極性値 $\overline{\mathbf{P}}^P(w_i)$ を，各単語に対して，平均 0，分散 1 に標準化する．

3.2 DLS モデル

次に，DLS モデルを提案する．DLS モデルでは分散表現学習時に，コーパスに含まれるドメイン情報に加えて，類義・対義語関係の情報を分散表現として埋め込み，得られた分散表現を元に k-近傍グラフを作成し，ラベル拡散法によって単語の推定極性値を得る．

DLS モデルでは，はじめに K. A. Nguyen et al.(2016) によって提案された，d-LCE 法を用いて，コーパスにおける単語のドメイン情報と，単語の類義性・対義性に関する情報を埋め込んだ分散表現を学習する [5]．d-LCE 法は，Skip-gram モデルの目的関数に，単語の類義性・対義性に関する制約項を加えたモデルであり，d-LCE 法の目的関数は以下である．

$$\sum_{w\in V}\sum_{c\in V}\{(\#(w,c)\log\sigma(sim(w,c)) +k\#(w)P_0(c)\log\sigma(-sim(w,c))) +(\frac{1}{\#(w,u)}\sum_{u\in W(c)\cap S(w)} sim(w,u) -\frac{1}{\#(w,v)}\sum_{v\in W(c)\cap A(w)} sim(w,v))\}$$

ここで，V はコーパスに含まれる単語集合，$\#(w,c)$ は単語 w と単語 w に対するコンテキスト c との共起回数，k はネガティブサンプリングにおけるパラメーター値，P_0 はユニグラム分布，$sim(w_1,w_2)$ は単語 w_1 と w_2 のベクトル間のコサイン類似度，$W(c)$ はコンテキスト c に対する LMI 値が正の単語集合，$S(w)$・$A(w)$ は単語 w に対してシソーラスから抽出した類義語・対義語集合を表す．

次に d-LCE 法によって得られた分散表現を元に，SMLS モデルと同様に，各単語についてコサイン類似度上位 k 個の単語をエッジで結んだ k-近傍グラフを構築し，得られた k-近傍グラフを対象として，シードとして与えた単語の極性ラベルをラベル拡散法によって拡散する．

4 実験

本章では，提案手法の有効性を検証するための，各種実験設定と評価方法について述べる．実験は二通りの実験を行い，はじめに極性辞書構築実験を，次に極性辞書拡張実験を行った．

4.1 極性辞書構築実験

提案手法の優位性を検証するために，提案手法である SMLS モデル・DLS モデル，そして先行研究の手法である SNWELP モデルを用いて辞書構築を行い，辞書構築精度を比較した．ここでは，コーパスとして米国の株式公開企業において提出される Form 8-K を，シソーラスとして Wordnet を用い，これらを元に各種辞書構築手法を用いて辞書の自動構築を行い，ファイナンス分野のセンチメント分析において多く用いられている，Loughran and McDonald (2011) の辞書（以下 LM 辞書）中に含まれる各単語の極性の方向性を正解として，評価指標を AUC として評価した [6]．

各モデルにおいて，分散表現の次元数は 300，ネガティブサンプリングの負例数を 15，k-近傍グラフにおける k の値として 10 を用いた．また，コーパスとして用いる Form 8-K は，Lee et al. (2014) によって公開されているデータ[1]を用い，前処理の結果 20,198,170 センテンスを抽出し，各種モデルの入力データとした [7]．さらに単語の類義・対義関係を記述したシソーラスとして，Wordnet を用い，ラベル拡散法における反復処理の回数は 10,000，パラメーター β の値は 0.99 とした．また，ラベル拡散法におけるシード単語として，以下の表 1 に含まれるシード単語を用いた．

表 1: ラベル拡散法において用いたシード単語

Positive 単語	Negative 単語
successful, excellent, profit, beneficial, improving, improved, success, gains, positive	negligent, loss, volatile, wrong, losses, damages, bad, litigation, failure, down, negative

4.2 極性辞書拡張実験

提案手法の有用性を検証するために，金融政策ドメインを対象として人手で作成された辞書を対象に，提案手法である SMLS モデルを用いて辞書拡張を行い，辞書拡張の有無によって辞書ベースのセンチメント分析の精度が向上するかという観点から精度評価を行った．

