ベクトル表現を用いた因果関係連鎖の抽出

Extraction of Causal Relation Chains using Vector Expressions

西村弘平 [1] *　坂地泰紀 [2] 和泉潔 [2]

[1] 東京大学工学部システム創成学科システムデザイン&マネジメントコース
[1] Department of Systems Innovation, Faculty of Engineering, The University of Tokyo
[2] 東京大学大学院工学系研究科
[2] Graduate School of Engineering, The University of Tokyo

Abstract: 複数のテキストデータから経済・金融事象を背景知識まで含めて可視化することは, 経済・金融事象の理解の助けになり有用である. しかしながら, 経済・金融事象の連鎖を手動で抽出することは非常に時間とコストがかかる. そこで, 本研究では経済・金融事象の連鎖を因果関係として扱い, 各事象を表したベクトル間の類似度を用いて因果関係の連鎖を構築する手法を提案する. また, 提案手法における問題点から今後の提案をまとめる.

1　はじめに

近年, 人工知能分野の手法や技術の金融市場の様々な場面への応用が期待されており, 膨大な金融情報を分析して投資・経営判断を支援する技術が注目されている. 特に, 投資家・経営者は経営・金融事象の情報を把握し, 各事象を正しく理解して投資や経営の判断をする必要があるため, 複数のテキストデータから抽出した因果関係を用いて因果ネットワークを構築し, 事象を背景知識とともに可視化する技術は有用である. 本論文では上記の背景を踏まえ, テキストデータから抽出した因果関係ノード間の原因表現と結果表現の表現類似度を測ることによってテキストデータから因果関係の連鎖を構築する手法を提案する.

2　先行研究

本章では, 提案手法に関連する因果関係抽出・因果関係連鎖構築の先行研究について述べる.

2.1　因果関係抽出

Khoo et al.[1], 乾ら [2] は接続関係や格フレームを用いて因果関係を抽出する手法を提案している. これらの手法は単文もしくは複文・重文からしか因果関係を抽出できないという問題点がある. 坂地ら [3] は文書中にある因果関係を, 手がかり表現を用いて抽出する手法を提案している. 坂地らの手法は因果関係が存在する構文パターンを列挙し, 手がかり表現を用いることによって単文, 複文・重文関係なく因果関係にある表現を抽出することができる. 本提案手法では坂地らの手法を利用して, 因果関係を抽出する. また, 本論文での因果関係の定義を坂地らと同様に, 「原因若しくは, 理由と結果を示し, 手がかり表現を伴って 1 文中, もしくは隣り合う 2 文中に表層的に出現するもの」とする.

2.2　因果関係連鎖構築

Ishii et al.[4] は SVO の構造と WordNet を用いてニュース記事から因果関係の連鎖を構築する手法を提案している. Ishii et al. は SVO 構造を利用することで注目する単語を決め, WordNet を用いた単語のマッチングによって概念的な単語の類似度を計算している. 津川ら [5], [6] は WebAPI を用いて要因結果検索を用いて, 事象間の共起度を測定し因果関係の連鎖を構築している. 津川らの WebAPI を用いる手法は検索エンジンのアルゴリズムによって影響を受けるため, 手法の再現性が低いと考えて, 比較手法には Ishii et al. の手法を用いた.

3　提案手法

本章では, テキストデータから因果関係の連鎖を構築する手法について述べる. 3.1 節から 3.4 節までで因果関係連鎖構築の概要を述べ, 3.5 節で因果関係連鎖構築に用いる因果ノード間類似度計算手法について述べる. 手法の概要は図 1 の通りである.

*東京大学工学部システム創成学科システムデザイン&マネジメントコース
〒 113 － 8656 東京都文京区本郷 7 － 3 － 1 工学部 8 号館 530 室
E-mail: b2017knishimura@socsim.org

