様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ４ 月 ２５ 日現在

機関番号：１０１０１
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１２０１０

研究課題名（和文）経済シナリオ分析のための因果関係インスタンス認識技術の確立

研究課題名（英文）Establishment of causal instance recognition techniques for economic scenario analysis

研究代表者
坂地　泰紀（Sakaji, Hiroki）

北海道大学・情報科学研究院・准教授

研究者番号：７０７２２８０９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：タグ付けを行った決算短信データ、タグ付与済みの英語ロイターニュース記事、FinCausalのデータセットを用いて因果関係インスタンス抽出実験を行い、日英の文書から因果関係インスタンスを抽出できる手法の開発に取り組んだ。結果的に、BERTとグラフニューラルネットワークを組み合わせることで既存の手法よりも高い精度で因果関係インスタンスを抽出できる手法の開発に成功した。
最終的に、本研究は、「FinancialCausality Extraction based on Universal Dependencies and Clue Expressions」というタイトルで査読付き論文誌に採録された。

研究成果の学術的意義や社会的意義
因果関係インスタンス抽出手法を作成するために、BERTモデルの改良の検討も行った。その結果、金融特化のBERTモデルの構築ができ、こちらをhugging faceにて公開した。また、その過程で得られた他のBERTモデルも公開し、公開したBERTモデルは幅広く利用されている。
作成した因果関係インスタンス抽出手法は、日本語と英語のみならず、学習データさえ存在すれば、他の言語でも利用可能であることから、今後の研究発展が期待される。

研究成果の概要（英文）：I conducted causal instance extraction experiments using tagged earnings release data, tagged English Reuters news articles, and FinCausal datasets to develop a method that can extract causal instances from Japanese and English documents. As a result, I succeeded in developing a method that can extract causal instances with higher accuracy than existing methods by combining BERT and graph neural networks.
Finally, this research was accepted for publication in a peer-reviewed journal under the title “ FinancialCausality Extraction based on Universal Dependencies and Clue Expressions.

研究分野：自然言語処理

キーワード：因果関係

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

COVID-19 を代表されるような社会的イベントが起こった場合、対策や結果は企業によって異なってくる。これは、ある事象が起こった場合に、考えられるシナリオが複数あることを意味し、個々の事象を把握する必要があることを意味する。すなわち、本研究で目指しているのは、一般的な知識を認識するのではなく、個々の対策や結果(これらを本研究では、インスタンスとする)を認識することである。例えば、「COVID-19 によって売り上げが減少した」という因果情報があった場合、これは飲食店や航空などの交通業界には当てはまる。しかしながら、任天堂などの家庭向けエンターテイメント事業を行っている会社は当てはまらない。さらに同業種においても、例えば、2021 年 3 月期の決算短信において、平和不動産は COVID-19 で売上減少としているのに対して、住友不動産は COVID-19 の影響で売り上げが下がった部門もあるが、全体的には過去最高益と報告している。このように、ある社会イベントに対して、個々の結果は様々であることから、「あるイベントから個々の結果へのパスを認識することで、それぞれのパスの違いを把握することができるのか？」という問いが生まれくる。もし、パスの違いを把握することができれば、すなわち、ここでは、各々の企業が考えている COVID-19 の影響が見えてくる。

上記を踏まえ、本申請では、今後のリスク対策や正確なシナリオ分析を行うために、個々の因果関係で表される情報を認識し、社会イベント発生から個々の影響へのパスを明らかにする。こちらが達成できれば、Robert Shiller 博士が提唱している Narrative Economics を実証でき、実経済における人々の期待形成の分析が可能になると考えらえる。

２．研究の目的

本研究の目的は、様々な文書から因果関係インスタンスを認識することで、社会イベント発生から個々の影響を把握することである。これを達成するためには、認識した因果関係インスタンスを結合することが必要である。「風が吹けば桶屋が儲かる」ではないが、本研究では、因果関係インスタンスの連鎖を因果チェーンと定義し、これを構築することで、影響のパスを把握する。また、申請者が指導学生と開発した解釈可能なニューラルネットワーク(Ito 2020)を用いて、文書を数値化することで、各因果関係インスタンスのセンチメント分析も可能になる。