コーパスとして，米国の金融政策を策定する委員会である Federal Open Market Committee（FOMC）[2]によって公開される議事録レビュー部分を用い，1993 年 1 月から 2016 年 12 月の間に公表された 192 件を対象とした．また辞書構築におけるシード単語として，人手によって作成された金融政策専門辞書を用いて，SMLS モデルによる辞書拡張を行った．この際，推定極性値が上位・下位 30 位以内に含まれ，かつ既存辞書に含まれず，名詞・動詞・形容詞の品詞に該当する単語を抽出し，既存辞書に追加をする事で辞書拡張を行った．また，SMLS モデルにおいて，分散表現の次元数は 300，ネガティブサンプリングの負例数を 15，k-近傍グラフにおける k の値として 10 を用いた．さらに，単語の類義・対義関係を記述したシソーラスとして，Wordnet を用い，ラベル拡散法における反復処理の回数は 10,000，パラメーター β の値は 0.99 とした．

次に，拡張された辞書を用いて伊藤ら (2017) の手法によって，経済成長・消費・生産・雇用・金融政策・金

[1] https://nlp.stanford.edu/pubs/stock-event.html
[2] https://www.federalreserve.gov

融市場・インフレ・貿易の 8 つのトピックに対する，トピック別センチメントの抽出を行った [8]．なお，伊藤ら (2017) のトピック別センチメントの抽出法は，トピック分類と，単語間の係り受け関係を考慮した辞書ベースのセンチメント分析を組み合わせた手法であり，文書を入力として，文書に含まれる各トピックに対するセンチメントスコアを算出する手法である．

ここで，FOMC 議事録におけるレビュー部分は，経済環境や金融市場の振り返りを行うセクションであるため，得られたトピック別センチメントが正確なものであれば，各トピックに対応するマクロ指標の実測値を，よく説明できるものと考えられる．そこで，辞書拡張の評価として，辞書拡張を行った場合と行わなかった場合とで，それぞれ算出されたトピック別センチメントの，各マクロ指標の実測値に対する説明力を以って評価を行った．説明力の算出においては，各トピックのセンチメントを説明変数，対応するマクロ指標を被説明変数として単回帰分析を行い，決定係数 R^2 による評価を行った．なお，レビューにおける各トピックのセンチメントを評価するマクロ指標は，以下の表 2 の対応となっている．

表 2: 各トピックに対応する検証用マクロ指標

トピック	マクロ変数
インフレ	インフレ率
雇用	非農業部門雇用者数
貿易	経常収支
消費	個人消費支出（PCE）
生産	鉱工業生産指数
経済成長	実質 GDP

5 結果と考察

本章では，各種実験の結果と考察について述べる．

5.1 極性辞書構築実験

表 3 は各モデルにおける，AUC の値を比較したものである．実験結果として，AUC の高い順に SMLS モデル，SNWELP モデル，DLS モデルとなり，提案手法である SMLS モデルが最も高い精度となった．

表 3: 各モデルにおける AUC の比較

SNWELP	DLS	SMLS
0.9096	0.8441	0.9190

これらの結果の理由として，まず SMLS モデルは SNWELP モデルと比較して高い AUC が得られているが，これは対義語の情報を考慮した上でラベル拡散を行なっているためだと考えられる．提案手法の章においても述べたように，ある単語が極性語である際に，その単語に対する対義語は， 元の単語とは反対の極性を有する場合が多いため，対義語の情報をモデルとして考慮することで，辞書構築の精度が向上したと考えられる．一方，DLS モデルでは他の 2 つのモデルと比較して，辞書構築の精度が低い結果となったが，これは分散表現の学習時において，単語の類義性・対義性に関する知識を十分に分散表現として埋め込む事が出来なかったためと考えられる．この点に関しては，d-LCE 法における，類義性・対義性に関する項の重要度を調整する事によって，より良い辞書構築精度を得る事ができると考える．

