The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

極性を考慮したリスク発見に向けた因果関係ネットワークの構築

Construction of Causal Networks for Risk Finding Considering Word Polarities

五十嵐 光秋 *1 Hiroaki Igarashi　坂地 泰紀 *1 Hiroki Sakaji　和泉 潔 *1 Kiyoshi Izumi　島田 尚 *1 Takashi Shimada　松島 裕康 *1 Hiroyasu Matsushima

須田 真太郎 *2 Shintaro Suda

*1東京大学 大学院工学系研究科 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 Mitsubishi UFJ Trust Investment Technology Institute Co., Ltd.(MTEC)

In this research, we conduct an experiment of constructing a causal networks within settlement account briefs. We extracted the causal relations from settlement briefs, and construct a network by connecting them by judging similarities. For calculating the similarities, we use a word2vec model created from the Japanese Wikipedia corpus.We use a method based on a combination of idf values which representing importance of words. In addition, by giving the polarities of causal expression using a polar dictionary, and judgment of synonyms that word2vec can't detect, we define so to speak, "negative relation".

100 edges between causal relationships obtained by experiments were randomly selected and evaluated visually. As a result, the ratio that is deemed to be a reasonable connection was 84%, and among them the ratio of edges where polarity inversion was correctly captured was 86%.

1. はじめに

近年，人工知能分野の技術の発展に注目が集まっており，これを金融領域に対しても応用することが期待されている．経済事象は複雑に絡み合い，様々なことが原因となり発生する．これを深く理解するには，様々な社会事象の因果関係を読み解いていく必要がある．こうした背景の中で，テキストデータの活用に対して注目が集まっている．複雑な因果関係を数値データだけで読み解くのは難しいが，人間の理解力をもって因果関係と解釈された関係性が，テキスト情報には蓄えられているからである．

社会事象の因果関係を読み解くことによるメリットの 1 つとして，あるニュースから企業の業績を予測することができることがある．これにより，例えば投資家はいち早く投資の意思決定に役立てることができるし，また企業自身としても業績予測に応じた対応策を講じることができるであろう．

世界中に衝撃を与えた社会事象として，リーマン・ショックは記憶に新しい．サブプライムローンというリスクの高い金融商品が出回っていたこと，信用の高い格付け企業がこの金融商品に高い評価を出していたこと，不動産価格の暴落など，様々な問題が密接に結びついて発生したと言われるリーマン・ショックであるが，因果関係を適切に把握することでこのようなリスクを事前に察知することができると考えられる．もちろん，リーマンブラザーズだけにリーマン・ショックの責任があるのではない．リスクの高い状態にあった金融市場に対して，投資家や証券会社等の様々なステークホルダーが楽観的すぎたことが最大の要因であると考える．したがって，社会全体として，リスク要因となるような因果関係を把握していくことが，健全な社会への第一歩となる．

とはいうものの，こうした複雑な因果関係は単一の文章に記載されているものではなく，膨大な数のテキストデータを網羅し，因果関係をつなぎ合わせることで初めて読み解くことができる．したがって，膨大なテキストから因果関係のネットワークを構築して，リスク事象を発見することに大きな価値がある．

この課題を解決するアプローチとして、テキストマイニングによる因果関係を抽出する研究，そして抽出した因果関係を元にそれらのネットワークを構築する研究がなされてきた．乾ら [乾 04] は，接続標識の「ため」に注目して，これが含まれる表層的な因果関係を抽出した．ここでは，事象を「事態」と「行為」に分けて捉え，抽出した因果関係を cause 関係，effect 関係，precond 関係，means 関係の 4 つに分類した．坂地ら [坂地 11] の研究では，乾らと同様の概念である「手がかり表現」を鍵として新聞記事から因果関係を抽出した．そしてこれを機械学習の手法の 1 つである SVM（サポートベクターマシン）によって分類し精度を高めた．これらの研究はあくまで単一の記事からの抽出にとどまり，因果関係のネットワークを作成するには至らなかった．そこで，佐藤ら [佐藤 06] は，Web 上の文書から因果関係ネットワークを構築することを試みた．ここでは因果関係表現のキーワードを抽出し，その一致率から検討していた．石井らは [石井 09]SVO 構造に注目した単語の一致による因果関係の接続について検討した．この手法では，それぞれ主語，動詞，目的語が一致した場合に限り因果関係の接続を行うという制約のもとで接続を行なっている．しかし，主語，動詞，目的語のセットというシンプルなフォーマット以外での表現による因果関係表現も存在し，本来的には接続されるべきであった因果関係が抜け落ちてしまう可能性がある．表現方法による制限をかけずに因果関係を接続するという考え方から，word2vec モデルの利用を検討した研究も存在する．西村ら [NISHIMURA 18] は，新聞記事から抽出した因果関係表現に対して，word2vec モデルを用いて因果表現のベクトル表現を算出し，コサイン類似度の計算から因果関係の接続を試みた．ただ，ここでは単語ベクトルの扱いの処理や品詞の選定などに改善点が見られた．