言語処理による既存の因果関係を代表される知識の抽出においては、一般的な知識の抽出が主である。統計的な処理をすることが多いことから、代表的な知識のみの抽出となってしまうという手法による影響も考えられる。また、これらの研究においては、一般的な知識を体系化することで、検索や質問応答に役立てようという背景がある。そのため、個々の知識というよりは、万人が理解している一般的な知識の抽出に着眼点がおかれる。それに対して、本研究では、因果関係インスタンスを認識することで、イベントからの個々への影響のパスを把握する。そもそも、文書に記述された因果関係とは、書いた人が認識した因果関係であり、これは書いた人の立場や環境によって、異なってくることから、文書から抜き出す因果関係という意味では、本研究の考え方は妥当であると言える。

３．研究の方法

各文書に対して因果関係インスタスのタグを付与し、タグ付きデータを作成する。その後、既存の因果関係インスタンス抽出手法を用いて、各文書から因果関係インスタンスの抽出を試みる。申請者は、既に特許文書（坂地 2010）や金融テキスト（坂地 2015）からの因果関係インスタンス抽出の経験がある。さらに、決算短信からの因果関係インスタンス抽出に関する研究は、金融情報学研究会において、2013 年度の優秀論文賞に選ばれるなど、完成度の高い研究となっている。ここで、問題となったのは、各文書で異なる「手がかり表現」が存在することである(坂地 2011)。手がかり表現とは、因果知識を抽出する際に用いる手がかりとなる表現であり、例えば、日本語であれば「ため」や「から」、英語であれば「because」や「as」などがある。同じ新聞記事においても「を背景に」という手がかり表現は、スポーツ記事であれば、「選手が富士山を背景に走った」と実際の背景のことを指し、経済記事であれば、「景気悪化を背景に消費が落ち込む」のように原因を示す。これは日本語だけの問題ではなく、英語においても同様で、「since」や「by」、「as」は原因・結果を表すだけではなく、他の意味を持つ。そこで、本研究では意味の同定を行うことで言語によらず、かつ、精度高く判定できるモデルを構築する。

因果チェーンを構築するためには、因果関係インスタンスを構成する原因と結果を抽出する必要がある。因果関係インスタンスを構成する原因と結果は、文ごとに出現する位置が異なってくることから、申請者は 5 種類の構文的なパタンを作成し、抽出する研究(Sakaji 2008)を行った。この研究において、申請者は、手がかり表現を含む分析を核文節、核文節の係り先の文節を基点文節と定義した。この定義を用いて、原因と結果を抽出するアルゴリズムを構築し、高い精度で抽出することができるようになった。これを発展させる形で、BERT による系列ラベリングを行うモデルを作成した。単語ごとに評価した性能は高かったが、しかしながら、単語ごとにラベルを付与するということから、全て正解となる場合（完全一致）が少なく、因果チェーンへの応用が難しいことから、構文パタンを組合わせた新たな手法の開発を行っている。調査の結果、

日本語においても、英語においても、核文節と基点文節が存在すること（英語の場合は正しくは、文節ではなく単語になるため、核単語と基点単語となる）から、この定義を BERT モデルに組み込むことで、日・英どちらの文書からも原因と結果を抽出できるモデルの構築を目指す。

因果チェーンを構築する手法に関しての試作版は、国際ワークショップ(FinNLP)で発表済みである（Izumi 2019）。さらに、こちらの発表はワークショップの Best Paper にも選ばれるなど、今後の期待が国際的にも高い。本研究では、こちらの研究を進めるため、異なる因果関係インスタンスの結合に関するデータセットを作成し、評価実験を進めていく。また、関連のある研究タスクとして、文の類似度を測る研究があることから、こちらの研究タスクにおいても高い性能を達成できるような手法構築を目指す。具体的には、Sentence-BERT をベースに、因果関係インスタンスを結合するために、fine-tuning する BERT の数を 2 つから 4 つに変更し(原因と結果が 2 つずつ存在することから)、Loss 関数をそれに合わせて変更することを考えている。これは、文対の類似性を判定する問題と異なり、因果関係インスタンスには原因と結果が含まれていることから、それぞれの類似性を計算できるようにすることで、経済分析につながるような手法の構築が可能になると考えたためである。