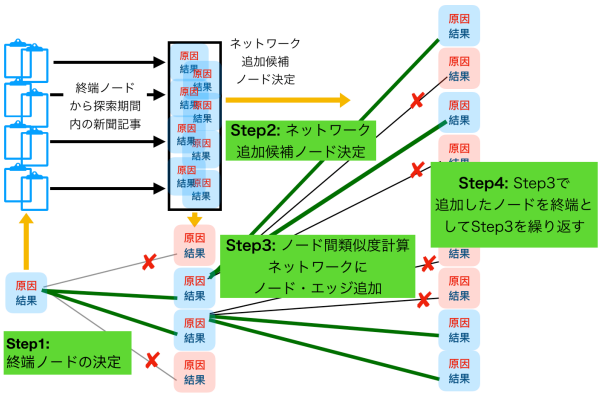

図 1: 因果関係連鎖構築手法の概略

3.1 因果関係連鎖の終端ノードの決定

決算短信のテキストから坂地ら [3] の手法を用いて因果関係表現を抽出する．抽出した因果関係ノードの中から市場の情勢や企業の業績を記した因果関係ノードを手動で選択し, 因果関係連鎖の終端ノードとする．

3.2 探索対象の因果関係ノードの探索

3.1 節で定めた終端ノードを終端とする因果関係連鎖への追加候補の因果関係ノードを日本経済新聞のテキストから抽出する．

因果関係抽出は 3.1 節と同じく坂地ら [3] の手法を用いる．3.1 節で定めた終端ノードよりも過去のもので、終端ノードから探索期間 S 以内の因果関係を因果関係連鎖の追加候補とする．さらに坂地らの手法で抽出した因果関係ノードのうち, 原因・結果表現の名詞・形容詞・動詞の単語数の合計がともに閾値 β 以上である因果関係ノードのみを因果関係連鎖の追加候補とした．

3.3 因果関係連鎖へのエッジ・ノードの追加

3.2 節で抽出した因果関係連鎖への追加候補の全ての因果関係ノードと終端ノードの組み合わせについて因果関係ノード間の類似度を計算する. 因果関係ノード間の類似度が閾値 α 以上であるときにノードを追加して因果関係連鎖を拡張する．

3.4 因果関係連鎖の更新

3.2 節で抽出した因果関係ノードを因果関係連鎖への追加候補ノード, 3.3 節で因果関係連鎖に追加したノードを終端ノードとして 3.2 節, 3.3 節の処理を $n-1$ 回繰り返す．ここで, n はあらかじめ定めた因果関係連鎖の更新回数である．

3.5 因果関係ノード間の表現類似度計算手法

本節では, 提案手法で用いている因果関係ノード間の計算方法について述べる．計算手法の概要は図 2 の通りである．因果関係ノード間の類似度は 2 種類の類似度を足し合わせることによって計算する．1 つは IDF 値を用いて作成したベクトル間のコサイン類似度で, もう 1 つは Bojanoski et al.[7] の FastText から求めた単語分散表現を用いた類似度である．IDF 値を用いたベクトル表現を用いることによって, 因果関係が含まれている領域やドメインなどのトピック情報の類似度を計算し, FastText から得た分散表現からなるベクトル表現を用いることで比較する因果関係の原因・結果表現間の表層的な類似度を計算する．ベクトル間類似度はコサイン類似度を 0 から 1 の値取るように正規化したものとした．

$$cosine\_similarity = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \quad (1)$$

$$vector\_similarity = \frac{(cosine\_similarity + 1)}{2} \quad (2)$$

IDF 値を用いた類似度は, 因果関係連鎖への追加対象のノードの結果表現と終端ノードの原因表現から IDF 値上位 3 つの単語を抽出し, IDF 値を用いてベクトルを作成しベクトル間の類似度を計算する．IDF 値は 1998 年から 2016 年までの全ての原因・結果表現から抽出した．FastText を用いた類似度は, 因果関係連鎖への追加対象ノードの結果表現と終端ノードの原因表現内にある名詞・動詞・形容詞の分散表現を足し合わせてベクトルを作成し, ベクトル間類似度を計算する．FastText の類似度は cbow と skip-gram の 2 つのアルゴリズムに対して計算した．IDF, FastText の合計 3 つのベクトル類似度の平均を取ることによってノード間の類似度とした．

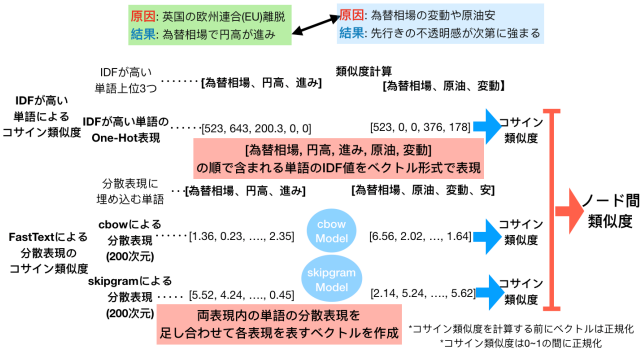

図 2: 因果関係ノード間類似度の計算方法

4 実験

実験には 1982 年から 2016 年までの 54206 件の決算短信テキストと 1998 年から 2016 年までの 6921 日分の記事テキストを用いた.

