1N3-OS-39b-4

社会課題における因果関係を表すLinked Dataの半自動的な構築手法の提案

A Method for Semi-automatic Constructing Linked Urban Problem Data with Causal Relations

江上周作 *1 Shusaku Egami　川村隆浩 *1*2 Takahiro Kawamura　古崎晃司 *3 Kouji Kozaki　大須賀昭彦 *1 Akihiko Ohsuga

*1電気通信大学大学院情報理工学研究科
Graduate School of Informatics and Engineering, The University of Electro-Communications

*2科学技術振興機構 情報分析室
Department of Information Planning, Japan Science and Technology Agency

*3大阪大学産業科学研究所
The Institute of Scientific and Industrial Research (ISIR), Osaka University

Municipalities have some urban problems such as traffic accidents, illegally parked bicycles, and noise pollution in Japan. However, since these data are not structurally managed, to solve urban problems using these data is not facilitated. Thus, we aim to construct the Linked Data infrastructure for solving urban problems. In this paper, we propose a method for semi-automatic constructing Linked Data with causal relations of urban problems based on Web pages and open government data. Specifically, we extracted causal relations using natural language processing and crowdsourcing. As the result, Linked Data that have causal relations of noise pollution, illegally parked bicycles and traffic accidents is constructed.

1. はじめに

現在，我が国では交通事故，放置自転車，騒音，ごみ問題など複数の社会課題を抱えており，自治体ごとに対策に向けた議論や施策検討が行われている．こうした社会課題の解決に向けたデータドリブンアプローチが注目されており，2013 年に行われた G8 サミットでの「オープンデータ憲章」への合意以降，自治体によるオープンデータの取り組みが普及している．大阪市ではオープンデータポータルサイト *1 で約 10,000 件のデータセットを公開しており，オープンデータを活用した課題解決に役立つアプリケーションの事例を紹介している．横浜市では，NPO 法人横浜コミュニティデザイン・ラボにより，横浜市オープンデータを活用した地域課題解決のビジネス創出を支援するクラウドファンディングサイトが運営されている *2．

このようにオープンデータを用いた社会課題解決が各地で注目されているが，より課題解決を進めるためには自治体の抱える各種社会課題の分類や因果関係，解決に向けた施策や効果に関するデータの整理が望まれる．そこで，本研究ではこれらのデータのスキーマを整備し，Linked Open Data (LOD) としてデータを蓄積することで，社会課題の分析基盤の構築を目指している．この社会課題の分析基盤の実現により，例えば社会課題の因果関係や階層関係のリンクをたどって課題の影響範囲を分析・予測することや，その結果を基に自治体の対策方針を検討することが可能になると考える．本稿では，社会課題の因果関係を表す LOD を半自動的に構築する手法を提案する．具体的には，社会課題の中でもまず放置自転車問題，交通事故，騒音問題を例に，Web ページや自治体オープンデータを基に因果単語の抽出を行い，クラウドソーシングによる因果単語の絞り込みを行う．その後，クラス・インスタンスのスキーマを設計し因果関係を表す LOD を構築する．

連絡先: egami.shusaku@ohsuga.lab.uec.ac.jp

∗1 https://data.city.osaka.lg.jp/

∗2 http://yokohama.localgood.jp/

以下，2 章では関連研究について述べ，3 章では社会課題因果関係 LOD の半自動的な構築手法について述べる．4 章では提案手法の結果と考察を行い，最後に 5 章でむすびとする．

2. 関連研究

社会課題解決に向けた LOD 化の取り組みとしては PUSH 大阪がある [古崎 16]．この研究では大阪市の広報情報 RSS(RDF Site Summary) を収集し LOD として公開しており，大阪市の課題解決に向けた LOD の活用を試行している．この LOD と本稿で述べる我々の LOD をリンクすることで更なる分析が可能になると考える．白松らは社会課題解決目標の LOD 化により，課題解決者のマッチングやシビックテックを促進させる研究 [白松 16] を行っており，本稿で述べる LOD とのリンクにより課題解決促進が期待できる．