以上のような背景から、我々は word2vec モデルを用いた因果関係ネットワークを構築する．これにより，企業の業績悪化

連絡先: 五十嵐光秋，東京大学大学院工学系研究科，東京都文京区本郷 7-3-1，m2018higarashi@socsim.org

などに対してその根拠となるような潜在的な事象を発見することを目指す．

2. 因果関係ネットワーク構築手法

文書集合の中から因果関係を抽出し，それらを接続する手続きについて説明する．本研究では word2vec のモデルと単語の重要度を用いて，因果関係ネットワークを構築する．図 1 に本手法の概要を示す．

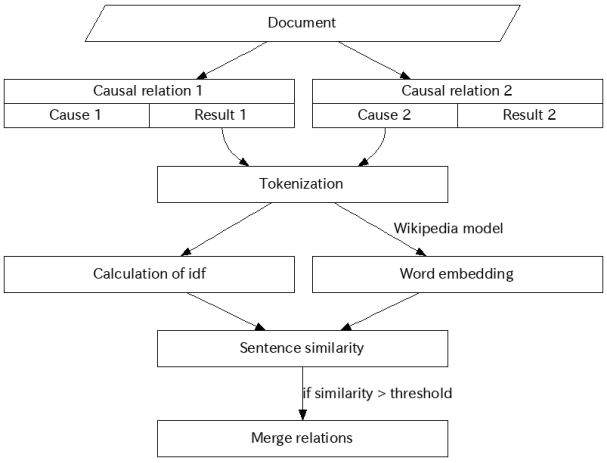

図 1: 本手法のフレームワーク

2.1 因果関係の抽出

決算短信からの因果関係の抽出には，坂地らの手法を用いた．[坂地 11] これにより，重文や複文にまたがって記述されるような因果関係を抽出することができる．詳しいアルゴリズムについては引用論文を参照されたい．この手法により決算短信から抽出した因果関係の例を以下に記述する．

原因表現	手がかり表現	結果表現
海外経済の改善	を背景に，	輸出も持ち直しの動きが見られています。
物流及び生産機能の低下	により，	当社の業績は影響を受ける可能性があります。

表 1: 決算短信から抽出した因果関係の一覧

2.2 因果表現から単語集合への変換

本手法では，単語をベクトルで表現することで，文を表現するベクトルを得る．日本語では，単語と単語が必ずスペースで区切られている英語などの言語とは異なり，単語が全てそのまま接続されて文をなす構造をしているために，これを単語集合へと変換する作業が必要である．この分かち書きの作業を，形態素解析ツール MeCab[*1] を用いて行った．辞書には，新語などに対応している mecab-ipadic-NEologd[*2] を採用した．

自然言語から単語集合へ変換する際に，品詞による単語の選定を行なった．有効とした品詞とその詳細は表 2 に示す．

2.3 因果関係の接続

複数の因果関係を接続することで，因果関係の連鎖を構築することを試みる．ある因果関係 $A \rightarrow B$ に対して，また別の因果関係 $B' \rightarrow C$ が存在する状況を想定する．$A \rightarrow B$ の関係性においては，A が原因表現，B を結果表現と呼ぶ．この時，B と B' が同一の事象について言及しているのであれば，これらを接続することで $A \rightarrow B \rightarrow C$ という連鎖が確立される．この因果関係の上流側には，企業の業績の要因となるような事象が現れることが期待される．したがって，文書群に対して因果関係の抽出，接続を行い，これらを遡ることで単一の文書だけでは発見できなかった潜在的な事象を発見することを目指す．