決算短信テキストから抽出した因果関係の中から 5 つの終端ノードを手動で選び, 日本経済新聞テキストから因果関係連鎖への追加候補ノードを抽出して因果関係の連鎖を構築する実験を行なった. 比較手法に Ishii et al.[4] の WordNet を用いたノード間類似度を計算する手法を用いた. 終端ノードは表 4 実験のパラメータは表 3 の通りである. 比較手法はノード間類似度の値が 0, 0.33, 0.66, 1 のいずれかなので 0.65 のみ実験を行なった.

S, 因果関係連鎖への追加候補ノード探索期間: 52 週間

n, 因果関係連鎖の更新回数: 4

α, 因果関係連鎖にマージするときのノード間類似度の閾値: 0.65, 0.70, 0.75

β, 因果関係連鎖への追加候補ノードの原因・結果表現の単語数についての閾値: 5

図 3: 因果関係連鎖構築に用いたパラメータ

5 実験結果と考察

手法の評価指標には Precision(精度) とノード数を用いた. 1 日の新聞テキストから約 300 件の因果関係ノードが抽出されるため、正解データを作成できず再現率の代わりにノード数を評価指標に加えた.

Precision の計算方法は, 構築した因果関係連鎖に対してランダムに 100 個のノードを抽出し, ノード間関係が「経済・金融事象を背景知識まで理解するために適切なつながりであれば正しい, そうでなければ誤り」という定義に従って, 因果関係の連鎖として適切か手動で判断し評価を行なった. 各ノード間関係の評価は著者 1 人のみが行なった.

Precision(精度), ノード数をそれぞれ表 1, 表 2 に記す. 手法名は比較手法, 提案手法をそれぞれ比較, 提案と記している.

精度, ノード数がともに高い手法がよりノード間の類似度をより正確に計算できていると言える. 提案手法では, 比較手法に比べて精度では全てのノードについて, ノード数にもおいても 3 つの終端ノードに対して比較手法よりも多くノード数を抽出できており, 提案手法が比較手法よりも精度高くノード間の類似度を計算できていると言える.

ID1: [**原因**] **為替相場の変動や原油安** [**結果**] **先行きの不透明感が次第に強まる, 中央魚類株式会社, 2016-11-2**

ID2: [**原因**] **海外経済の減速懸念や個人消費の低迷, 資源価格安の長期化,** [**結果**] **景気の先行きに不透明感が出てまいりました, 矢作建築工業株式会社, 2016-5-9**

ID3: [**原因**]**3 中国や韓国を中心に全世界で鉄鋼生産能力増強が進行し, 過剰な生産設備による供給過剰問題が顕在化する,** [**結果**] **世界的な鉄鋼需給バランスが大きく崩れた, 合同製鐵株式会社, 2016-4-28**

ID4: [**原因**] **米国のゼロ金利政策解除による金融市場の変動, 中国経済の減速, 原油価格の下落などの影響,** [**結果**] **先行きが不透明な状況で推移しました, クエスト株式会社, 2016-4-1**

ID5: [**原因**]**EU 情勢不安や中国経済減速の影響,** [**結果**] **売り上げが減少しています, 石塚硝子株式会社, 2016-4-25**

図 4: 終端ノードに用いた因果関係ノード

表 1: 各実験における精度の差

手法 (α)	ID1	ID2	ID3	ID4	ID5
比較 (0.65)	0.00	0.02	0.03	0.00	0.00
提案 (0.65)	**0.12**	**0.08**	**0.10**	**0.11**	**0.05**
提案 (0.70)	0.39	0.44	0.39	0.46	0.30
提案 (0.75)	1.00	0.48	0.83	0.56	0.60