3. 社会課題因果関係 LOD の半自動構築

我々はこれまで社会課題の一つである放置自転車問題解決に向けてデータを収集し，継続的に放置自転車 LOD を構築してきた [Egami 16]．その際に，日常的に生じる社会課題の LOD スキーマ設計の手順を方法論としてまとめた．さらに因果関係の構築に向けて LOD スキーマ設計の方法論を次のように拡張した [江上 17]．本稿では放置自転車の他に交通事故，騒音についても本手法を適用した．

1. 因果単語の抽出
   a. 検索エンジンを用いた記事検索および収集
   b. 収集した記事から因果単語を抽出
   c. 抽出した単語を基にワードクラウドを生成
   d. クラウドソーシングによるキーワードの選択
2. スキーマの設計

a. 対象とする社会課題をモデル化する既存オントロジーを選択

b. 既存オントロジーを軸として因果単語を基にクラスを設計

c. 設計したクラスを基にインスタンスを設計

3.1 因果単語の抽出

まず，対象とする社会課題名とその同意語を一番目の検索語とし，“要因” の同意語を二番目の検索語として検索エンジンを用いた記事検索を行う．以下，騒音の要因を例として説明する．騒音の場合，一番目の検索語は “騒音” を使用し，二番目の検索語は “要因” とその同意語であり，日本語 WordNet[*3] から “因子”，“素因”，“要素”，“導因”，“もと”，“原因”，“誘因”，“起こり”，“起り” を取得して使用する．同意語として取得したが日本語文書において使用頻度の少ないと考えられる，“factor”，“ファクタ”，“エレメント”，“ファクター” に関しては，これらの検索語がメインとして検索されてしまったため除外した．因果関係における本研究では検索エンジンとして Google を使用し，Google Custom Search API[*4] に「騒音 要因」のようなキーワードパラメータを与えて結果を取得する．取得するファイルタイプとして HTML と PDF を別々に指定し，一番目の検索語と二番目の検索語の全ての組み合わせで記事を 20 件ずつ検索する．HTML と PDF を分けて検索した意図としては，社会課題の因果関係について言及している記事を，行政の調査報告や議事録等と一般人の SNS やブログ等から満遍なく収集するためである．また，記事中における社会課題名の出現回数が低い場合は，この後の要因単語抽出処理でノイズが多く混ざってしまうため，社会課題名の出現回数でフィルタリングを行う．このような記事は，例えば地域の抱える様々な問題（育児，教育，医療，交通等）について全て述べた長文記事が該当する．今回は，“騒音” の出現回数が合計 3 回未満の記事を解析対象外とした．

次に，収集した記事を一文ごとに分割し，文ごとに形態素解析を行い名詞を抽出する．形態素解析には mecab-ipadic-NEologd[*5] を使用した．この後の処理であるクラウドソーシングによる要因単語選択の処理において，ある程度意味が通じる名詞となるように，サ変接続名詞の直前・直後に名詞がある場合はそれらを連結して一つの名詞とする．例えば “駐輪料金” は通常 “睡眠 (名詞-サ変接続)” と “妨害 (名詞-サ変接続)” に分割されてしまうが，本手法ではサ変接続を結合して “睡眠妨害” とする．

次に TEXT2LOD[Kawamura 16] を用いて各文を RDF トリプルに分解する．TEXT2LOD は自然言語処理，ルール適用，条件付き確率場により，与えられた文からチャンク（句）を生成して関係抽出を行い，主語，述語，目的語（トリプル）を Linked Data として生成する Web API である．TEXT2LOD から得られた結果に対して，二番目の検索語を含む句を有するトリプルから，該当する句以外の句を全て取得する．取得した句の中に形態素解析を実行して得られた名詞が存在する場合，その名詞を要因単語として抽出する．