品詞	詳細
名詞	サ変接続，一般，固有名詞，形容動詞語幹，副詞可能
動詞	自立
形容詞	自立

表 2: 決算短信から抽出した因果関係の一覧

2.4 word2vec モデル

前節で述べた手続きを行う上で，結果表現 B と原因表現 B' の類似度を評価する必要がある．この時の文間類似度を算出する手続きでは，word2vec モデルを用いる．word2vec とは，「文書集合において，類似した単語は同一の文脈において現れる」という理解に基づいて構築したニューラルネットワークモデルを介して作成した，単語のベクトル表現である．従来の自然言語処理においては，テキストに登場する各単語を表現する方法が語彙数次元の onehot ベクトル（文書集合内での登場回数を用いる tf-idf という考え方もあるが，依然として次数は語彙数に一致する）であり，文書のサイズが大きくなるほどに計算量が膨大になってしまうという問題点があった．しかしこの word2vec モデルで学習することにより，単語同士の意味的な繋がりへの表現力に強く，かつ低次元なベクトル表現の実装が可能になった．本手法では，この word2vec による単語のベクトル表現を拡張して，文に対するベクトル表現を獲得する．なお，モデルの中に含まれない単語に関しては，全ての成分が 0 のベクトルにより代用した．

2.5 idf 値の算出

例えば「する」,「いる」のような語は，非常に多種の単語との親和性が高く，多くの文書内で登場する．今回のような文の意味的類似性を測る上で，このような単語のマッチングがノイズとなることが予想される．したがって，複数の単語から構成される文に対してベクトル表現を獲得するにあたってこのような重要度の低い単語の影響は小さく抑え，かつ重要度の高い単語は大きい影響を持つように調整することが必要である．

この点を考慮するために，本手法では各単語の idf 値を単語ベクトルに対する重みとして扱うこととする．この idf 値は以下の式により与えられる．

$$idf_i = log_2 \frac{D}{freq_i} \tag{1}$$

ここで，D は全文数，$freq_i$ は単語 i の出現する文数である．

各単語に対する idf 値を対象の決算短信データから算出した．各単語の idf 値の算出結果について，下位 10 単語を表 3 にまとめる．

決算短信で頻繁に現れる単語の idf 値が，確かに下位に現れていることが確認できる．

2.6 文間類似度の算出

前々節で述べた単語ベクトルと前節で述べた idf 値を用いて文ベクトルを獲得する．文 s_j を単語集合に変換する手続きにより，そこに含まれる全ての単語に対して単語ベクトルと idf 値が獲得できる．それらを用いて，以下のように文ベクトルを定める．ただし，一文中に登場する同一の単語は，登場する回

∗1 http://taku910.github.io/mecab/
∗2 https://github.com/neologd/mecab-unidic-neologd

単語	idf 値
する	0.96
なる	2.429
四半期	2.874
ある	2.93
当社	3.19
連結会計	2.93
事業	3.26
平成	3.315
増加	3.352
セグメント	3.384

表 3: idf 値の下位 10 項目

数だけ重複して加算される．

$$vec_{s_j} = \sum_i vec_i \cdot idf_i \tag{2}$$

ここで，vec_i と idf_i は，それぞれ文 s_j 内に含まれる単語 i に相当する単語ベクトルと idf 値である．

文間類似度に関しては，以上の手続きによって獲得された文ベクトルのコサイン類似度を採択した．

$$sim(s_1, s_2) = \frac{vec_{s_1} \cdot vec_{s_2}}{|vec_{s_1}||vec_{s_2}|} \tag{3}$$

算出された類似度が閾値を超える因果関係のペアを接続した．

2.7 極性辞書の利用

類似した文を探索するタスクにおいて，word2vec モデルでは対義語と類義語の区別ができないため欠点となる．これは，word2vec モデル構築の際に設定している「類似単語が同一の文脈において登場するという」仮定によるものであり，例えば「増加」と「減少」は類似単語として学習されてしまう．それにより，本来真逆の概念として捉えたい対義語が，類義語であるかのように扱われてしまう．

この問題に対するアプローチとして，我々は単語の極性を用いた．極性とは，単語が示す概念がポジティブなのかネガティブなのかについて定量評価をしたものである．単語に極性を付与するにあたって，本手法では伊藤ら [Ito 18] による金融領域に特化した極性辞書を用いた．伊藤らによる極性辞書では，金融分野におけるコーパスから極性を獲得しているため，決算短信内に現れる単語に対して高精度で適切な極性を与えることが可能であると考える．本手法では，単語ごとに付与されている極性の総和を取り，文に対する極性として定義する．類似度と極性の正負に注目することにより，対義語により誤って類似文と判定された文を特定することができる．