４．研究成果

研究成果として得られた、構文情報と手がかり表現の両方を使用して原因と結果の表現を抽出する提案モデルについて説明する。提案モデルは， bidirectional encoder representations from transformers (BERT)とグラフニューラルネットワークを用いている．また，構文情報を有効に利用するために，グラフニューラルネットワークとして graph attention networks (GAT) を採用する．さらに，入れ子の因果関係を処理するために，提案モデルは手がかり表現を利用する．ここで，提案モデルは， gated recurrent unit (GRU)を使用してエンコードされた手がかり表現を用いる．図 1 に入れ子の因果関係例を示し，図 2 に，提案モデルの概要を示す．図 1 において，<b>タグは原因表現，<r>タグは結果表現，<c>タグは手がかり表現をそれぞれ示す．また，タグ内の数字は，対になっている原因・結果表現を識別するためのものである．図 1 では，1 文内に 3 つの因果関係が含まれており，それぞれ，原因「震災後の消費マインドの冷え込み」，結果「低価格志向が強まり」の対と，原因「原発事故」，結果「放射能汚染」の対，原因「原発事故による放射能汚染の影響」，結果「消費者の食の「安全・安心」に対する意識が一層高まる」の対となっている．

食品業界におきましては、<b1>震災後の消費マインドの冷え込み</b1><c1>から</c1><r1>低価格志向が強まり</r1>、また<b3><b2>原発事故</b2><c2>による</c2><r2>放射能汚染</r2>の影響</b3><c3>から、</c3><r3>消費者の食の「安全・安心」に対する意識が一層高まる</r3>など厳しい事業環境となりました。

図 1：入れ子の因果関係例

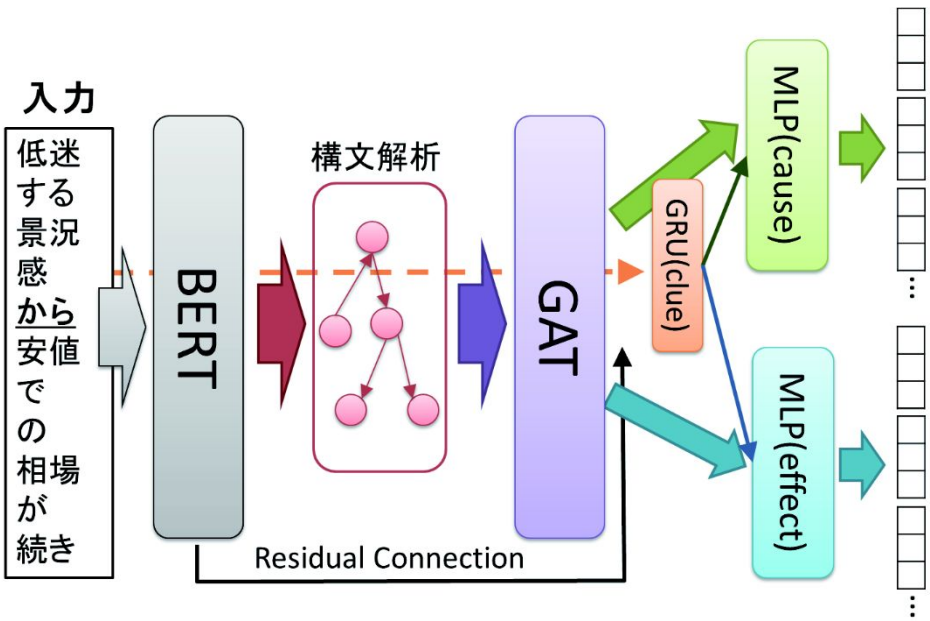

図 2：手法の概要

表 1：ロイターニュース記事を用いた実験結果

	P	*R*	F1	EM
BiLSTM [26]	0.678	0.619	0.625	0.160
BERT	0.695	0.652	0.654	**0.322**
MB	0.794	0.795	0.794	0.071
CMB	0.797	0.802	0.799	0.247
GCMB (proposed model)	**0.798**	**0.808**	**0.801**	0.258
GCMB_GCN	0.788	0.796	0.791	0.217