同様に社会課題が引き起こす事象や影響（以下，影響単語）に関しての抽出も行う．日本語 WordNet から “引き起こす”，“惹起”，“もたらす”，“来たす”，“招く”，“誘発”，“生み出す” などの同意語を取得し，これらを二番目の検索語として要因抽出と同様の処理を行う．

*3 http://nlpwww.nict.go.jp/wn-ja/
*4 https://developers.google.com/custom-search/?hl=ja
*5 https://github.com/neologd/mecab-ipadic-neologd

図 1: 騒音の要因単語のワードクラウド

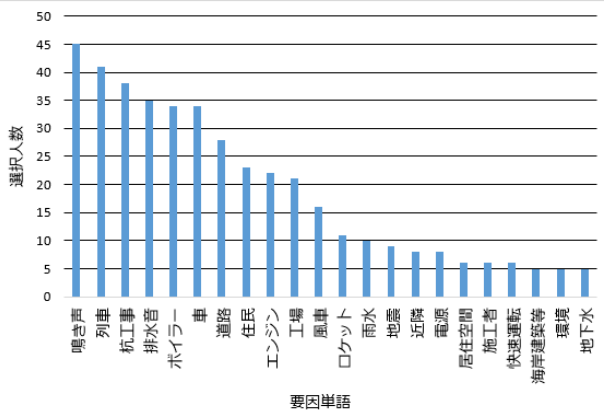

図 2: 最終的に選択した騒音の要因単語

3.2 クラウドソーシングによる因果単語の選択

次に抽出した因果単語を基にワードクラウドを生成し，クラウドソーシングによる因果単語の選択を行う．図 1 に騒音の要因単語のワードクラウドを，図 3 に騒音の影響単語のワードクラウドを示す．出現回数が多いほど中心に位置し文字サイズが大きくなる．文字の色はランダムに決定する．このワードクラウドから「騒音の原因として考えられるものを 10 個選択する」「騒音が引き起こすものや影響範囲として考えられるものを 10 個選択する」というタスクを設定し，クラウドソーシングによる因果単語の選択を行う．今回は金銭的報酬を与えるマイクロタスク型クラウドソーシングを行った．クラウドソーシングサービスとしてランサーズ [*6] を使用し，上記 2 つのタスクの遂行報酬を 50 円として，一つの社会課題につき最大 50 人に作業を依頼した．このうち作業人数の 1 割以上に選択された単語を使用する．騒音の場合 50 人に作業されたため，5 人以上に選択された単語を使用する．選択された単語の中に類似の単語がある場合は，選択した人数の多い単語に統一する．最終的に抽出された放置自転車の要因単語を図 2 に，影響単語を図 4 に示す．

3.3 因果関係を含むスキーマの設計

3.2 節で最終的に選択した単語を基に因果関係を含むスキーマを設計する．まず，対象とする社会課題をモデル化する既存オントロジーを選択する．本研究では放置自転車，交通事故，騒音をイベントとみなし，Event Ontology (EO)[*7] を選択した．次に，対象とする社会課題名のクラスを作成し，EO における Event クラスのサブクラスとする．次に，抽出した単語を基にクラスを作成し，因果関係のプロパティでリンクする．