表 2: 各実験におけるノード数の差

手法 (α)	ID1	ID2	ID3	ID4	ID5
比較 (0.65)	**82392**	24779	**1970587**	56027	11073
提案 (0.65)	80323	**97814**	90978	**100273**	**46429**
提案 (0.70)	44	424	41	288	115
提案 (0.75)	2	26	6	70	10

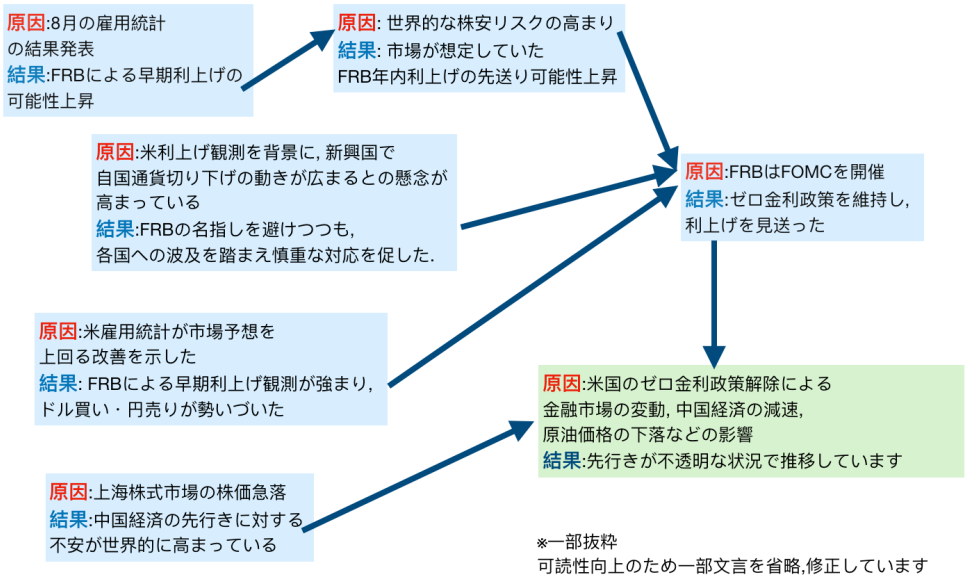

図 5: 構築した因果関係連鎖の例

終端ノードを ID5 としたときの具体的な因果関係連鎖の構築事例から抜粋したものを図 5 に示す.

構築した因果関係では, 「米国のゼロ金利政策解除による金融市場の変動, 中国経済の減速, 原油価格の下落」という経済事象の原因となったアメリカの金利政策の流れ, 上海市場の動向を因果関係の連鎖として取得できていることがわかる.

6 まとめ

本研究では, ベクトル表現の類似度を用いて決算短信と日本経済新聞のテキストから因果関係連鎖の構築手法を提案した. また, 提案手法が WordNet を用いた既存手法よりも精度高く因果関係間の類似度を計算できることを実験で確認した. 提案手法のベクトル表現は WordNet といった事前知識を必要としないため, 新たな単語を含むテキストからでも因果関係の連鎖を構築することができ有用である.

今後の課題としては類似だけでなく, 包含・例示・時系列推移といった類似以外の事象間関係を区別して因果関係の連鎖を構築すること, また, 因果表現のみだと経済・金融事象の情報量が少ないことがあるため, 周辺テキストから精度高く経済・金融事象の情報を抽出することの 2 点が挙げられる.

参考文献

[1] C. S. Khoo, S. Chan, and Y. Niu, "Extracting causal knowledge from a medical database using graphical patterns," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 336–343, Association for Computational Linguistics, 2000.

[2] 乾孝司, 乾健太郎, 松本裕治, *et al.*, "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.

[3] 坂地泰紀, 酒井浩之, and 増山繁, "決算短信 pdf からの原因・結果表現の抽出," **電子情報通信学会論文誌** *D*, vol. 98, no. 5, pp. 811–822, 2015.

[4] H. Ishii, Q. Ma, and M. Yoshikawa, "Incremental construction of causal network from news articles," *Journal of information processing*, vol. 20, no. 1, pp. 207–215, 2012.

[5] 津川敦朗, 新妻弘崇, and 太田学, "交絡事象の発見による因果関係ネットワークの改良." DEIM Forum, E2-3, 2015.

[6] 津川敦朗, 新妻弘崇, and 太田学, "共起関係に着目した因果関係ネットワークの拡張." DEIM Forum, F3-2, 2016.

[7] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.