この判定によって，結果表現 $A \rightarrow$ 原因表現 B の極性の反転が認められた場合，A を含む事象の発生に対して B を含む事象が「起こらない」という関係性が得られたことに相当する．したがって，次章で示す結果ではあえてこの極性判定によるエッジの削除は行わず，エッジの種類を変えることでそれを表現することとする．

3. 因果関係ネットワーク構築実験

3.1 実験仕様

ある事象が影響する企業の業績に焦点を当てるために，決算短信を対象として因果関係を抽出した．決算短信には，各企業の売上や利益といった数値情報の他に，その利益の原因となった事象や経営上のリスクに関する記述などが記載されている．2012 年 10 月 9 日から 2018 年 5 月 11 日までに発行された，4359 企業の決算短信データから因果関係を抽出したものを対象に分析を行った．これらの決算短信から，前章に述べた idf 値の獲得と因果関係の抽出，因果関係ネットワークの構築を行った．

また，単語のベクトル表現に関しては，乾らによる Wikipedia の日本語コーパスを用いて作成した「日本語 Wikipedia エンティティベクトル」[*3] を用いた．このモデルでは，通常の word2vec の学習に加えて，Wikipedia の記事になった固有表現に付される記事本文中のハイパーリンクを用いて学習している．同一表記であるが違うものを指す言葉（例えば魚の「スズキ」と四輪・二輪車メーカーの「スズキ」など）の違いを反映することができる．

本実験では「国際情勢への不安」を起点とした因果関係ネットワークを構築した．因果関係接続のための閾値を 0.9 で定め，起点事象から広がる因果関係の波及を 2 段階目まで確認した．

3.2 実験結果

坂地ら [坂地 11] の手法により抽出された因果関係の数は 92 万 5908 件であった．これらの候補の中から前節で述べた条件で探索をし，得られた因果関係ネットワークの一部を図 2 に示す．

算出された類似度をエッジの近傍に付した．また，極性の反転が認められたエッジは点線により表現している．構築したネットワークのうちランダムに 100 件を抽出し，結ばれたエッジの妥当性を目視により検証した．目視による判定は，「反転なし」，「反転あり」，「因果関係なし」の 3 種類である．表 4 にその結果を示す．

表 4: 因果関係エッジの評価

		目視による判定		
		反転なし	反転あり	因果関係なし
実験結果	反転なし	42	2	11
	反転あり	9	31	5

抽出された 100 件の因果関係エッジのうち，妥当なつながりであると認められる割合は 84%，それらの中で，正しく反転を捉えられた割合は $(42 + 31)/84 = 0.86$ より 86%であった．

4. 考察

因果関係が正しく結ばれるケースのほぼ全てで，その事象を表す核となる単語が一致している（例えば「原材料価格の上昇」における「原材料」，「上昇」など）．逆に，正しく結ばれないケースでは文章中でさほど重要でない単語が一致していた（例えば「円相場は１０１円台まで円高が進行しました。」における「進行」など）．したがって，因果関係を適切に接続するには文におけるキーワードを適切に抽出することが重要であると言える．

また，Wikipedia のモデルには収録されていない，決算の専門用語（「繰延税金資産」など）がキーワードとなっているケースが存在し，Recall を下げている可能性が考えられる．こうしたベクトルは 0 ベクトルとして置換するため，素性をつかむことができない．これに対しては，決算短信を学習させた word2vec モデルを用いた場合，決算短信に固有な単語もカ

*3 http://taku910.github.io/mecab/

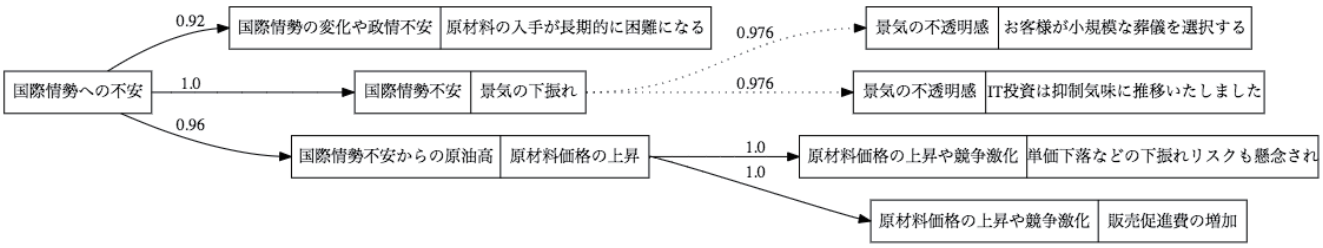

図 2: 「国際情勢への不安」を起点とした因果関係ネットワーク （可読性のため，一部を省略している）

バーすることができる．本研究で利用した Wikipedia によるモデルに対して，決算短信のデータを追加学習させることで，単語のカバー範囲を広げることが有効であると考えられる．