*6 http://www.lancers.jp/
*7 http://motools.sourceforge.net/event/event.html

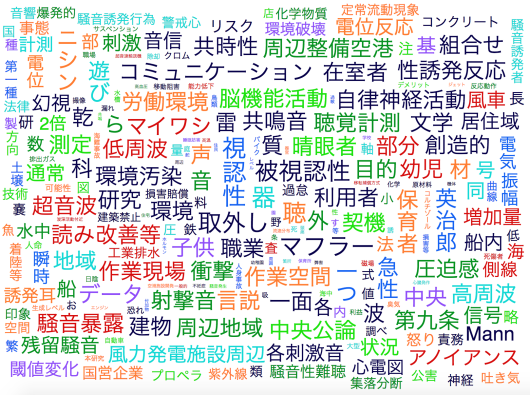

図 3: 騒音の影響単語のワードクラウド

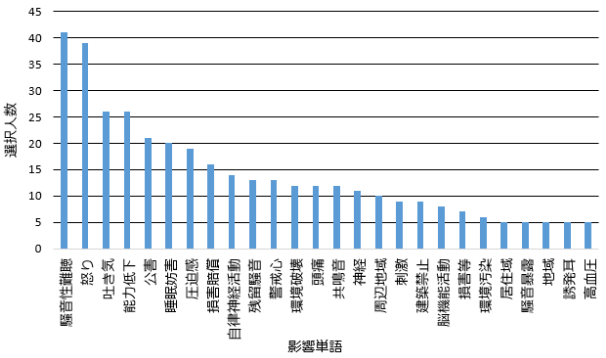

図 4: 最終的に選択した騒音の影響単語

EO には要因として factor プロパティが存在するためこれを再利用する．しかし，社会課題が引き起こす物や影響に相当するプロパティが EO に存在しないため，affect プロパティを新たに定義してリンク付けする．また，place，time，agent，product，sub_event プロパティを EO から再利用し，EO と同様の構造で設計する．さらに，対象とする社会課題において重要であると思われる項目がある場合，新たにプロパティとその値を追加する．以下に社会課題のクラス要件を記述論理で示す．i, j は要因単語，影響単語において最終的に選択された単語の数である．

社会課題 $\sqsubseteq$ Event

社会課題 $\sqsubseteq \exists$place.SpatialThing

社会課題 $\sqsubseteq \exists$time.TemporalEntity

社会課題 $\sqsubseteq \exists$agent.Agent

社会課題 $\sqsubseteq \exists$product.Thing

社会課題 $\sqsubseteq \exists$sub_event.Event

社会課題 $\sqsubseteq \exists$factor.(要因 1 $\sqcup$ 要因 2 $\sqcup$ ... $\sqcup$ 要因 i)

社会課題 $\sqsubseteq \exists$affect.(影響 1 $\sqcup$ 影響 2 $\sqcup$ ... $\sqcup$ 影響 j)

次に各クラスのインスタンス部分を設計する．各クラスのインスタンスは実データを基に作成されるため，データとして取得することが困難な場合は LOD 化する際にインスタンスを作成しない．最終的に設計した LOD スキーマの一部を図 5 に示す．一般的に考えて因果関係があると考えられるが，本手法ではリンクすることができなかった関係を破線で示している．最終的に構築した LOD をオープンライセンスで Web 上に公開している *8．

表 1: 因果単語抽出結果

	全記事 (使用)	全文 (抽出文)	表示単語	一致度
騒音要因	210 (44)	20,771 (65)	204	0.468
騒音影響	400 (68)	54,816 (115)	233	0.289
交通事故要因	304 (81)	41,932 (112)	358	0.443
交通事故影響	343 (71)	61,147 (108)	244	0.194
放置自転車要因	330 (83)	64,676 (103)	237	0.175
放置自転車影響	323 (57)	164,238 (94)	214	0.218

4. 結果と考察

因果関係抽出処理において，最終的に収集された記事数，使用した記事数，文数，抽出文，ワードクラウド表示単語数，単語選択の一致度の結果を表 1 に示す．使用した記事数が減った理由は，リンク切れページや画像 PDF が多く存在したことや，Google 検索において二番目の検索語を重視してマッチした記事が取得されてしまったことにある．特に “素因” や “誘因” などは日常生活において使用頻度が少ないため，これらの単語が重視されて検索されることとなり，因果関係記述を多く含む医療関係の記事がノイズとして多く取得された．今後は「“騒音の原因”」など，検索語をスペースで区切らずにフレーズとして検索することを検討している．

また，クラウドソーシングによる因果単語選択の一致度を算出するために Fleiss の kappa 係数 [Fleiss 73] を計算した．表 1 より，騒音が与える影響，交通事故の要因に関しては適度な一致度が確認でき，それ以外に関しては弱い一致度が確認できた．この結果から，因果単語の抽出結果およびワードクラウドの表示はある程度妥当であったと考えられる．