表 2：決算短信を用いた実験結果

	P	*R*	F1	EM
BiLSTM [26]	0.634	0.592	0.592	0.033
BERT	0.684	0.677	0.677	0.103
MB	0.795	0.777	0.775	0.063
CMB	0.879	0.854	0.864	0.204
GCMB (proposed model)	**0.890**	**0.890**	**0.890**	**0.234**
GCMB_GCN	0.880	0.871	0.875	0.199

表 1 にロイターニュース記事のデータセット（英語）を用いた実験結果、表 2 に決算短信データセット（日本語）を用いた実験結果をそれぞれ示す。表 1 と表 2 において、MB はマルチタスク学習と BERT を組み合わせたモデル、CMB は手がかり表現エンコーダーと MB を組み合わせモデル、GCMB は GAT と CMB を組み合わせたモデルを示す。また、GCMB_GCN は、GAT に代わりに GCN を用いたモデルとなっている。表 1 と表 2 から、提案手法（GCMB）が日本語においても、英語においても最も高い因果関係抽出性能を示した。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Masahiro Suzuki, Hiroki Sakaji, Masanori Hirano, Kiyoshi Izumi	４．巻 60
２．論文標題 Constructing and analyzing domain-specific language model for financial text mining	５．発行年 2023年
３．雑誌名 Information Processing & Management	６．最初と最後の頁 103194～103194
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.ipm.2022.103194	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Sakaji Hiroki、Izumi Kiyoshi	４．巻 41
２．論文標題 Financial Causality Extraction Based on Universal Dependencies and Clue Expressions	５．発行年 2023年
３．雑誌名 New Generation Computing	６．最初と最後の頁 839～857
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00354-023-00233-2	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計9件（うち招待講演　1件／うち国際学会　3件）

１．発表者名 Hiroki Sakaji, Masahiro Suzuki, Kiyoshi Izumi, Hiroyuki Mitsugi
２．発表標題 Gradual Further Pre-training Architecture for Economics/Finance Domain Adaptation of Language Model
３．学会等名 2022 IEEE International Conference on Big Data (IEEE BigData 2022)（国際学会）
４．発表年 2022年

１．発表者名 高柳剛弘，坂地泰紀，和泉潔
２．発表標題 銘柄特徴と投資家特性を考慮した株式銘柄推薦の個別化
３．学会等名 言語処理学会第29回年次大会(NLP2023)
４．発表年 2023年

１．発表者名
小林涼太郎，坂地泰紀，和泉潔

２．発表標題
BERTとGATを用いた金融テキストにおける因果関係を含む文の判定

３．学会等名
言語処理学会第29回年次大会(NLP2023)

４．発表年
2023年

１．発表者名
高柳剛弘，坂地泰紀，和泉潔

２．発表標題
個別銘柄情報と銘柄間情報を利用したテーマ株抽出手法の提案

３．学会等名
2022年度人工知能学会全国大会

４．発表年
2022年

１．発表者名
坂地泰紀，和泉潔，加藤惇雄，長尾慎太郎

２．発表標題
系列ラベリングによる原因・結果表現抽出の試み

３．学会等名
第18回テキストアナリティクス・シンポジウム

４．発表年
2021年

１．発表者名
金融ドメインにおける事前学習BERTモデルの性能検証

２．発表標題
鈴木雅弘，坂地泰紀，平野正徳，和泉潔

３．学会等名
第18回テキストアナリティクス・シンポジウム

４．発表年
2021年

１．発表者名
鈴木雅弘，坂地泰紀，平野正徳，和泉潔

２．発表標題
金融文書を用いた事前学習言語モデルの構築と検証

３．学会等名
人工知能学会第27回金融情報学研究会

４．発表年
2021年

１．発表者名
Hiroki Sakaji, Noriyasu Kaneda

２．発表標題
Indexing and Visualization of Climate Change Narratives Using BERT and Causal Extraction

３．学会等名
2023 IEEE International Conference on Big Data (IEEE BigData 2023)（国際学会）

４．発表年
2023年

１．発表者名
Hiroki Sakaji

２．発表標題
Economic Causal-Chain Search using Text Mining Technology

３．学会等名
Knowledge Graphs in Finance and Economics（招待講演）（国際学会）

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
Hiroki's page
https://testuwaka.net/
事前学習言語モデル
https://sites.google.com/socsim.org/izumi-lab/tools/language-model
Hiroki's page
http://tetsuwaka.net/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関