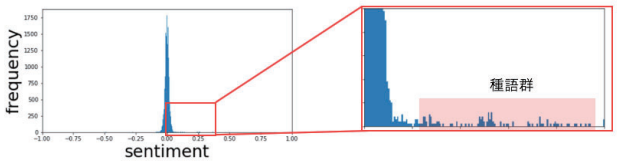

図 3: 極性辞書内の極性の分布

図 3 に示したように，極性辞書の中に収録されている単語の極性の分布は，極めて多くの単語が 0 近傍に集中しており，ごく少数の単語がそれより外側に位置するという構成になっている．この非連続性は，極性辞書の作成時の手順による．伊藤ら [Ito 18] の手法では，あらかじめ人手によりリストアップされた極めて強い極性をもつ単語（「種語」という）を起点に，ニューラルネットワークモデルによりその他の単語へ伝搬させている．そのため，これら種語群とその他の単語群では極性が大きく異なる．これにより，「下振れ」のような，種語として設定されていないが十分に強い極性を示す単語が適切に処理されなかったと考える．こうした問題に対しては，石井ら [石井 09] の手法のように，単語のシソーラスを用いて同概念を示す類似表現に置き換えることが有効である可能性がある．

5. まとめ

本研究では，決算短信内での因果関係ネットワークの構築実験を行った．

決算短信内に存在する因果関係を抽出し，これらの類似度を判定して接続することによりネットワークを構築した．類似度の算出には，Wikipedia 日本語コーパスから作成した word2vec モデルと，単語の重要度を表す idf 値による組み合わせによる手法を用いた．また，極性辞書を用いて因果表現の極性を付与することにより，word2vec が苦手とする対義語の判定を行い，いわば「負の関係性」を特定した．

この手順により得られた因果関係間のエッジから 100 件について目視で判定したところ，妥当なつながりであると認められる割合は 84%，またそれらの中で，正しく反転を捉えられた割合は 86%であった．

上述のアルゴリズムによって作成された因果関係の連鎖について，目視による判定を導入して正解データを集積したのち，機械学習によるフィルタリングを導入することにより Precision の向上が期待される．

また評価方法に関しても改善が必要である．本研究では，インプットした事象に対して，因果関係ネットワークの終端に企業の業績に関する表現が現れることは確認したが，これが実社会において妥当な結果なのかは不明である．外部の数値データを正解として，実際にそのような因果関係が働いているのかを確認することが重要だと考える．

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．

参考文献

[Ito 18] Ito, T., Sakaji, H., Tsubouchi, K., Izumi, K., and Yamashita, T.: Text-Visualizing Neural Network Model: Understanding Online Financial Textual Data, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 247–259Springer (2018)

[NISHIMURA 18] NISHIMURA, K., SAKAJI, H., and IZUMI, K. : Creation of Causal Relation Network using Semantic Similarity, 人工知能学会全国大会論文集 2018 年度人工知能学会全国大会 (第 32 回) 論文集, pp. 1P104–1P104 一般社団法人人工知能学会 (2018)

[乾 04] 乾 孝司, 乾 健太郎, 松本 裕治 他 : 接続標識 「ため」 に基づく文書集合からの因果関係知識の自動獲得, 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–933 (2004)

[佐藤 06] 佐藤 岳文, 堀田 昌英 : Web マイニングを用いた因果ネットワークの自動構築手法の開発, 社会技術研究論文集, Vol. 4, pp. 66–74 (2006)

[坂地 11] 坂地 泰紀, 増山 繁 : 新聞記事からの因果関係を含む文の抽出手法, 電子情報通信学会論文誌 D, Vol. 94, No. 8, pp. 1496–1506 (2011)

[石井 09] 石井 裕志, 馬 強, 吉川 正俊 他 : SVO 構造を用いた因果関係ネットワーク構築手法について, 研究報告データベースシステム (DBS), Vol. 2009, No. 10, pp. 1–8 (2009)