5.2 極性辞書拡張実験

SMLS モデルによる辞書拡張の結果，既存の辞書には含まれない 52 単語が抽出されたが，表 4 は，新たに辞書へ追加された単語の一部である．ポジティブ単語としては，gain・growth・lift などの，ポジティブな極性を有すると考えられる単語が追加され，一方ネガティブ単語としては，discourage・downgrade・depressing などの，ネガティブな極性を有すると考えられる単語が追加された．一方，substantial などのように，本来は極性を持たない単語も極性を持つ単語として，辞書に追加される結果となった．

表 4: SMLS によって，新たに辞書へ追加された単語の一部

Positive	Negative
gain, growth, foster, stability, accommodation, lift, substantial, strengthening	discourage, thin, cut, downgrade concerned, depressing, wan, lessening

図 1 は，辞書拡張の有無による，各マクロ変数の説明力の比較を示したものである．表から分かるように，辞書拡張の結果，鉱工業生産指数に対応する生産トピックを除いた全てのトピックにおいて，マクロ変数に対する説明力が向上するという結果が得られた．とりわけ，消費トピックに対応する PCE や，経済成長トピックに対応する GDP 成長率において，説明力を大きく向上させる事が出来ている．

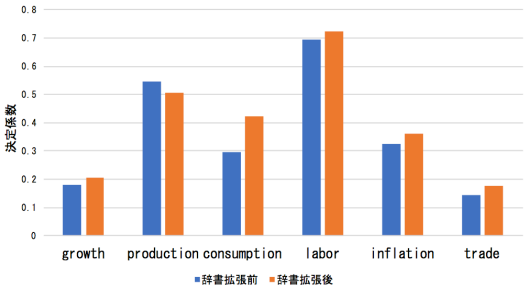

図 1: 辞書拡張の有無による，各マクロ変数に対する説明力の比較

辞書拡張によって精度が向上した理由として，既存の辞書には含まれていない極性語を獲得する事が出来，対象ドメインに対するより網羅性のある辞書を構築する事が出来たためと考えられる．一方で，本来極性語とは関係のない単語も極性語として追加されており，このような単語を機械的に排除する手法の検討や，さらなる辞書構築精度の向上は今後の課題である．

6 まとめ

本研究は，解析対象となるドメインに特化した，センチメント分析のための極性辞書を自動構築することを目的とする，ドメイン特化型極性辞書自動構築手法をした．提案手法である SMLS は，先行研究の SNWELP よりも辞書構築精度が上回る結果となり，提案手法の有効性が確認された．これは，ある単語が極性語である際に，その単語に対する対義語は， 元の単語とは反対の極性を有する場合が多いため，対義語の情報をモデルとして考慮することで，単語の極性値推定タスクにおいて，精度が向上したと考えられる．

また，SMLS を用いて既存の辞書を拡張した結果，拡張辞書を用いたトピック別センチメントのマクロ変数に対する説明力は一つのトピックを除き向上しており，提案手法のモデルの有効性が確認された．これは，辞書を拡張した結果，対象ドメインに対するより網羅性のある辞書を構築する事が出来たためと考えられる．

本研究の課題として，辞書構築精度向上のために，コパースの知識，単語の類義性・対義性に関する知識のみならず，センテンスの構文知識をモデルとして考慮する事が挙げられる．極性語どうしの文法的類似性から，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ており，このような構文情報・コーパスのドメイン知識・単語の類義対義性を同時に分散表現として埋め込むことで，極性語獲得タスクにおいてよりよい分散表現になると期待される．

参考文献

[1] Ding, Y., and Foo, S. (2002). Ontology research and development, *Part 1-a review of ontology generation. Journal of information science* **28(2)**: pp.123-136.

[2] Alhothali, A., and Hoey, J. (2017). Semi-Supervised Affective Meaning Lexicon Expansion Using Semantic and Distributed Word Representations, *arXiv preprint arXiv*:1703.09825.

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

[4] Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency, *In Advances in neural information processing systems*: pp.321-328.

[5] 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.*

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

[7] Lee, H., Surdeanu, M., MacCartney, B., and Jurafsky, D. (2014). On the Importance of Text Analysis for Stock Price Prediction, *In LREC*: pp.1170-1175.

[8] 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: pp.60-71.