交通事故の要因の一致度が比較的高かったのは，交通事故の要因について警視庁，教育機関，報道機関などが一般に向けて説明する機会が多いため，ワーカーの予備知識が共通していたのではないかと考える．一方で，騒音の要因に関しても比較的高い一致度が得られたのは，ワーカーが日常的に騒音を経験していることや，ワードクラウドに表示された単語が連想しやすいものであったと考えられる．その他が弱い一致度となった原因としては，各社会課題と関連しない単語が大きく中心に集まってしまったことが原因である．その結果，本来関係のある単語が目立たなくなり選択されづらくなったと考えられる．例えば図 3 に示した騒音の影響のワードクラウドでは，“視認性” や “マイワシ” などの騒音との関連が不明な単語が目立つ．“視認性” は不正改造車に関する記事に出現し，着色フィルム等のガラスへの装着が “視認性” の低下を “招く” ことに関する文章から抽出された．同記事では，マフラーを外すことが騒音の増大を招くことについても記述している．また，“マイワシ” については魚類の聴性誘発反応に関する文献に出現し，“マイワシ” の聴性 “誘発” 反応の計測に関する文章から抽出された．同記事では，低周波騒音が魚の行動に影響を与えることについても記述している．このように，記事中に騒音の記述があるが別の箇所から因果関係が抽出された例や，社会課題とは異なる記述から因果関係が抽出された例が多く見られた．これらのノイズを除去するために，単一文書内から複数回出現した単語より，複数記事に出現した単語を重視して抽出することを検討している．また，一致度の低下からワーカーの予備知識が共通していないことも考えられ，社会課題の要因と影響の予備知識・問題意識が不足しているからこそ，社会課題が解決して

*8 http://www.ohsuga.lab.uec.ac.jp/socialproblem/

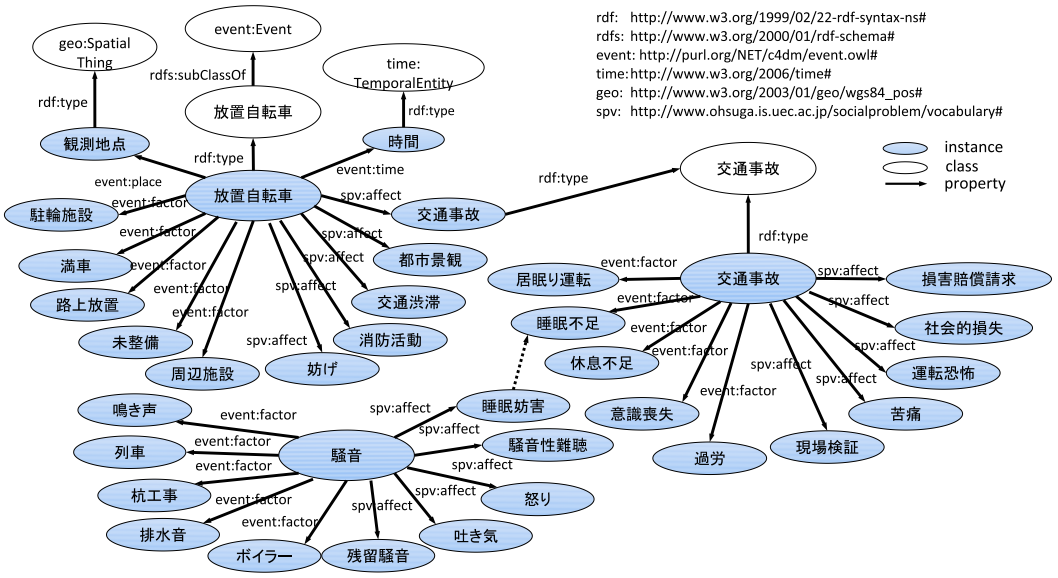

図 5: 設計した LOD スキーマの一部

いないことも読み取れる．

さらに，形態素解析と TEXT2LOD を用いた因果単語抽出処理において，本来社会課題と関係のある単語が抽出漏れしていることも一致度を下げた原因である．特に TEXT2LOD で SVO に分割された句の内，要因について記述している句と今回抽出した句が一致しない場合があることがわかった．これは行政の発行する記事に冗長的な言い回しが多く，一つの文が複文で構成されていることが多いからである．対象とする社会課題の因果単語をより高精度に抽出するために，因果単語の抽出ルールを新たに設定する必要があると考える．

次に LOD スキーマについて考察する．本稿で設計したスキーマとこれまでに設計していた放置自転車 LOD[Egami 16] のスキーマを比較すると，スキーマ設計の段階では要因リソースと結果リソースの数が大幅に増加している．さらに，放置自転車と交通事故という異なる社会課題間の因果関係を抽出することもできた．今後さらに多くの社会課題の因果関係を LOD として構築した際に，一つの事象が複数の社会課題に及ぼす影響を推論できる可能性を確認できた．また，図 5 のように “睡眠妨害” と “睡眠不足” のように，異なる社会課題の因果単語から新たな因果関係を発見した．このような関係を自動でリンク付けすることは今後の課題としたい．

しかしながら，設計した LOD スキーマにおけるインスタンス部分は実データとして取得できるものが少なく，LOD 構築の段階ではこれまでの放置自転車 LOD と比べて要因インスタンスの数が減少するという結果になった．そのため，クラスレベルでの因果関係分析に使用できる可能性があるが，インスタンスレベルでの因果関係分析を目的として使用することは現状では難しい．今後は，現在データとして取得することができないものについて，新たにオープンデータ化するように行政に働きかける必要がある．また，要因単語クラスと影響単語クラスの粒度を，実データを取得できる範囲で統一化・階層化する必要があると考える．

5. むすび

本稿では社会課題の分析基盤の実現に向けて，社会課題の因果単語を Web 記事や行政のオープンデータ等から抽出し，抽出結果をクラウドソーシングを用いて絞り込むことで，社会課題の因果関係を含む LOD を構築した．特に，本稿では数ある社会課題の中でもまず放置自転車，交通事故，騒音を対象として LOD を構築した．今後さらに多くの社会課題を対象とし，因果単語抽出精度を向上させて LOD を構築する予定である．さらに，自治体の政策内容やその効果などをオープンデータ，SNS 等から抽出して LOD に追加することで，自治体の抱える課題の解決に向けた政策決定を支援する分析基盤の構築を目指す．

謝辞

本研究は JSPS 科研費 16K12411,16K00419,16K12533 の助成を受けたものです．

参考文献

[江上 17] 江上周作，川村隆浩，古崎晃司，大須賀昭彦，放置自転車問題における因果関係を含む LOD の半自動的な構築手法の提案, 人工知能学会研究会資料，SIG-SWO-041-10 (2017)

[古崎 16] 古崎晃司，山本泰智，自治体広報情報の RSS に基づく地域課題分析の試み，人工知能学会全国大会（第 30 回）論文集，1N5-OS-19b-5 (2016)

[白松 16] 白松俊, Tossavainen, T., 大囿忠親, 新谷虎松，社会課題とその解決目標の Linked Open Data 化による目標マッチングサービスの開発，人工知能学会論文誌, Vol.30, No.1 (2016)

[Egami 16] Egami, S., Kawamura, T., Ohsuga, A.: Building Urban LOD for Solving Illegally Parked Bicycles in Tokyo. In: Proc of the 15th International Semantic Web Conference (ISWC), pp.291-307 (2016)

[Fleiss 73] Fleiss, J. L., Cohen, J.: The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient As Measures of Reliability. Educational and Psychological Measurement, Vol.33, No.3, pp.613-619 (1973)

[Kawamura 16] Kawamura, T., Ohsuga, A.: Development of Web Service for Japanese Text Triplification. New Generation Computing, Vol.34, No.4, pp.307-322 (2016)